Methodology of Educational Measurement and Assessment

Matthias von Davier
Young-Sun Lee *Editors*

Handbook of Diagnostic Classification Models

Models and Model Extensions, Applications, Software Packages

Springer

Methodology of Educational Measurement and Assessment

This book series collates key contributions to a fast-developing field of education research. It is an international forum for theoretical and empirical studies exploring new and existing methods of collecting, analyzing, and reporting data from educational measurements and assessments. Covering a high-profile topic from multiple viewpoints, it aims to foster a broader understanding of fresh developments as innovative software tools and new concepts such as competency models and skills diagnosis continue to gain traction in educational institutions around the world. Methodology of Educational Measurement and Assessment offers readers reliable critical evaluations, reviews and comparisons of existing methodologies alongside authoritative analysis and commentary on new and emerging approaches. It will showcase empirical research on applications, examine issues such as reliability, validity, and comparability, and help keep readers up to speed on developments in statistical modeling approaches. The fully peer-reviewed publications in the series cover measurement and assessment at all levels of education and feature work by academics and education professionals from around the world. Providing an authoritative central clearing-house for research in a core sector in education, the series forms a major contribution to the international literature.

More information about this series at http://www.springer.com/series/13206

Matthias von Davier • Young-Sun Lee
Editors

Handbook of Diagnostic Classification Models

Models and Model Extensions, Applications, Software Packages

Springer

Editors
Matthias von Davier
National Board of Medical
Examiners (NBME)
Philadelphia, PA, USA

Young-Sun Lee
Teachers College
Columbia University
New York, NY, USA

ISSN 2367-170X ISSN 2367-1718 (electronic)
Methodology of Educational Measurement and Assessment
ISBN 978-3-030-05583-7 ISBN 978-3-030-05584-4 (eBook)
https://doi.org/10.1007/978-3-030-05584-4

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The *Handbook of Diagnostic Classification Models* represents a collection of chapters reviewing diagnostic models, their applications, and descriptions of software tool, written by leading experts in the field. This volume covers most (one can never claim completeness) of the current major modeling families and approaches as well as provides a resource that can be used for self-study, teaching, or research that applies or extends the materials included in the book.

While virtually any project of this type takes longer than expected, and many will be tempted to remind us that Murphy's law strikes almost surely, we were amazed by the willingness of all contributors to put in the hours to finish their chapters and to review other chapters and, finally, to revise their contributions in order to help putting together a coherent volume. We hope that this process, together with some occasional assistance from the editors and the publisher, helped to compile a multi-authored work together that covers most aspects of doing research around diagnostic modeling.

We also want to remind readers as well as ourselves of colleagues who passed away and who leave a void in the research community. We lost Kikumi Tatsuoka, of whom one can truthfully say that her rule space approach is one of the major roots, maybe even the most important one, of this field. In her long career, she shaped many aspects of diagnostic modeling, and we should recall that, among these, the Q-matrix is one of the central building blocks present in the vast majority of these methods. The rule space method is described along with other early approaches in Chap. 1.

We furthermore would like to remember Lou DiBello, who made important contributions to the field, notably in his modified rule space work, and his work on the unified model together with colleagues. The work around extensions of the unified model is described in Chap. 3. We also want to remind readers of Wen-Chung Wang who just recently passed away. Wen-Chung and his coauthors worked on many topics around diagnostic models and other psychometric approaches. His work around DIF methods for use with diagnostic modeling approaches is found in Chap. 18. We hope that the friends we lost would have liked this volume.

Ending on a more positive note: working in a dynamic field that produces new knowledge every day, we are aware that the handbook is one stepping stone on the long path to fully understanding the potential of these powerful modeling approaches. We are expecting to see books that extend the material we have put together here; moreover, we expect to see this handbook be replaced or superseded by a new edition in a couple of years. If we are lucky, we may be involved in putting together some of the chapters of these future collections describing what will then be the state of the art in diagnostic modeling.

Philadelphia, PA, USA — Matthias von Davier
New York, NY, USA — Young-Sun Lee

Contents

Contributors

Stephan Abele Institute of Vocational Education and Vocational Didactics, Technische Universität Dresden, Dresden, Germany

Russell G. Almond Department of Educational Psychology and Learning Systems, Florida State University, Tallahassee, FL, USA

Daniel M. Bolt Department of Educational Psychology, University of Wisconsin – Madison, Madison, WI, USA

Li Cai University of California, Los Angeles, CA, USA

Vector Psychometric Group, LLC, Chapel Hill, NC, USA

Hua-Hua Chang Department of Educational Studies, Purdue University, IN, USA

Yunxiao Chen London School of Economics and Political Science, London, UK

Ying Cheng Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

Chia-Yi Chiu Department of Educational Psychology, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

Pravin Chopade ACTNext ACT Inc., Iowa City, IA, USA

Steven Andrew Culpepper Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Jimmy de la Torre Division of Learning, Development and Diversity, University of Hong Kong, Hong Kong, China

Lawrence T. DeCarlo Department of Human Development, Teachers College, Columbia University, New York, NY, USA

Benjamin Deonovic ACTNext ACT Inc., Iowa City, IA, USA

Lou DiBello Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Jeff Douglas Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Susan E. Embretson School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

Meghan Fager National University, Precision Institute, La Jolla, CA, USA

Ann Cathrice George Federal Institute for Educational Research, Innovation and Development of the Austrian School System, Salzburg, Austria

Zhuangzhuang Han Department of Human Development, Teachers College, Columbia University, New York, NY, USA

Robert Henson Educational Research Methodology (ERM) Department, The University of North Carolina at Greensboro, Greensboro, NC, USA

Carrie R. Houts Vector Psychometric Group, LLC, Chapel Hill, NC, USA

Hong Jiao Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA

Matthew S. Johnson Educational Testing Service, Princeton, NJ, USA

Hyeon-Ah Kang Department of Educational Psychology, University of Texas, Austin, TX, USA

Lale Khorramdel National Board of Medical Examiners (NBME), Philadelphia, PA, USA

Hans-Friedrich Köhn Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Young-Sun Lee Teachers College, Columbia University, New York, NY, USA

Xiaomin Li The University of Hong Kong, Pokfulam, Hong Kong

Xiaoou Li School of Statistics, University of Minnesota, Minneapolis, MN, USA

Dandan Liao American Institutes for Research, Washington, DC, USA

Jingchen Liu Department of Statistics, Columbia University, New York, NY, USA

Xiang Liu Department of Human Development, Teachers College, Columbia University, New York, NY, USA

Diego A. Luna-Bazaldua School of Psychology, National Autonomous University of Mexico, Mexico City, Mexico

Wenchao Ma The University of Alabama, Tuscaloosa, AL, USA

Nathan D. Minchen Pearson, Bronx, NY, USA

Jesse Pace University of Kansas, Lawrence, KS, USA

Yoon Soo Park Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, IL, USA

Xue-Lan Qiu The University of Hong Kong, Pokfulam, Hong Kong

Alexander Robitzsch Department of Educational Measurement, IPN Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Centre for International Student Assessment, Munich, Germany

Gina Rossi Research Group Personality and Psychopathology, Vrije Universiteit Brussel, Brussels, Belgium

Benjamin Shear Research and Evaluation Methodology, University of Colorado, Boulder, CO, USA

Hyo Jeong Shin Educational Testing Service, Princeton, NJ, USA

Klaas Sijtsma Department of Methodology and Statistics, TSB, Tilburg University, Tilburg, The Netherlands

Sandip Sinharay Educational Testing Service, Princeton, NJ, USA

William Stout Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Jonathan L. Templin Educational Measurement and Statistics Program, University of Iowa, Iowa City, IA, USA

L. Andries van der Ark Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

Alina A. von Davier ACTNext ACT Inc., Iowa City, IA, USA

Matthias von Davier National Board of Medical Examiners (NBME), Philadelphia, PA, USA

Shiyu Wang Department of Educational Psychology, University of Georgia, Athens, GA, USA

Wen-Chung Wang The University of Hong Kong, Pokfulam, Hong Kong

Gongjun Xu Department of Statistics, University of Michigan, Ann Arbor, MI, USA

Xueli Xu Educational Testing Service, Princeton, NJ, USA

Xiaofeng Yu Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

Jiangxi Normal University, Nanchang, China

Michael Yudelson ACTNext ACT Inc., Iowa City, IA, USA

Juan-Diego Zapata-Rivera Educational Testing Service, Princeton, NJ, USA

Peida Zhan Department of Psychology, College of Teacher Education, Zhejiang Normal University, Zhejiang, China

Susu Zhang Department of Statistics, Columbia University, New York, NY, USA

Chapter 1
Introduction: From Latent Classes to Cognitive Diagnostic Models

Matthias von Davier and Young-Sun Lee

Abstract This chapter provides historical and structural context for models and approaches presented in this volume, by presenting an overview of important predecessors of diagnostic classification models which we will refer to as DCM in this volume, or alternatively cognitive diagnostic models (CDMs). The chapter covers general notation and concepts central to latent class analysis, followed by an introduction of mastery models, ranging from deterministic to probabilistic forms. The ensuing sections cover knowledge state and rule space approaches, which can be viewed as deterministic skill-profile models. The chapter closes with a section on the multiple classification latent class model and the deterministic input noisy and (DINA) model.

1.1 Introduction

This chapter provides historical and structural context for models and approaches presented in this volume, by presenting an overview of important predecessors of diagnostic classification models which we will refer to as DCM in this volume, or alternatively cognitive diagnostic models (CDMs). We are attempting to organize the growing field somewhat systematically to help clarify the development and relationships between models. However, given the fact that DCMs have been developed based on at least two, if not three traditions, not all readers may necessarily agree with the order in which we put the early developments. While there is a multitude of approaches that can be considered predecessors of current

M. von Davier (✉)
National Board of Medical Examiners (NBME), Philadelphia, PA, USA
e-mail: mvondavier@nbme.org

Y.-S. Lee
Teachers College, Columbia University, New York, NY, USA
e-mail: sly2003@columbia.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_1

approaches to cognitive diagnostic modeling, there are many connections between these seemingly different approaches, while several different lines of development may be later understood as variants of one common more general approach (e.g., von Davier, 2013, 2014). In essence, any attempt to organize the many different approaches that exist today may lead to simplifications, and potentially omissions of related approaches.

The aim of all these approaches, however, can be summarized as the attempt to provide powerful tools to help researchers learn about how observed behaviors, such as responding to test items, can be used to derive information about generalizable behavioral tendencies.

We begin the chapter with a section on general notation and concepts central to latent class analysis, followed by an introduction of mastery models, ranging from deterministic to probabilistic forms. The ensuing sections cover knowledge state and rule space approaches, which can be viewed as deterministic skill-profile models. The chapter closes with a section on the multiple classification latent class model and the deterministic input noisy and (DINA) model.

1.2 Notation, Log-linear Models, and Latent Class Analysis

This section introduces notation used in subsequent chapters. We use the case of binary observed variables as a standard example but note that all definitions can be directly extended to polytomous nominal or ordinal response variables. Let $\boldsymbol{X} = (X_1, \ \ldots \ , X_K)$ denote K binary (or polytomous) response variables and let $\boldsymbol{x}_n = (x_{n1}, \ \ldots \ , x_{nK})$ denote the observed responses for test takers $n = 1, \ \ldots \ , N$. Let G denote a grouping variable with $g_n \in \{1, \ \ldots \ , M\}$ for all test takers. In the case of discrete mixture (or latent class) models, g_n is unobserved, while for multiple group models, g_n is completely or partially observed (von Davier & Yamamoto, 2004; von Davier & Carstensen, 2007).

The probability of observing $\boldsymbol{x} = (x_1, \ \ldots \ , x_K)$ will be denoted by

$$P\left(\boldsymbol{X} = \boldsymbol{x}\right) = P\left(x_1, \ldots, x_K\right).$$

Obviously, these probabilities are unknown, while we may have some idea which observed variables have higher or lower probability of exhibiting certain values. For cognitive tasks, we may have some idea about the order of items with respect to the likelihood of successful completion, but typically, there is no complete knowledge about the joint distribution of response variables.

The aim of modeling response data is to provide a hypothesis of how this unknown joint distribution can be constructed in a systematic way, either through associations and interactions between observables, or by means of predictors, or through assumed unobserved variables, or a combination of these.

1.2.1 Log-linear Models

One customary way to model the joint distribution of the responses $x_1, \ldots, x_K$ is using log-linear models (e.g., Haberman, 1979; Hagenaars, 1993). Log-linear models can be used with or without assuming latent variables. Log-linear models describe transformed probabilities, using the natural logarithm. We can write

$$ln\,P\,(x_1, \ldots, x_K) = f\,(x_1, \ldots, x_K)\,,$$

where $f(x_1, \ldots, x_K)$ is a function of the observed responses. One possible assumption is that the log of the response probabilities can be expressed as

$$f\,(x_1, \ldots, x_K) = \lambda_0 + \sum_{i=1}^{K} \lambda_{1i} x_i + \sum_{\{i \neq j\}} \lambda_{2ij} x_i x_j + \ldots + \sum_{\{i \neq \ldots \neq k\}} \lambda_{Ki\ldots k} \left[\prod_{v=1}^{K} x_v\right].$$

Log-linear models in the context of CDMs have been discussed for example by von Davier and Yamamoto (2004) and Xu and von Davier (2008) for dichotomous and ordinal skill attribute variables. von Davier (2018) showed how certain log-linear models used in the context of CDMs can be considered generalizations of models recently discussed under the term *network psychometrics* (e.g., Marsman et al., 2018; von Davier, 2018). In the example above, all products of any possible subset of observed variables are included, however, it is customary to also consider models that only include terms up to a certain degree D, assuming for higher degrees $E > D$ that $\lambda_{Ei\ldots,i_E} = 0$.

One central issue when estimating log-linear models for large numbers of observables is that a normalization factor is needed. Since, $1 = \sum_{(x_1,\ldots,x_K)} P\,(x_1, \ldots, x_K) = \sum_{(x_1,\ldots,x_K)} \exp f\,(x_1, \ldots, x_K)$, it follows that

$$\lambda_0 = log\left[\sum_{(x_1,\ldots,x_K)} \exp\left[\sum_{i=1}^{K} \lambda_{1i} x_i + \sum_{\{i \neq j\}} \lambda_{2ij} x_i x_j + \ldots + \sum_{\{i \neq \ldots \neq k\}} \lambda_{Ki\ldots k} \left[\prod_{v=1}^{K} x_v\right]\right]\right].$$

This normalization factor involves a sum over all possible configurations $(x_1, \ldots, x_K)$. For K binary variables, this is a sum involving 2^K terms, for $K = 30$ items this is a summation over 1,073,741,824 terms. von Davier (2018) describes how noise contrastive estimation (e.g., Guttmann & Hyvärinen, 2010, 2012) can be used for estimation of log-linear and network psychometrics models, as well as extended log-linear models for polytomous and dichotomous CDMs.

Log-linear models can be extended by assuming latent variables (Haberman, 1979; Hagenaars, 1993) so that the distribution of observed response variables depends on an unobserved variable α,

$$f\,(x_1, \ldots, x_K|\alpha) = \lambda_0\,(\alpha) + \sum_{i=1}^{K} \lambda_{1i}\,(\alpha)\, x_i + \ldots + \sum_{\{i \neq \ldots \neq k\}} \lambda_{Ki\ldots k}\,(\alpha) \left[\prod_{v=1}^{K} x_v\right]$$

and by definition

$$P(x_1, \ldots, x_K|\alpha) = \exp\left[\lambda_0(\alpha) + \sum_{i=1}^{K} \lambda_{1i}(\alpha) x_i + \ldots + \sum_{\{i\neq\ldots\neq k\}} \lambda_{Ki\ldots k}(\alpha) \left[\prod_{v=1}^{K} x_v\right]\right].$$

If the latent variable is discrete, it can be written as $\alpha \in \{g_1, \ldots, g_G\}$, with G sets of each parameter type $\lambda_{dgi_1\ldots,i_d} = \lambda_{di_1\ldots,i_d}(g)$ for $g = g_1, \ldots, g_G$ *and* $d = 0, \ldots, K$. With this modification, the model becomes more complex. All parameters may depend on some unobserved quantity α, some grouping variable g, or some combination of both.

1.2.2 Latent Class Analysis

Latent Class Analysis (LCA) can be understood as an approach that assumes the dependence of response probabilities on an unobserved discrete variable, which we denote by c. In this sense, LCA is a direct application of the definition of conditional response probabilities, as introduced above. LCA assumes a latent categorical variable that cannot be directly observed. The LCA model equation follows from a set of three assumptions, some of which match assumptions commonly used in other latent variable models:

1. Class dependent response probabilities: For response variables x_i, LCA assumes class specific response probabilities. While there is no direct constraint that imposes

$$P(x_i|c_v) \neq P(x_i|c_w),$$

   it is a prerequisite for class separation so that respondents who are members of different classes $c_v \neq c_w$ can be reliably classified given their observed responses.

2. Local independence: A central assumption is local independence given class membership c,

$$P(x_1, \ldots, x_K|c) = \prod_{i=1}^{K} P(x_i|c).$$

In LCA the class membership variable c is the latent variable that is expected to 'explain' the dependencies between observed responses. Once conditional probabilities are considered, the dependencies between observed variables vanish, under this assumption.

3. Classes are mutually exclusive and exhaustive: For each examinee v there is one, and only one, 'true' latent class membership $c_v \in \{1, \ldots, G\}$. While the latent variable in LCA is nominal, this assumption is analogous to the assumption of a true (but unobserved expected) score in classical test theory (CTT) or a true ability θ in item response theory (IRT).

These three assumptions make the LCA a discrete mixture distribution model, since it follows from this set of assumptions that the marginal probability of a response pattern is given by

$$P(x_1, \ldots, x_K) = \sum_{c=1}^{G} \pi_c P(x_1, \ldots, x_K | c) = \sum_{c=1}^{G} \pi_c \prod_{i=1}^{K} P(x_i | c)$$

with mixing proportions (class sizes) $\pi_c = P(C = c)$. A logarithmic transform following assumption 2 above yields,

$$\ln P(x_1, \ldots, x_K | c) = \sum_{i=1}^{K} \ln P(x_i | c) = \sum_{i=1}^{K} [x_i \ln P(X_i = 1 | c) + (1 - x_i) \ln P(X_i = 0 | c)]$$

and further, using standard rules for the logarithm,

$$\ln P(x_1, \ldots, x_K | c) = \sum_{i=1}^{K} \ln P(X_i = 0 | c) + \sum_{i=1}^{K} x_i \left[\ln \frac{P(X_i = 1 | c)}{P(X_i = 0 | c)} \right].$$

As such, LCA can be understood as a log-linear model without interactions (as local independence is assumed), conditional on a nominal latent variable. This can be seen by setting $\lambda_{1ci} = \left[\ln \frac{P(X_i=1|c)}{P(X_i=0|c)} \right]$ (a term that represents the *log-odds* for item i conditional on class membership c) and $\lambda_{0c} = \sum_{i=1}^{K} \ln P(X_i = 0 | c)$ (an intercept term) and observing that

$$\ln P(x_1, \ldots, x_K | c) = \lambda_{0c} + \sum_{i=1}^{K} x_i \lambda_{1ci}.$$

Note that the log-odds λ_{1ci} and the conditional response probabilities have the following relationship:

$$\frac{\exp(\lambda_{1ci})}{1 + \exp(\lambda_{1ci})} = P(X_i = 1 | c).$$

While the within-class model of LCA is rather restrictive, as independence of all responses is assumed, the LCA is a very flexible model, since the number of

classes C is not specified a priori. Any dependence between observed variables can be modeled by increasing the number of classes, however, identifiability may be an issue (e.g., Goodman, 1974; Allman, Matias, & Rhodes, 2009; Xu, this volume). Therefore, this flexibility is also a weakness of the LCA. With the addition of classes to the model the fit between model predictions and observed data will always improve, which may result in a LCA solution that overfits the observed dependencies. In addition, the increase in number of classes leads to a substantial increase in the number of parameters to be estimated. For additional details on applications of LCA, see the volumes by Langeheine and Rost (1988), Rost and Langeheine (1997), and Hagenaars and McCutcheon (2002), as well as the chapter by Dayton and Macready (2006).

Confirmatory approaches to LCA constrain the number of classes and often also impose inequality or equality constraints on class specific response probabilities (e.g., Croon, 1990). Most DCMs covered in this volume can be written as constrained variants of LCA (von Davier, 2009). Some constrained versions of LCA share many interesting similarities with (M-)IRT models (e.g., Haberman, von Davier, & Lee, 2008) and can be used to replace these models.

1.3 Mastery Models

Mastery models assume a skill domain for which we can sort any person into one of two classes: expert versus novice, master versus non-master, or professional versus amateur. This may not be adequate for most domains, even if there is a distinct 'can do' versus 'cannot do'; there are often gradual differences in the 'can do'. In this section, however, we use this notion of mastery and assume all respondents can be classified into two groups without further distinction.

While these types of distinctions may be oversimplifications, can they still be useful categories to describe how test takers respond to a test? If we consider ideas from developmental psychology (e.g., Piaget, 1950; Wilson, 1989), we find that some things in life are thought of as being acquired or learned in terms of qualitative jumps. We may want to entertain the idea of mastery learning for a while and examine where this leads us in terms of how a latent variable model may represent this concept. For example, young children cannot perform or solve task X until they mature and 'get it', after which the same task becomes quite easy for them.

The mastery-state can be represented by a random variable that takes on two values: '1' = mastery and '0' = non-mastery. Formally, we define a latent variable A, with $a_v \in \{0, 1\}$ for all respondents $v = 1, \ldots, n$, and with

$$a_v = 1 \text{ if person } v \text{ masters the skill of interest}$$

and

$$a_v = 0 \text{ if person } v \text{ does not master the skill.}$$

The two mastery levels are expected to differ with respect to the probabilities of success, just as in assumption 1 presented in the section on LCA above. However, in mastery models, there is an order expectation, or even an order restriction in place: it is expected (and potentially specified directly in the model) that for all response variables the probability of success is larger for masters than for non-masters. More formally,

$$P\left(X_i = 1|a = 1\right) = 1 - s_i \geq g_i = P\left(X_i = 1|a = 0\right)$$

may be assumed for all response variables $X_1, \ldots, X_K$. For each item, there are four probabilities to consider, the conditional probabilities of success and failure under mastery and non-mastery. These are often denoted as follows (e.g., Dayton & Macready, 1977):

- Guessing correctly by non-masters: $g_i = P(X_i = 1 \,|\, a = 0)$
- Incorrect response by non-masters: $1 - g_i = P(X_i = 0 |\, a = 0)$
- Slipping = unexpected incorrect response by masters: $s_i = P(X_i = 0 \,|\, a = 1)$
- Correct response by masters: $1 - s_i = P(X_i = 1 |\, a = 1)$

A variety of constraints on these parameters have been suggested in the literature, some examples are discussed by Macready and Dayton (1977). Nowadays, the term 'slipping' is often used instead of 'unexpected error' while 'guessing' is still in use (Junker & Sijtsma, 2001). Just like LCA, mastery models also assume local independence and that masters and non-masters are mutually exclusive and exhaustive. Based on the equivalency shown in the previous section, a mastery model with two levels can be written either in the form of a 2-class LCA or as a log-linear model with latent variables:

$$P\left(x_1, \ldots, x_K|a\right) = \prod_{i=1}^{K} P\left(X_i = 0|a\right) \left[\frac{P\left(X_i = 1|a\right)}{P\left(X_i = 0|a\right)}\right]^{x_i}$$

and with the definitions above, we have $P\left(X_i{=}1|a\right) {=} (1{-}s_i)^a g_i^{[1-a]}$, and for the complement we have $P\left(X_i = 0|a\right) = s_i^a (1 - g_i)^{[1-a]}$. A logarithmic transformation and insertion of the definitions yields the following:

$$\ln P\left(x_1, \ldots, x_K|a\right) = \sum_{i=1}^{K} \ln s_i^a (1{-}g_i)^{[1-a]} + \sum_{i=1}^{K} x_i \ln \left[\frac{(1{-}s_i)}{s_i}\right]^a \left[\frac{g_i}{(1{-}g_i)}\right]^{[1-a]}.$$

As before, by setting $\sum_{i=1}^{K} \ln s_i^a (1-g_i)^{[1-a]} = \lambda_{0a}$ and $\ln \left[\frac{(1-s_i)}{s_i}\right]^a \left[\frac{g_i}{(1-g_i)}\right]^{[1-a]} = \lambda_{1ai}$, the equivalency of the mastery model to a log-linear model with a binary latent variable is obtained. Note that λ_{1ai} can be written as

$$a \ln \left[\frac{P\left(X_i=1|a=1\right)}{P\left(X_i=0|a=1\right)}\right] + [1-a] \ln \left[\frac{P\left(X_i=1|a=0\right)}{P\left(X_i=0|a=0\right)}\right] = \lambda_{10i} + a\left[\lambda_{11i} - \lambda_{10i}\right],$$

which again contains the log-odds for masters and non-masters, multiplied by the mastery status.

1.4 Located Latent Class or Multi State Mastery Models

The additional model specifications needed to move from LCA, which is characterized by a nominal latent class variable, to located classes are easily introduced. The last section that examined mastery models provides the basis for these developments. For a correct response $x_i = 1$, the term $\lambda_{1ai} = \lambda_{10i} + a[\lambda_{11i} - \lambda_{10i}]$ is part of the sum. This term is linear in the mastery level $a \in \{0, 1\}$ and if $\lambda_{11i} > \lambda_{10i}$ or equivalently, $P(X_i = 1 | a = 1) > P(X_i = 1 | a = 0)$, the term λ_{1ai} is monotone increasing over the (in the case of mastery models: two) ordered mastery levels.

With more than two levels of mastery, for example an ordinal variable that represents non-mastery as zero, but allows multiple levels of mastery represented as successive integer, i.e., $a' \in \{0, 1, 2, \ldots M\}$, a model can be defined as

$$\ln P\left(x_1, \ldots, x_K | a'\right) = \lambda_{0a'} + \sum_{i=1}^{K} x_i \lambda_{1a'i}$$

with

$$\lambda_{1a'i} < \lambda_{1a''i} \text{ for all } a' < a''.$$

This ensures that

$$P\left(X_i = 1 | a'\right) < P\left(X_i = 1 | a''\right) \text{ for all } a' < a''.$$

This produces a monotone increasing sequence of response probabilities over $a' \in \{0, 1, 2, \ldots M\}$. Note, however, that this type of constraints (still) produces a comparably large number of quantities that need to be estimated. However, this model includes equality constraints (e.g., Formann, 1985, 1992) which may be imposed via additional assumptions about how model parameters relate to the ordered levels of mastery. Essentially, each latent class in this model becomes an ordered mastery level, but the distances between classes differ by item i and class

level $a^{'} \in \{0, 1, 2, \ldots M\}$. This model requires $(M + 1)K$ parameters one set of K item parameters for each class. As before, probabilities can be derived using the equivalency

$$P\left(X_i = 1|a'\right) = \frac{\exp\left(\lambda_{1a'i}\right)}{1 + \exp\left(\lambda_{1a'i}\right)}.$$

A more parsimonious model can be implemented by imposing the following constraint

$$\lambda_{1a'i} = \beta_i + \gamma_i \theta_{a'}$$

which requires $2K$ item location β_i and slope parameters γ_i and $M + 1$ ordered class specific locations $\theta_{a'} < \theta_{a''}$ for $a^{'} < a^{''} \in \{0, \ldots, M\}$. With the transformation

$$\beta_i + \gamma_i \theta_{a'} = a\left(\theta - b\right)$$

it can be easily observed that located latent class models define the class specific response probabilities as

$$P\left(X_i = 1|\theta_{a'} = \theta\right) = \frac{\exp\left(a\left(\theta - b\right)\right)}{1 + \exp\left(a\left(\theta - b\right)\right)}$$

which is very similar to IRT (Lord & Novick, 1968), while assuming a discrete latent variable with located latent classes (e.g., Formann, 1992; Haberman et al., 2008).

1.5 Rule Space Methodology and Knowledge Spaces

Rule space (RS; e.g., Tatsuoka, 1983, 1990, 2009) and knowledge spaces (KS; Doignon & Falmagne, 1985, 1998; Albert & Lukas 1999) are independently developed approaches to the question of how the association between performance on heterogeneous tasks and multiple skills can be conceptualized. Much like mastery models, RS and KS assume that a respondent who masters a certain number of skills is on a regular basis capable of solving tasks that require these skills. In contrast to the first generation of mastery models, both RS and KS assume that there are multiple skills to be considered, and that each respondent is characterized by a skill pattern or attribute pattern – or a *knowledge state* – and that every task requires a subset of the skills represented in the skill space of respondents.

Consider an example with two skills, addition and multiplication, ignoring for a moment that there is an additional skill required that tells us in what order these operations have to be executed. If asking examinees to solve tasks of the type

(a) $3 + 4 =?$
(b) $4 * 5 =?$
(c) $3 * 3 + 2 =?$

one could argue that there are four potential groups of test takers. Group 1 does neither master addition nor multiplication and cannot solve any of the task types; Group 2 masters only addition and can solve tasks of type (a) only; Group 3 only masters multiplication (no matter how unlikely that may seem to a math educator) and hence can solve only tasks of type (b); and Group 4 masters both addition and multiplication, and hence can solve tasks of type (a), (b), and (c) on a regular basis.

More formally, for tasks that require a subset of D skills, we can assign to each task $i = 1, \ldots, K$ a vector of skill requirements $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{iD}) \in \{0, 1\}^D$ that indicates which skill (or attribute) is required for that task. The matrix

$$Q = \begin{pmatrix} q_{11} & \cdots & q_{1D} \\ \cdots & \cdots & \cdots \\ q_{K1} & \cdots & q_{KD} \end{pmatrix}$$

is referred to as Q-matrix and represents a hypothesized relationship of how a skill vector (skill state) or attribute pattern $\boldsymbol{a} = (a_1, \ldots, a_D)$ is connected to expected performance on each task. The ideal (the most likely, or expected given a skill pattern) response on item i given can be written as

$$x_i^{[I]}\left(\boldsymbol{q}_i, \boldsymbol{a}\right) = \prod_{d=1}^{D} a_d^{q_{id}} \in \{0, 1\}$$

which equals one if the attribute mastery pattern $\boldsymbol{a}$ matches or exceeds non-zero entries of the skill requirements $\boldsymbol{q}_i$, i.e., if at least all required skills are mastered, and is zero otherwise. The above equation can be applied to all items to construct an ideal response pattern

$$\boldsymbol{x}^{[I]}\left(\boldsymbol{a}\right) = \left(\prod_{d=1}^{D} a_d^{q_{1d}}, \ldots, \prod_{d=1}^{D} a_d^{q_{Kd}}\right)$$

for each attribute mastery pattern $\boldsymbol{a}$**.** The observed response pattern $\boldsymbol{x}_v$ produced by respondent v can then be compared to each of these ideal response vectors, and the closest match determined. This can be done in a variety of ways; for example, Tatsuoka (1983, 1985) discussed methods based on distance measures, but also presents classification based on IRT ability estimates and person fit. von Davier, DiBello, and Yamamoto (2008) provide a summary of the IRT and fit based approach. A simple measure of agreement can be defined as

$$sim\left(\boldsymbol{x}_v, \boldsymbol{a}\right) = \frac{\sum_{i=1}^{K} x_{vi} * x_i^{[I]}\left(\boldsymbol{q}_i, \boldsymbol{a}\right)}{\sqrt{\left(\sum_{i=1}^{K} x_{vi}^2\right)\left(\sum_{i=1}^{K}\left[x_i^{[I]}\left(\boldsymbol{q}_i, \boldsymbol{a}\right)\right]^2\right)}}$$

which equals the cosine similarity of the observed and ideal vectors. The cosine similarity is a correlation related measure commonly used in data mining, machine learning and natural language processing (Tan, Steinbach, & Kumar, 2005). Respondents can be assigned to the attribute pattern that produces the largest similarity measure relative to the observed vector $\boldsymbol{x}_v$.

Tatsuoka's RS has demonstrated its utility in many applications over the years. Recently, the method gained new interest under the name 'attribute hierarchy method' (AHM; Leighton, Gierl, & Hunka, 2004). The authors describe the AHM as being an instantiation of rule space that differs from Tatsuoka's (1983, 1985, 1990, 2009) methodology in that it allows attribute hierarchies. Attribute hierarchies limit the permissible attribute space, as some attributes have to be mastered before other can be mastered, by definition of what a hierarchy encompasses. von Davier and Haberman (2014) show how the assumption of hierarchical attributes restricts the number and type of parameters of diagnostic classification and multiple mastery models.

Both RS and KS were initially conceptualized as deterministic classification approaches. Respondents would be classified according to their similarity to ideal response patterns, regardless of the observation that only very few respondents will produce exactly the 'ideal' patterns that can be expected based on the Q-matrix. Attempts to produce a less deterministic version of these approaches have been made, and Schrepp (2005) describes similarities between KS approaches and latent class analysis. The next section describes models that share many of the features of RS and KS approaches, but provide a structured latent attribute space, and a probabilistic approach to define how multiple mastery levels relate to response probabilities in a systematic way, rather than by means of unstructured class profiles as used in LCA.

1.5.1 Multiple Classification Models and Deterministic Input Noisy and (DINA) Models

Latent class models with multiple latent variables (Haberman, 1979; Haertel, 1989) or multiple classification latent class models (MCLCM; Maris, 1999) extend latent class analysis (LCA) in such a way that multiple nominal or ordinal latent variables can be identified simultaneously. This approach retains the defining properties of LCA, local independence given latent class, assumption of an exhaustive and disjunctive latent classification variable, and distinctness of conditional probabilities across classes.

The MCLCM approach can be viewed as a non-parametric precursor to many of the diagnostic models introduced in subsequent chapters. For a MCLCM with two latent variables $c_1 \in \{0, \ldots, C_1\}$, $c_2 \in \{0, \ldots, C_2\}$ denote the joint distribution of these with π_{c_1,c_2} and define

$$P(x_1, \ldots, x_K) = \sum_{c_1=0}^{C_1} \sum_{c_2=0}^{C_2} \pi_{c_1,c_2} \prod_{i=1}^{K} P(x_i|c_1, c_2).$$

This is a well-defined LCA that can be rewritten as a single latent[1] variable LCA with 'attribute' $\boldsymbol{a} = \{c_1, c_2\}$ and MNCL $= (C_1 + 1)\,(C_2 + 1)$ latent classes representing all possible combinations. However, one may introduce additional structure – constraints on the response probabilities – for the two-variable case to specify whether the conditional probabilities may for some items depend on only one or the other component c_1 or c_2. More specifically, one may assume

$$P\left(x_i|c_1, c_2\right) = P\left(x_i|f_{i1}\left(c_1\right), f_{i2}\left(c_2\right)\right).$$

As a special case with specific relevance to diagnostic models, we will consider the following form of these constraints in the example

$$f_{id}\left(c_d\right) = c_d^{q_{id}}$$

for $d = 1, 2$ and with $q_{i1}, q_{i2} \in \{0, 1\}$.

Basically, if one or the other q_{i_*} is zero, the dependency on that component of the multiple classification LCM variable vanishes from the conditional probability of item response x_i. This is true because

$$c_d^0 = 1$$

for all levels of c_d whenever $q_{i1} = 0$. With this constraint, the conditional probabilities of a response variable may depend on both c_1, c_2 in MNCL levels for some items, on c_1 only in $(C_1 + 1)$ levels for some other items, or on c_2 with $(C_2 + 1)$ levels for a third set of items, or on neither one of them in a fourth group of response variables.

Two additional restrictions lead to the model that is commonly known as the DINA (Deterministic Input, Noisy And) model (Macready & Dayton, 1977; Junker & Sijtsma, 2001). First, all components of the latent skill pattern $\boldsymbol{a}$ are assumed to be binary (and as before, we use a_d for binary attributes, while for nominal classes, we use $c_1, c_2, \ldots$), that is

$$\boldsymbol{a} = (a_1, \ldots, a_D) \in \{0, 1\}^D$$

and for the conditional probabilities we assume

$$P\left(x_i|a_1, \ldots, a_D\right) = P\left(x_i|\prod_{d=1}^{D} a_d^{q_{id}}\right).$$

[1] Class variables are represented as integers, but the use of integers do not imply any ordering here; only equivalence classes are used in the context of LCA.

Note that the conditional probability depends on a binary variable $\xi_{\boldsymbol{a}q_i} = \prod_{d=1}^{D} a_d^{q_{id}} \in \{0, 1\}$ which is a function of the skill pattern $\boldsymbol{a}$ and one row of the Q-matrix, a vector that specifies the skill requirements for a specific item. Just as in the section on mastery models, applying this definition leads to the following expressions:

$$P\left(X_i = 1 | \xi_{\boldsymbol{a}q_i} = 1\right) = 1 - s_i$$

and

$$P\left(X_i = 1 | \xi_{\boldsymbol{a}q_i} = 0\right) = g_i.$$

The DINA model is said to be conjunctive because it reduces the respondent-skill by item-attribute comparison to only two levels $\prod_{d=1}^{D} a_d^{q_{id}} = 1$ or $\prod_{d=1}^{D} a_d^{q_{id}} = 0$. With this, we can write

$$P\left(x_i | a_1, \ldots, a_D\right) = \left[(1 - s_i)^{\xi_{\boldsymbol{a}q_i}} g_i^{1-\xi_{\boldsymbol{a}q_i}}\right]^{x_i} \left[s_i^{\xi_{\boldsymbol{a}q_i}} (1 - g_i)^{1-\xi_{\boldsymbol{a}q_i}}\right]^{1-x_i}.$$

Only those respondents who possess all necessary skills have a "high" probability $1 - s_i$ of solving an item, while respondents who lack at least one of the required skills have a "low" probability g_i —the same "low" probability no matter whether only one or all required skills are not mastered.

Note that the g_i and the s_i denote the item parameters in the DINA model, so that there are two parameters per item in this model. In addition, the skill vectors $a_v = (a_{v1}, \ldots a_{vK})$ are unobserved, so we typically have to assume that the distribution of skills $P\left(A = (a_1, \ldots a_K)\right) = \pi_{(a_1, \ldots a_K)}$ is unknown. Therefore, there are $\|\{0, 1\}^K\| - 1 = 2^K - 1$ independent skill pattern probabilities with $\sum_{(a_1, \ldots a_K)} \pi_{(a_1, \ldots a_K)} = 1.0$ if an unconstrained estimate of the skill distribution is attempted. There may be fewer parameters if a constrained distribution over the skill space (von Davier & Yamamoto, 2004; Xu & von Davier, 2008) is used. For model identification, no constraints are needed on the guessing and slipping parameters (even though it is desirable that $1 - s_i > g_i$ for somewhat sensible results).

While de la Torre (2009) does not make statements about identifiability of the DINA model and the uniqueness of the model parameters, Junker and Sijtsma (2001) discuss (a lack of) empirical identification in the context of their data example used in conjunction with Markov chain Monte Carlo (MCMC) estimation. Haertel (1989) describes identification of latent class skill patterns in the DINA model, and notes that *"it may be impossible to distinguish all these classes empirically using a given set of items. Depending upon the items' skill requirements, latent response patterns for two or more classes may be identical (p.303)."* One of the remedies Haertel (1989) suggests is the combination of two or more latent classes that cannot be distinguished. In subsequent chapters, identifiability of diagnostic models is discussed in more detail (Xu, this volume; Liu & Kang, this

volume; DeCarlo, this volume) and von Davier (2014) provides an example of how the (lack of) empirical identifiability of diagnostic models can be checked.

The DINA model is a very restrictive model as it assumes only two parameters per item, and skill attributes only enter the item functions through conjunction function $\xi_{\boldsymbol{a}q_i} = \prod_{d=1}^{D} a_d^{q_{id}}$. This restricts the probability space so that different attribute mastery patterns, in particular those that are not a perfect match of the Q-matrix for an item, are all mapped onto the same low "guessing" probability. There are several issues with the assumption made in the DINA model. Formally, this assumption is equivalent to assuming a log-linear model (see Eq. 1) in which all parameters are set to zero except the one that parameterizes the highest order interaction term. Additionally, from the point of view of most applications of skills, compensation happens: Multiplication can be replaced by repeated addition, a lack of vocabulary when acquiring a new language, or even learning disabilities can be compensated for (and eventually remedied) by higher general intelligence (e.g., Reis, McGuire, & Neu, 2000), etc. In total darkness, hearing can be used to, admittedly poorly, compensate for lack of vision. For diagnostic models and compensatory and non-compensatory MIRT models, it was found that real data examples are often fit better (in terms of item fit, or overall goodness of fit assessed with information criteria or similar) with additive/compensatory models rather than conjunctive models (de la Torre & Minchen, this volume; von Davier, 2013).

In addition, it was found that the DINA model may be affected by model identification issues. DeCarlo (2011) and Fang, Liu, and Ying (2017) show that the DINA model is not identified unless there are what some may call 'pure' items in the Q-matrix, that is, items that only measure a single attribute. DeCarlo (2011) shows that the DINA model with the Q-matrix provided for the Fraction Subtraction data (Tatsuoka, 1985) is not able to identify all attribute patterns. Fang, Liu, and Ying (2017) provide more general results on the requirements for the Q-matrix. Xu (this volume) and Liu and Kang (this volume) provide further results and more recent examples.

1.6 Summary

The notation and models introduced in this chapter form the basis for many of the subsequent chapters. Most, if not all DCMs can be written as constrained latent class models or alternatively, log-linear models with discrete latent variables.

This introduction does not provide an in-depth coverage of how to evaluate the different approaches. However, all models presented in this volume are approaches that provide marginal probability distributions for multivariate discrete observables. This means that methods from categorical data analysis can be used to compare models and to evaluate model data fit.

While some of the models introduced above may be considered approaches for diagnostic classification and may have been used as such, many more sophisticated

approaches have been developed since, based on these initial modeling attempts. The aim of the current volume is providing a systematic overview of these more recent approaches.

The *Handbook Diagnostic Classification Models* aims at capturing the current state of research and applications in this domain. While a complete overview of this broad area of research would require a multi-volume effort, we tried to capture a collection of major research streams that have been developed over several years and that continue to produce new results.

The first part of the volume covers major developments of diagnostic models in the form of chapters that introduce the models formally, provide information on parameter estimation and on how to test model-data fit, and applications or extensions of the approach.

The second part of the volume describes special topics and applications. Special topics such as Q-matrix issues are covered, including the data driven improvement and construction, as well as issues around model identifiability. The third part presents applications of diagnostic models, as these are a centerpiece to reasons why not only methodologists but also applied researchers may want to study the volume. These applications show how diagnostic models can be used to derive more fine-grained information about respondents than what traditional methods such as CTT or IRT can provide.

The fourth part of the book includes a range of available software packages, including the use of general purpose statistical software, specialized add-on packages, and available stand-alone software for estimation and testing of CDMs.

In many cases, latent class analysis, customary IRT, and other latent variable models can directly be considered alternatives to diagnostic models, as these are often more parsimonious (in the case of IRT) or do not make as strong (parametric) assumptions about the latent structures and how these structures are related to the conditional response probabilities in the levels of the latent variables. Standard procedure should therefore be used as a comparison of more complex modeling approaches with customary standard examples of latent variable models such as IRT or LCA. Such a practice will ensure that researchers can compare their findings to those obtained from less complex models to check whether the increased model complexity provides added value, through improved model-data fit, and by means of more useful derived quantities such as estimated mastery states.

References

Albert, D., & Lukas, J. (1999). *Knowledge spaces: Theories, empirical research and applications*. Mahwah, NJ: Erlbaum.

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics, 37*(6A), 3099–3132. https://doi.org/10.1214/09-AOS689

Croon, M. A. (1990). Latent class analysis with ordered classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171–192.

Dayton, C. M., & Macready, G. (2006). Latent class analysis in psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Elsevier Science B.V.: Amsterdam, The Netherlands.

de la Torre, J. (2009, March). DINA Model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

Decarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26.

Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies, 23*, 175–196.

Doignon, J. P., & Falmagne, J. C. (1998). *Knowledge spaces*. Berlin, Germany: Springer.

Fang, G, Liu, J., & Ying, Z. (2017). *On the identifiability of diagnostic classification models* (Submitted on 5 Jun 2017) arXiv:1706.01240 [math.ST].

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38*, 87–111.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*(418), 476–486. https://doi.org/10.1080/01621459.1992.10475229

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.

Gutmann, M. U., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Gutmann, M. U., & Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research, 13*, 307–361.

Haberman, S. J. (1979). *Qualitative data analysis: Vol. 2. New developments*. New York, NY: Academic.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. RR-08-45. ETS Research Report. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.

Hagenaars, J. A. (1993). Loglinear models with latent variables. In *Sage University Papers* (Vol. 94). Newbury Park, CA: Sage Publications.

Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis models*. Cambridge, UK: Cambridge University Press.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models*. New York, NY: Plenum Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin Company.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., … Maris, G. (2018). An introduction to network psychometrics: Relating ising network models to item response theory models. *Multivariate Behavioral Research, 53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

Piaget, J. (1950). *The psychology of intelligence* (M. Piercy, Trans.). London, UK: Lowe & Brydone (Original work published 1947).

Reis, S. M., McGuire, J. M., & Neu, T. W. (2000). Compensation strategies used by high ability students with learning disabilities. *The Gifted Child Quarterly, 44*, 123–134.

Rost, J., & Langeheine, R. (Eds.). (1997). *Applications of latent trait and latent class models in the social sciences*. Waxmann Publishers.

Schrepp, M. (2005). About the connection between knowledge structures and latent class models. *Methodology, 1*, 92–102.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA: Addison-Wesley.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55–73. https://doi.org/10.2307/1164930. https://www.jstor.org/stable/1164930

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York, NY: Routledge.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement, 7*, 67–74.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model - two variants of a model equivalency. *BJMSP, 67*, 49–71. http://onlinelibrary.wiley.com/doi/10.1111/bmsp.12003/abstract

von Davier, M. (2014*). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (ETS research report series). http://onlinelibrary.wiley.com/doi/10.1002/ets2.12043/abstract

von Davier, M. (2018). Diagnosing diagnostic models: From von Neumann's elephant to model equivalencies and network psychometrics. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 59–70. https://doi.org/10.1080/15366367.2018.1436827

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*(6), 389–406.

von Davier, M., & Carstensen, C. H. (2007). *Multivariate and mixture distribution Rasch models–extensions and applications*. (Editor of the volume). New York: Springer. ISBN: 978-0387-32916-1.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'Diagnostic' classification models – A commentary. *Psychometrika, 79*, 340. https://doi.org/10.1007/s11336-013-9363-z

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Chapter 7: Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.) *Assessment of competencies in educational contexts* (pp. 151–176). Hogrefe & Huber Publishers.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Xu, X., & von Davier, M. (2008). *Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model*. RR-08-35, ETS Research Report.

Part I
Approaches to Cognitive Diagnosis

Chapter 2
Nonparametric Item Response Theory and Mokken Scale Analysis, with Relations to Latent Class Models and Cognitive Diagnostic Models

L. Andries van der Ark, Gina Rossi, and Klaas Sijtsma

Abstract As the focus of this chapter, we discuss nonparametric item response theory for ordinal person scales, specifically the monotone homogeneity model and Mokken scale analysis, which is the data-analysis procedure used for investigating the compliance between the monotone homogeneity model and data. Next, we discuss the unrestricted latent class model as an even more liberal model for investigating the scalability of a set of items, producing nominal scales, but we also discuss an ordered latent class model that one can use to investigate assumptions about item response functions in the monotone homogeneity model and other nonparametric item response models. Finally, we discuss cognitive diagnostic models, which are the core of this volume, and which are a further deepening of latent class models, providing diagnostic information about the people who responded to a set of items. A data analysis example, using item scores of 1210 respondents on 44 items from the Millon Clinical Multiaxial Inventory III, demonstrates how the monotone homogeneity model, the latent class model, and two cognitive diagnostic models can be used jointly to understand one's data.

L. A. van der Ark (✉)
Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands
e-mail: L.A.vanderArk@uva.nl

G. Rossi
Research Group Personality and Psychopathology, Vrije Universiteit Brussel, Brussels, Belgium
e-mail: grossi@vub.be

K. Sijtsma
Department of Methodology and Statistics, TSB, Tilburg University, Tilburg, The Netherlands
e-mail: k.sijtsma@tilburguniversity.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_2

2.1 Introduction

Nonparametric item response theory (IRT; Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & van der Ark, 2017; van Schuur, 2011), which is the focus of this chapter, is a set of psychometric measurement models implying ordinal person measurement using the sum score on a set of items. The sum score provides a summary of the ability or the trait the items measure but does not inform us about sub-attributes needed for one or more subsets of items in the test. Latent class models (LCMs) aim at classifying persons in unordered or nominal classes based on the set of scores on the items that comprise the test (Hagenaars & McCutcheon, 2002; Heinen, 1996); in LCMs the sum score does not play a role. Although parametric LCMs exist (e.g., Goodman, 1974; Formann & Kohlmann, 2002), the typical LCM is nonparametric. Nonparametric IRT and LCMs may appear to be different, but Croon (1990) and Vermunt (2001) showed how imposing an ordering of the latent classes renders LCM analysis suitable for assessing the fit of a nonparametric IRT model to the data. This application identifies an interesting link between LCMs and nonparametric IRT. Haertel (1989) argued that LCMs are stepping-stones to what later became known as cognitive diagnostic models (CDMs; Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010; von Davier, 2010, 2014). CDMs constitute the core of this volume. The models classify persons based on a set of skills, abilities or attributes the researcher hypothesizes persons need to solve a set of items. In this sense, CDMs provide information about persons' proficiency to solve particular sets of items that is finer-grained than the summary sum score that nonparametric IRT models provide. Junker and Sijtsma (2001) demonstrated how nonparametric IRT and CDMs are related.

The three types of models—nonparametric IRT models, LCMs, and CDMs—have in common that they rely on assumptions about the data that are sufficiently strong to classify or order persons. On the other hand, the assumptions are not so demanding that they all too easily lead to the rejection of items that may not satisfy stronger models but contribute to reliable person classification or person ordering. In this sense, the models are "item-preserving", asking as little as possible from the data and still being able to measure people's attributes at the nominal and ordinal levels (Michell, 1999; Stout, 2002). Although there is much to say about the relationships between the three types of models and much more work that remains to be done to further unravel these relationships, given the goal of this volume we will only briefly discuss the models' assumptions and main ideas. We focus on nonparametric IRT, specifically the version Mokken (1971) introduced and many other researchers further developed. We conclude the chapter discussing a real-data example of a Mokken scale analysis, with brief reference to LCM and CDM data analysis.

2.2 Three Types of Models, Their Properties, and Their Relations

We focus on tests and questionnaires that use a set of K items to measure an attribute, such as a cognitive ability or a personality trait (psychology), an educational achievement and skills (educational measurement), quality of life or pain experience (health sciences), an attitude (sociology) or an opinion (political science). The measurement of a certain type of attributes is not the privilege of a particular research area, hence educational measurement may also measure cognitive abilities (e.g., verbal ability), psychology may also measure attitudes (e.g., towards significant others), health science may also measure personality traits (e.g., introversion), et cetera. Random variable X_k ($k = 1, \ldots, K$) represents the score on item k, and attains ordered scores $x_k = 0, \ldots, m_k$. For simplicity, we assume that within one measurement instrument all items are scored similarly, so that $m_k = m$. Items are often scored dichotomously, for example, incorrect/correct, no/yes, or disagree/agree, in which case $x = 0, 1$. The test score or the sum score summarizes the performance on the K items, $X_+ = \sum_{k=1}^{K} X_k$.

Nonparametric IRT models have in common that they use X_+ to order persons on a scale for the attribute. Each of the models does this by ordinally restricting the relation between the score on an item and the scale of measurement represented by one or more latent variables, but without the use of a parametric function such as the normal ogive or the logistic; see van der Linden (2016) for examples. Mokken (1971) considered the use of IRT models based on parametric functions for the relation between the item score and the latent variable, called item response functions (IRFs), for short, prohibitive of successful measurement of attributes for which foundational theory often was absent or poorly developed, and proposed his nonparametric IRT models (also, see Sijtsma & Molenaar, 2016). Nonparametric IRT models differ with respect to the assumptions they posit to describe the structure of the data, such that they imply an ordinal person scale. Stout (1990, 2002) developed assumptions that were as weak as possible, that is, imposing as few restrictions as possible on the data, and still enabling the ordering of persons. Ramsay (1991, 2016) used kernel smoothing and spline regression to arrive at an ordinal scale for person measurement. Holland and Rosenbaum (1986) derived a broad class of what one might call nonparametric IRT models and studied the mathematical properties of these models. Other work is due to, for example, Junker (1993), Douglas (2001), and Karabatsos and Sheu (2004), and recent work is due to, for example, Straat, van der Ark, and Sijtsma (2013), Tijmstra, Hessen, van der Heijden, and Sijtsma (2013), Ellis (2014) and Brusco, Köhn, and Steinley (2015). Each of the nonparametric approaches has their merits, but in this chapter, we focus on Mokken's approach and present the state of the art of this line of research.

2.2.1 Monotone Homogeneity Model

Mokken's model of monotone homogeneity (Mokken, 1971, pp. 115–169) for ordering persons using the sum score X_+, is based on three assumptions:

1. *Unidimensionality (UD)*. One latent variable denoted Θ stands for the attribute the K items measure.
2. *Monotonicity (M)*. The probability of obtaining a score of at least x on item k, $X_k \geq x$, increases or remains constant but cannot decrease as latent variable Θ increases: $P(X_k \geq x|\,\Theta)$ is non-decreasing in Θ, for $x = 1, \ldots, m$, while $P(X_k \geq 0|\,\Theta) = 1$ by definition; hence, it is uninformative about the relation between the item score and the latent variable. Conditional probability $P(X_k \geq x|\,\Theta)$ is called the item step response function (ISRF), and Fig. 2.1 shows an example of two items each with $x = 0, \ldots, 3$; hence, both items have three IRSFs for $x = 1, 2, 3$. For dichotomous items, $P(X_k = 1|\,\Theta)$ is non-decreasing in Θ, while $P(X_k = 0|\,\Theta) = 1 - P(X_k = 1|\,\Theta)$ and thus is uninformative when $P(X_k = 1|\,\Theta)$ is known.
3. *Local Independence (LI)*. When latent variable Θ explains the relations between the K items and no other latent variables or observed variables such as covariates explain the relations between at least two of the other items, conditioning on Θ renders the K-variate distribution of the item scores equal to the product of the K marginal item-score distributions. This property is called local independence (LI),

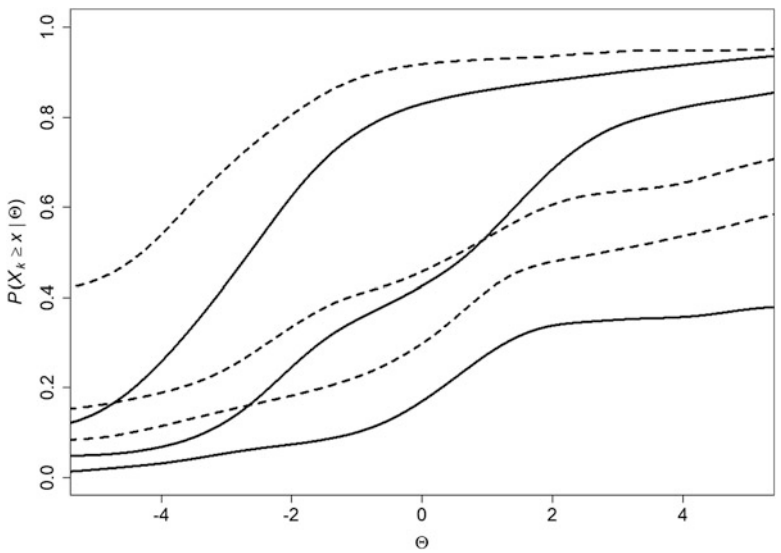


Fig. 2.1 Three nondecreasing item step response functions for two items; solid lines for one item, dashed lines for the other item

$$P(X_1 = x_1, \ldots, X_K = x_K|\Theta) = \prod_{k=1}^{K} P(X_k = x_k|\Theta)\,. \tag{2.1}$$

LI implies weak LI, meaning that the conditional covariance between any pair of items j and k equals 0 (Stout, 1990); that is, Eq. (2.1) implies $\sigma(X_j, X_k|\,\Theta) = 0$.

The three assumptions UD, M, and LI together do not enable direct estimation of Θ. However, for dichotomous items, the monotone homogeneity model implies that sum score X_+ orders persons stochastically on latent variable Θ; that is, for any value θ of Θ and any pair of values x_{+a} and x_{+b} of X_+ such that $0 \leq x_{+a} < x_{+b} \leq K$,

$$P(\Theta > \theta|X_+ = x_{+a}) \leq P(\Theta > \theta|X_+ = x_{+b}) \tag{2.2}$$

(Grayson, 1988). Hemker, Sijtsma, Molenaar and Junker (1997) called the property in Eq. (2.2) stochastic ordering of the latent trait by means of the sum score (SOL). SOL is important, because it shows that if one orders persons by their sum scores, they are also stochastically ordered by Θ. Hence, the monotone homogeneity model implies that Θ is an ordinal scale for persons, and that one can use X_+ to order persons on this scale. An interesting and insightful implication of Eq. (2.2) pertains to conditional expectations; that is,

$$\mathcal{E}(\Theta|X_+ = x_{+a}) \leq \mathcal{E}(\Theta|X_+ = x_{+b})\,, \tag{2.3}$$

meaning that sum score X_+ orders persons by expectation, that is, subgroups characterized by increasing mean Θs. Obviously, because random error affects measurement, one cannot unambiguously conclude at the level of individuals n_1 and n_2 that when one observes $x_{+n_1} < x_{+n_2}$, then $\theta_{n_1} \leq \theta_{n_2}$. Random measurement error may obscure the real ordering $\theta_{n_1} > \theta_{n_2}$, but for homogeneous sum-score groups, identified by θ_a and θ_b, Eq. (2.3) assures an ordering by mean Θs.

For polytomous items, Hemker et al. (1997) showed that the monotone homogeneity model does not imply SOL, hence the model does not produce an ordinal person scale; also see Hemker, van der Ark and Sijtsma (2001). They further showed that among the class of parametric IRT models for polytomous items, only the parametric partial credit model (Masters, 1982), and its special cases such as the rating scale model (Andrich, 1978), implies SOL. Other well-known polytomous IRT models, such as the generalized partial credit model (Muraki, 1992) and the graded response model (Samejima, 1969) do not possess the SOL property. This result suggested that sum score X_+ may not be useful for ordering people on Θ in most polytomous IRT models, but one may also argue that this is not a problem because such models allow the assumption of a real-valued variable Θ and its estimation, thus enabling person measurement using Θ and without X_+.

However, two additional results provide hope for X_+. First, based on multiple simulated latent variable distributions and ISRFs, van der Ark (2005) found that, as a rule, X_+ correctly orders people on Θ. When reversals with respect to Θ happen, they mostly concern X_+ values that are close, often just one unit apart. When K and

m decrease, and ISRFs are more similar, the proportions of person pairs showing ordering violations decrease. Reversely, short tests containing items with, say, five ordered scores, and ISRFs that vary greatly produced more ordering violations than long tests containing items with, say, three ordered scores and ISRFs that are similar.

One may argue that SOL must hold for models to justify ordinal person scales and that failure of SOL is unacceptable, thus rendering X_+ useless as a statistic that orders persons on Θ. Two arguments mitigate this position. One argument is that for realistic K, say, $K \leq 40$, measurement of psychological attributes suffers greatly from measurement error in any measurement value including X_+, which probably causes many accidental ordering reversals that cannot be distinguished from systematic reversals caused by failure of SOL (Eq. 2.2), and may even have a greater impact on ordering. Hence, irrespective of whether the IRT model implies SOL, random measurement error probably overshadows the damage a violation of SOL does to person ordering. The other argument concerns an ordering property van der Ark and Bergsma (2010) called weak SOL, which is an implication of Eq. (2.2), SOL, and which the authors proved holds for all polytomous IRT models assuming UD, M, and LI. Hence, weak SOL provides some relief when a model fails to imply the stronger SOL.

Weak SOL is defined as follows. Assume polytomous items, a fixed integer value x_{+c}, such that $1 \leq x_{+c} \leq Km$, and assume UD, M, and LI; then weak SOL means

$$P(\Theta > \theta | X_+ < x_{+c}) \leq P(\Theta > \theta | X_+ \geq x_{+c}) . \tag{2.4}$$

It may be noted that for $x_{+c} < 1$ and $x_{+c} > Km$, Eq. (2.4) is undefined. Weak SOL does not imply Eq. (2.2), SOL, and is thus a weaker ordering property; see van der Ark and Bergsma (2010) for a computational example showing that SOL can fail while weak SOL is satisfied. SOL Eq. (2.2) implies Eq. (2.3) concerning expected values, and weak SOL Eq. (2.4) implies a similar ordering property concerning expected values,

$$\mathcal{E}(\Theta | X_+ < x_{+c}) \leq \mathcal{E}(\Theta | X_+ \geq x_{+c}) . \tag{2.5}$$

Equation (2.5) shows that, for $x_{+c} = 1, \ldots, Km$, weak SOL enables the ordering of two groups defined by $X_+ < x_{+c}$ and $X_+ \geq x_{+c}$ on Θ. For example, if one selects the 20% best students from a sample using the test scores as a selection criterion, then weak SOL implies that the expected Θ value for the selected respondents is at least as high as the expected Θ value for the respondents who were not selected. However, weak SOL does not allow the ordering of more than two mutually exclusive groups (e.g., three groups defined by $X_+ < x_{+c}$, $x_{+c} \leq X_+ < x_{+c} + u$, and $X_+ \geq x_{+c} + u$; $u \in \{1, 2, \ldots, Km - x_{+c} - 1\}$); two non-exclusive groups (e.g., two groups defined by $X_+ < x_{+c} + u$ and $X_+ \geq x_{+c}$) or two non-exhaustive groups (e.g., two groups defined by $X_+ < x_{+c}$ and $X_+ \geq x_{+c} + u$; $u \geq 1$) (van der Ark & Bergsma, 2010, proposition; also see Douglas, Fienberg, Lee, Sampson, & Whitaker, 1991). One can check that for three persons, n_1, n_2, and n_3, with $x_{+n_1} < x_{+n_2}$, $x_{+n_2} < x_{+n_3}$, and consequently, $x_{+n_1} < x_{+n_3}$, for each person pair

one can always find cut scores x_{+c}, such that for each person pair weak SOL implies a pairwise ordering, but one can also check that an ordering of all three persons is not possible because three subgroups based on two cut scores always overlap.

We conclude that, based on theoretical considerations, the monotone homogeneity model for polytomous items only allows pairwise person ordering but not complete person ordering. Van der Ark's (2005) computational results give us enough confidence to use sum scores X_+ to order people on Θ in practical applications of tests and questionnaires.

2.2.2 *Latent Class Model*

The LCM assumes a discrete latent variable but refrains from specifying its dimensionality, thus defining unordered measurement values that represent latent classes. The model can be used to identify subgroups characterized by the same pattern of scores on K observables, and here we assume that N persons provide discrete integer scores on K items, just as with the discussion of the monotone homogeneity model. Like the monotone homogeneity model, the LCM assumes LI, but now given class membership. Let latent variable Φ have W discrete values denoted $w = 1, \ldots, W$; then LI is defined as in Eq. (2.1), but for $\Phi = w$.

Only assuming a discrete latent variable and LI would provide too little structure to restrict the probability structure governing the data (Suppes & Zanotti, 1981). Assuming every observation falls into one of just a few latent classes, W, restricts the LCM and makes it a feasible approach. One can write the probability of a particular pattern of item scores, denoted $\mathbf{X} = (X_1, \ldots, X_K)$ with realization $\mathbf{x} = (x_1, \ldots, x_K)$, and being in class $\Phi = w$, $P(\mathbf{X} = \mathbf{x} \wedge \Phi = w)$, as the product of the probability of being in class $\Phi = w$, $P(\Phi = w)$, and the probability of obtaining score pattern $\mathbf{X} = \mathbf{x}$ conditional on class membership, $P(\mathbf{X} = \mathbf{x} | \Phi = w)$; that is,

$$P(\mathbf{X} = \mathbf{x} \wedge \Phi = w) = P(\Phi = w)\, P(\mathbf{X} = \mathbf{x} | \Phi = w)\,. \tag{2.6}$$

Applying LI to conditional probability $P(\mathbf{X} = \mathbf{x} | \Phi = w)$ in Eq. (2.6), and summation across discrete classes yields the foundational equation of LCM analysis,

$$P(\mathbf{X} = \mathbf{x}) = \sum_{w=1}^{W} P(\Phi = w) \prod_{j=1}^{J} P\big(X_j = x_j | \Phi = w\big)\,. \tag{2.7}$$

The discrete IRFs (rather, for each item, W separate response probabilities), $P(X_k = x_k | \Phi = w)$, appear on the right-hand side in Eq. (2.7). The model is typically used in an exploratory fashion, because the classes are unknown, hence latent, and the quest is for the number W that explains the data structure best using the model in Eq. (2.6). Conducting an analysis involves estimating the class weights, $P(\Phi = w)$, for each w, and the item response probabilities, $P(X_k = x_k | \Phi = w)$, for each k and w.

One can use these probabilities in conjunction with Bayes theorem to assign people to the class with best fit; that is, for person n, one finds the class w for which

$$P(\Phi = w|\mathbf{X}_n = \mathbf{x}_n) = \frac{P(\mathbf{X}_n = \mathbf{x}_n|\Phi = w)\, P(\Phi = w)}{\sum_{w=1}^{W} P(\Phi = w) \prod_{k=1}^{K} P(X_k = x_k|\Phi = w)}, \quad (2.8)$$

is maximized and assigns person n to this class. This application of the model assigns individuals to latent classes, thus producing a nominal scale. Another application of LCM analysis is to identify latent classes in an effort to understand the structure of the data. Different applications use the LCM to impute scores (Vermunt, van Ginkel, van der Ark, & Sijtsma, 2008), to model population ability distributions (Wetzel, Xu, & von Davier, 2015), to smooth large sparse contingency tables (Linzer, 2011), and to estimate the reliability of sum scores on tests (van der Ark, van der Palm, & Sijtsma, 2011) and of the scores on individual items (Zijlmans, van der Ark, Tijmstra, & Sijtsma, 2018).

LCMs have been extended with explanatory structures, such as regression models, multilevel models and factor models (Hagenaars & McCutcheon, 2002), but also IRT models such as the partial credit model (e.g., Bouwmeester, Vermunt, & Sijtsma, 2007), and led to numerous applications in a variety of research areas. In an effort to tie the LCM to the monotone homogeneity model, we briefly focus on Ligtvoet and Vermunt (2012; also, see Croon 1990, 1991; Hoijtink & Molenaar, 1997; van Onna, 2002; Vermunt, 2001) who used ordered LCM analysis to investigate assumption M of the monotone homogeneity model.

The unconstrained LCM (Eq. 2.7) is typically estimated using an EM algorithm, but can be estimated using a Gibbs sampler. Both methods yield estimates for the class weights $P(\Phi = w)$ and the item-response probabilities $P(X_k = x_k|\,\Phi = w)$. Ligtvoet and Vermunt (2012) explain how to use the LCM to test assumption M of the monotone homogeneity model by rephrasing that assumption as follows. Replace continuous latent variable Θ with discrete latent variable $\Phi = w$, $w = 1, \ldots, W$, and define the expectation

$$\mathcal{E}(X_k|\Phi = w) = \sum_{x=1}^{m} x \cdot P(X_k = x|\Phi = w)\,, \quad (2.9)$$

(Sijtsma & Hemker, 1998). We assume that $\mathcal{E}(X_k|w)$ is non-decreasing in Φ. The conditional expected item score, $\mathcal{E}(X_k|\Phi = w)$, summarizes the m item step response functions, $P(X_k \geq x|\,\Phi = w)$, for each item, while losing information present at the lower aggregation level, but simplifying the investigation of assumption M. Because for one item, conditional probabilities are dependent, in the Gibbs sampler, investigating assumption M by means of $\mathcal{E}(X_k|\Phi = w)$ entails sampling transformations of conditional probabilities, $P(X_k \geq x|\,\Phi = w)$, that are independent of one another, and together satisfy assumption M at the higher aggregation level of conditional expected item scores. Parameter estimates can be generated after convergence of the algorithm from the posterior distributions of the parameters.

A standard goodness of fit statistic is available for assessing the overall fit of the constrained LCM relative to competing models, and specialized fit statistics assess the fit of individual items. A model fitting strategy first entails choosing a value for W, the number of latent classes based on the best overall fit, and in the second analysis round determining for which items assumption M is satisfied. This is done by comparing the fit of the constrained W-class LCM to the unconstrained W-class LCM. The constrained model fits worse by definition but if the discrepancy between models is large, item fit statistics may be used to suggest which badly fitting items should be removed to improve the overall fit (rather than removing the item, the constraint M on the item is removed). Because inactivating constraint M for one or two items probably affects overall fit, the first analysis round is redone and depending on the fit, other items may be flagged for removal. After some iterations, the result is a W-class LCM for K^* items ($K^* \leq K$) for which assumption M holds, if applicable.

2.2.3 Cognitive Diagnostic Model

CDMs allow the assessment of mastery or non-mastery of multiple attributes or skills needed to solve items. CDMs have been applied most frequently to cognitive items in an educational context, but applications are also known to the evaluation and diagnosis of pathological gambling (Templin & Henson, 2006) and the understanding and scoring of situational judgment tests (Sorrel et al., 2016). Several models are available that have in common that they assume that the solution of an item depends on the availability of a set of latent attributes, and for different items different albeit partly overlapping subsets of latent attributes may be required. The most important difference between the two models we discuss here is that one is conjunctive or non-compensatory, and the other disjunctive or compensatory. Conjunctive models assume the tested person needs to master all attributes necessary to solve an item, and non-mastery of a required attribute cannot be compensated by mastery of another attribute. Disjunctive models require a subset of attributes to solve an item but not all attributes, and non-mastery of one or more attributes can be compensated by mastery of others. The models have in common that they compare a person's ideal item-score pattern with her observed item-score pattern, and posit an IRF that relates the two patterns and allows persons lacking attributes the item requires for its solution to solve it correctly (guessing), and likewise persons in possession of the necessary attributes to fail the item (slipping). Von Davier (2014) studied the relationship between non-compensatory and compensatory models, and showed mathematically how their differences may be understood in more detail.

Some notation needed for both models is the following. Let $\mathbf{X}_n = (X_{n1}, \ldots, X_{nK})$ be the vector of binary (incorrect/correct) scores for person n's responses on the K items, and let $\mathbf{A}_n = (A_{n1}, \ldots, A_{nD})$ be the binary latent attribute vector, where $A_{nd} = 1$ means that person n possesses attribute d, and $A_{nd} = 0$ that the person does not possess the attribute. Another building block of CDMs is the Q-matrix, which contains for each item (rows) and attribute (columns) elements q_{kd} indicating whether item k requires attribute d for its solution ($q_{kd} = 1$) or not ($q_{kd} = 0$). We discuss two ways in which $\mathbf{A}_n$ and matrix $\mathbf{Q}$ can be combined to produce a latent item-score vector, $\boldsymbol{\Xi}_n = (\Xi_{n1}, \ldots, \Xi_{nK})$, with realization $(\xi_{n1}, \ldots, \xi_{nK})$. $\boldsymbol{\Xi}_n$ may be considered ideal and can be compared to the observed item-score vector, $\mathbf{X}_n$, to determine how well the model fits the data. The two models we discuss are representatives of conjunctive and disjunctive approaches, and we discuss the models for didactical reasons, but notice that other, more flexible models are available. These alternative models are discussed elsewhere in this book.

The deterministic inputs, noisy "and" gate model (DINA; Junker & Sijtsma, 2001; Haertel, 1989; Macready & Dayton, 1977) is a conjunctive model that defines binary latent response variable, Ξ_{nk}, to indicate whether person n possesses all the attributes needed for solving item k ($\Xi_{nk} = 1$) or not ($\Xi_{nk} = 0$). The ideal responses are defined as

$$\Xi_{nk} = \prod_{d=1}^{D} A_{nd}^{q_{kd}}. \tag{2.10}$$

Equation (2.10) shows that if item k requires an attribute d (i.e., $q_{kd} = 1$) that person n lacks (i.e., $A_{nd} = 0$), then $A_{nd}^{q_{nd}} = 0^1 = 0$, yielding $\Xi_{nk} = 0$; otherwise, $A_{nd}^{q_{nd}} = 1$, and only if all power terms equal 1 we obtain $\Xi_{nk} = 1$. The IRFs relate the ideal latent item-score vector to the fallible real-data item-score vector by allowing masters ($\Xi_{nk} = 1$) to fail an item accidentally, called slipping, and quantified by the slipping parameter,

$$s_k = P(X_{nk} = 0 | \Xi_{nk} = 1), \tag{2.11}$$

and non-masters ($\Xi_{nk} = 0$) to succeed accidentally, quantified by the guessing parameter,

$$g_k = P(X_{nk} = 1 | \Xi_{nk} = 0). \tag{2.12}$$

Using the definitions in Eqs. (2.10), (2.11), and (2.12), the IRF of the DINA model is defined as

$$P(X_{nk} = 1 | \mathbf{A}_n, s_k, g_k) = (1 - s_k)^{\xi_{nk}} g_k{}^{1-\xi_{nk}}. \tag{2.13}$$

Equation (2.11) shows that for non-masters ($\xi_{nk} = 0$), we have $P(X_{nk} = 1 | \mathbf{A}_n, s_k, g_k) = g_k$ and for masters ($\xi_{nk} = 1$), we have $P(X_{nk} = 1 | \mathbf{A}_n, s_k, g_k) = 1 - s_k$.

Hence, the class of non-masters has a probability at the guessing level to solve the item correctly, and the class of masters has a probability reflecting non-slipping or, indeed, mastery. A feature of the IRF in Eq. (2.13) is that it is coordinate-wise monotone in $\mathbf{A}_n$ if and only if $1 - s_k > g_k$. One can check this monotonicity property by checking that changing zeroes in $\mathbf{A}_n$ in ones can change $\xi_{nk} = 0$ into $\xi_{nk} = 1$, but not vice versa; hence, by adding attributes, a non-master can become a master, but this makes sense only if scoring $X_{ij} = 1$ becomes more likely, i.e., if $1 - s_k > g_k$.

We briefly consider the disjunctive deterministic input, noisy "or" gate model (DINO; Templin & Henson, 2006) to illustrate a disjunctive process model. The DINO model assumes that the person needs to master only one attribute, A_{nd}, and the latent response variable is defined as

$$\Psi_{nk} = 1 - \prod_{d=1}^{D} (1 - A_{nd})^{q_{kd}}. \tag{2.14}$$

From Eq. (2.14) it can be seen that the combination of the item requiring an attribute that the person masters ($A_{nd} = q_{kd} = 1$), is the only combination that produces $(1 - A_{nd})^{q_{kd}} = 0$, hence a product equal to 0 and latent response, $\Psi_{nk} = 1$. Thus, one needs to master at least one attribute necessary for item k to produce a latent response $\Psi_{nk} = 1$. Several authors have suggested flexible frameworks that include the DINA and DINO models and several other CDMs (e.g., de la Torre, 2011; von Davier, 2008). This volume witnesses the wealth of CDMs and we therefore refrain from further discussion, except for two notes.

First, the joint distribution of the data conditional on the latent variables, here the D binary attributes, is the product of the conditional distributions of the item scores; that is, LI is assumed,

$$P(\mathbf{X}_n|\mathbf{a}_n) = \prod_{k=1}^{K} P(X_{nk}|\mathbf{a}_n)\,. \tag{2.15}$$

Also assuming that the data records of different persons are independent, the conditional likelihood of the data matrix $\mathbf{X}$ is written as

$$L(\mathbf{X}|\mathbf{a}) = \prod_{n=1}^{N} L(\mathbf{X}_n|\mathbf{a}_n)\,. \tag{2.16}$$

This joint likelihood can be maximized for the parameters $\mathbf{g} = (g_1, \ \ldots \ , g_K)$ and $\mathbf{s} = (s_1, \ \ldots \ , s_K)$, but because they are known to have unfavorable statistical properties, alternatively one rather uses the marginal likelihood approach,

$$L(\mathbf{X}) = \prod_{n=1}^{N} L(\mathbf{X}_n) = \prod_{n=1}^{N} \sum_{h=1}^{H} L(\mathbf{X}_n|\mathbf{a}_h)\, P(\mathbf{a}_h)\,, \tag{2.17}$$

where the number of possible skills patterns equals $H = 2^D$. The latent class structure is apparent on the right-hand side of Eq. (2.17). De la Torre (2009) discussed two estimation algorithms based on Eq. (2.17). One estimation method uses EM, which is labor-intensive due to the huge number of different latent attribute

vectors $\mathbf{A}_h$. The other estimation method avoids this problem by assuming that the elements of vector $\mathbf{A}$ are locally independent given a continuous higher-order latent variable Θ, having the structure of Eq. (2.1),

$$P(\mathbf{A}|\theta) = \prod_{d=1}^{D} P(a_d|\theta)\,, \tag{2.18}$$

and $P(a_d|\,\theta)$ is modeled as a two-parameter logistic model,

$$P(a_d|\theta) = \frac{\exp(\lambda_{0d} + \lambda_1\theta)}{1 + \exp(\lambda_{0d} + \lambda_1\theta)}, \tag{2.19}$$

in which λ_{0d} is the intercept, $\lambda_1 > 0$ is the slope, and $\theta \sim \mathcal{N}(0, 1)$ by assumption. Equation (2.19) renders the probability monotone in θ and dependent on an intercept parameter and a slope parameter that is equal across attributes. This is the higher-order DINA (de la Torre & Douglas, 2004), and an MCMC algorithm is used to estimate $D - 1$ intercept and 1 slope parameter. Yang and Embretson (2007) discussed an equation similar to Eq. (2.18) for inferring a person's most likely latent class $\mathbf{a}_h$ given her item-score pattern $\mathbf{X}_n$, the item parameters $\mathbf{g}$ and $\mathbf{s}$, and design matrix $\mathbf{Q}$.

Second, Junker and Sijtsma (2001) studied the properties of the DINA model from the perspective of the monotone homogeneity model and focused on monotonicity properties. Before we consider their result, we first notice a stochastic ordering result different from SOL, which reverses the roles of latent and manifest variables, and therefore is called stochastic ordering of the manifest variable by the latent variable (SOM). Starting from UD, M, and LI, and Eq. (2.4), for any pair of persons with $\theta_{n_1} < \theta_{n_2}$, Hemker et al. (1997) defined SOM as

$$P\big(X_+ \geq x_{+c}|\theta_{n_1}\big) \leq P\big(X_+ \geq x_{+c}|\theta_{n_2}\big)\,. \tag{2.20}$$

The monotone homogeneity model thus supplies a latent structure justifying ordering people on the observable X_+ total score. Older approaches, such as classical test theory, did not supply such a justification, but simply recommended the use of X_+. With the exception of the Rasch model, modern IRT approaches based on UD, M, and LI missed that they also justify SOM and even the more useful SOL, which allows one making inferences about latent, explanatory structures—an ordinal latent scale—from observable data. Holland and Rosenbaum (1986) introduced the notion of non-decreasing summaries of the item scores, denoted $g(\mathbf{X}_n)$, non-decreasing coordinate-wise in X_k ($k = 1, \ldots, K$), and Junker and Sijtsma (2001) noticed that in the DINA model,

$$P[g(\mathbf{X}_n)\,|\mathbf{a}_n] \text{ is coordinate-wise non-decreasing in } \mathbf{a}_n \tag{2.21}$$

Obviously, this is a SOM property, meaning that the mastery of more attributes yields a higher summary score. The authors were unable to derive similar SOL properties for the DINA model.

2.3 Example

2.3.1 Data: Millon Clinical Multiaxial Inventory-III

We used the item scores of 1210 Caucasian patients and inmates in Belgium (61% males) on 44 items of the Dutch version of the Millon Clinical Multiaxial Inventory-III (MCMI-III; Millon, Millon, Davis, & Grossman, 2009; Dutch version by Rossi, Sloore, & Derksen, 2008). For more details about the sample, see Rossi, Elklit, and Simonsen (2010), and for a previous data analysis, see de la Torre, van der Ark, and Rossi (2018). The MCMI-III consists of 175 dichotomous items defining 14 personality scales, 10 clinical syndrome scales, and 5 correction scales. The 44 items we used pertain to the clinical syndrome scales anxiety (A), somatoform (H), thought disorder (SS) and major depression (CC). Several items are indicative for more than one clinical disorder (Table 2.1). For example, a positive response to Item 148 ("Few things in life give me pleasure") is believed to be an indicator for somatoform, thought disorder, and major depression. The 44 × 4 Q-matrix (Table 2.2) reflects the contributions of each item to each scale.

Table 2.1 The number of items per scale measuring one, two, or three disorders

Scale	Number of disorders			Total
	1	2	3	
A	9	5	0	14
H	2	9	1	12
SS	11	5	1	17
CC	6	11	1	18

Note: *A* Anxiety, *H* Somatoform, *SS* thought disorder, *CC* major depression

Table 2.2 Q-matrix of 44 items by four clinical disorders

	Disorder					Disorder					Disorder					Disorder			
k	A	H	SS	CC	*k*	A	H	SS	CC	*k*	A	H	SS	CC	*k*	A	H	SS	CC
1	0	1	0	1	61	1	0	1	0	108	1	0	0	0	147	1	0	0	0
4	0	1	0	1	68	0	0	1	0	109	1	0	0	0	148	0	1	1	1
11	0	1	0	0	72	0	0	1	0	111	0	1	0	1	149	1	0	0	1
22	0	0	1	0	74	0	1	0	1	117	0	0	1	0	150	0	0	0	1
34	0	0	1	1	75	1	1	0	0	124	1	0	0	0	151	0	0	1	1
37	0	1	0	0	76	1	0	1	0	128	0	0	0	1	154	0	0	0	1
40	1	0	0	0	78	0	0	1	0	130	0	1	0	1	162	0	0	1	0
44	0	0	0	1	83	0	0	1	0	134	0	0	1	0	164	1	0	0	0
55	0	1	0	1	102	0	0	1	0	135	1	0	0	0	168	0	0	1	0
56	0	0	1	0	104	0	0	0	1	142	0	0	1	1	170	1	0	0	0
58	1	0	0	0	107	0	1	0	1	145	1	1	0	0	171	0	0	0	1

Note: *k* Item number, *A* Anxiety, *H* Somatoform, *SS* thought disorder, *CC* major depression

We screened the data for outliers using the number of Guttman errors as an outlier score (Zijlstra, van der Ark, & Sijtsma, 2007), which identifies inconsistent item-score patterns given the item ordering based on item-total scores. Two respondents had an unexpectedly large number of Guttman errors, well beyond the cutoff value suggested by the adjusted boxplot (Hubert & Vandervieren, 2008). These respondents were removed from the data, yielding a final sample size of $N = 1208$. The data contained no missing values.

2.3.2 Analysis of the Data

Nonparametric IRT Analysis. We investigated the assumptions of the monotone homogeneity model by means of Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & van der Ark, 2017), using the R package `mokken` (van der Ark, 2007, 2012). First, we conducted a confirmatory Mokken scale analysis assuming all items belonged to the same scale. For all 946 item-pairs, scalability coefficient H_{jk} was significantly greater than 0 ($.10 < H_{jk} < .88$). Except for item 117 ($H_{117} = .28$), item 154 ($H_{154} = .25$), and item 168 ($H_{168} = .29$), for each of the other 41 items, item-scalability coefficient H_k was significantly greater than .30 ($.25 < H_k < .60$). These results support the fit of the monotone homogeneity model and suggest that the 44 items form a unidimensional scale. Total-scale scalability coefficient $H = .42$ ($se = .01$), a value which Mokken labeled as a medium scale. Hence, based on scalability coefficients alone, we did not find support that the four scales represent different clinical disorders. This first result might imply that a CDM with four attributes is superfluous.

Second, we conducted an exploratory Mokken scale analysis. For lower-bound values $c \in \{.00, .05, .10, .15, \ldots, .60\}$, we partitioned the 44 items into scales requiring that items admitted to a scale have $H_k > c$. This means that items may drop out of scales and remain unscalable. For $c \leq .20$, all 44 items constituted a single scale. For $.20 < c \leq .35$, some items were unscalable (i.e., item 154 was unscalable at $c = .25$, items 22, 117, 154, and 168 were unscalable at $c = .35$) but the remaining items constituted a single scale. For $c > .35$, the one-scale structure fell apart into multiple scales (3 scales at $c = .40$ to 11 scales at $c = .60$) and up to 7 unscalable items. At first glance, these results also support the hypothesis that the data are approximately unidimensional suggest. However, when one applies stricter criteria for scalability, the items represent a smaller number of attributes, and a closer look may be in order.

Third, we inspected local independence using the W indices (notation W not to be confused with the number of latent classes) proposed by Straat, van der Ark, and Sijtsma (2016). Space limitations do not permit a discussion of these indices; hence, we refer the interested reader to Straat et al. (2016). Index W_1, which is used for the detection of positive locally dependent item pairs, flagged 106 of the 946 item pairs; index W_3, which is used for the detection of negative locally dependent item pairs, flagged two of the 946 item pairs. Because we did not have benchmarks

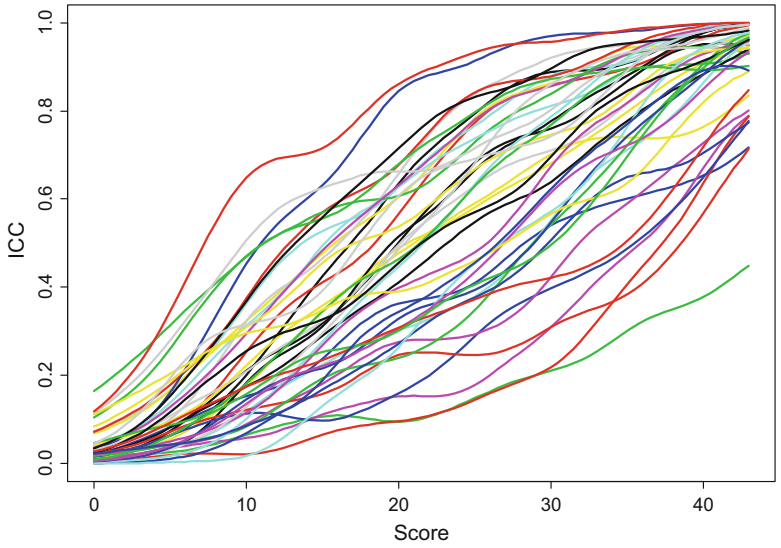


Fig. 2.2 Forty-four item response functions estimated by means of kernel smoothing ($h = 2.5$)

for W indices for so many items, we tentatively concluded that some items may be positive locally dependent, which suggests that within the unidimensional scale a more refined structure may be present.

Fourth, another in-depth analysis concerned the investigation of monotonicity using the property of manifest monotonicity (Junker & Sijtsma, 2000; Sijtsma & van der Ark, 2017). We did not find violations. This finding supports the fit of the monotone homogeneity model. Figure 2.2 shows the 44 IRFs estimated by means of kernel smoothing (Ramsay, 1991, 1996), using smoothing parameter $h = 2.5$.

To conclude, using the monotone homogeneity model, different clinical disorders remained unidentified and the data were unidimensional, but the unidimensionality signal was moderate. When we used stricter scaling criteria, the item set did not break down into four smaller scales that related to the four clinical disorders. To find out whether this result was a nonparametric-method effect, we also computed Yen's (1981) Q_1 statistic based on ten groups for testing goodness of fit of the parametric two-parameter logistic model (Table 2.3). The two-parameter logistic model is a special case of the monotone homogeneity model. Using Bonferroni correction (i.e., $p \approx .001$), based on the Q_1 statistic, none of the items showed misfit. Thus, the global goodness of fit measure for the two-parameter logistic model produced a result similar to that obtained from the confirmatory analysis in the context of the monotone homogeneity model, so that we could exclude a method effect.

Latent Class Analysis. We estimated twelve LCMs with $W = 1, 2, \ldots, 12$ classes using the R package poLCA (Linzer & Lewis, 2011), and computed information indices AIC3 (Andrews & Currim, 2003) and BIC (Schwarz, 1978). For large sample sizes and modest numbers of latent classes, both AIC3 and BIC are known

Table 2.3 Q_1 statistic for fit of two-parameter logistic model to each of 44 items

k	χ^2	p	k	χ^2	p	k	χ^2	p	k	χ^2	p
1	4.3	.828	61	14.3	.075	108	13.9	.085	147	16.3	.038
4	10.6	.226	68	6.8	.561	109	12.8	.120	148	8.9	.347
11	5.2	.735	72	9.7	.289	111	5.3	.727	149	16.5	.036
22	17.5	.025	74	15.5	.049	117	16.6	.035	150	16.9	.031
34	6.8	.561	75	10.6	.228	124	11.8	.162	151	6.6	.575
37	17.6	.025	76	10.0	.267	128	13.6	.093	154	2.2	.973
40	9.3	.320	78	23.1	.003	130	10.2	.252	162	18.2	.020
44	5.5	.703	83	10.3	.244	134	8.6	.378	164	10.0	.263
55	15.1	.057	102	11.8	.162	135	2.8	.948	168	9.9	.269
56	9.1	.337	104	12.5	.128	142	6.6	.580	170	19.9	.011
58	15.8	.046	107	13.8	.088	145	10.8	.213	171	9.5	.299

Note: k Item number, χ^2 chi-squared statistic with 10 degrees of freedom, p p value

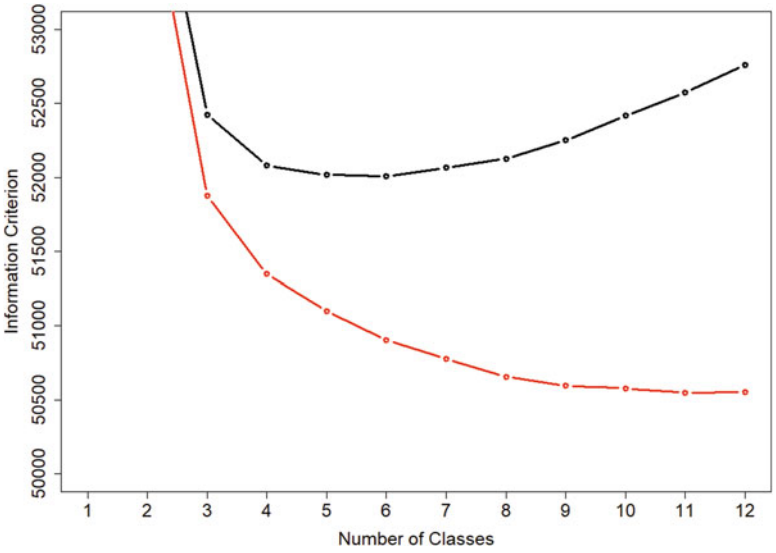


Fig. 2.3 BIC (black) and AIC3 (red) values for LCMs with 1, 2, . . . , 12 latent classes

to identify the correct number of classes reasonably well, but BIC tends to be conservative (Yang & Yang, 2007). To decrease the risk of local maxima, we estimated each model 10 times. We discuss results for the four-class LCM, because the number of classes is conveniently small, its interpretation relatively easy, while the fit of the model in terms of BIC seems adequate, and for the six and eleven-class LCMs, because these models provided the smallest values of AIC3 and BIC, respectively (Fig. 2.3).

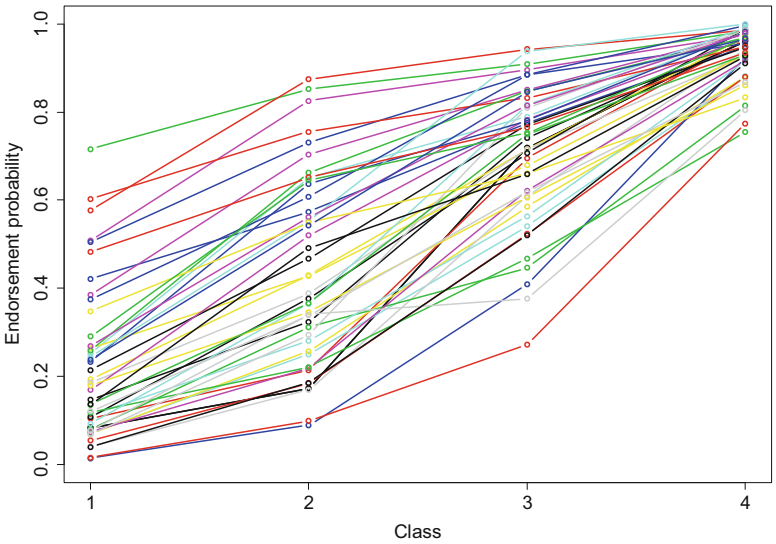


Fig. 2.4 *Endorsement probabilities for the four-class LCM*

Because it did not provide the smallest AIC3 and BIC values, the four-class LCM may have missed some of the heterogeneity in the data but the model corroborated the results from the nonparametric IRT data analysis. The four classes, with class probabilities $P(\Phi = w)$ equal to .288, .266, .159, and .291, are strictly ordinal, because for all 44 items, the estimated endorsement probability $P(X_i = 1 | \Phi = w)$ increased as w increased (Fig. 2.4).

Except for classes 3 and 4, the six-class LCM showed increasing endorsement probabilities (Fig. 2.5). The class probabilities $P(\Phi = w)$ equal .112, .264, .194, .111, .134, and .184. Figure 2.5 shows absence of consistent ordering between classes 3 and 4: For 17 of the 44 items (solid lines) the endorsement probability was larger in Class 4 than in Class 3. Fourteen of the 17 items relate to major depression or somatoform disorder. Hence, in addition to the ordinal trend, the six-class LCM seemed to distinguish a class with moderate endorsement probabilities leaning towards major depression and somatoform disorders (Class 3) and a class with moderate endorsement probabilities leaning towards anxiety disorders and thought disorders (Class 4). The eleven-class LCM was too difficult to interpret without an a priori hypothesized structure. Next, we investigated whether CDMs can provide additional information about the data structure.

Cognitive Diagnosis Models. Because this chapter discusses CDMs relative to nonparametric IRT and nonparametric LCMs, we compared these models with nonparametric CDMs. Our ambition was not to be complete with respect to the discussion of CDMs, but to discuss the general idea using a few simple models. The choice of two nonparametric CDMs, the basic DINA and DINO, reflect this modest ambition. Because these models are rather restrictive, we did not expect them to fit the data but used them instead for didactical purposes. We estimated the models

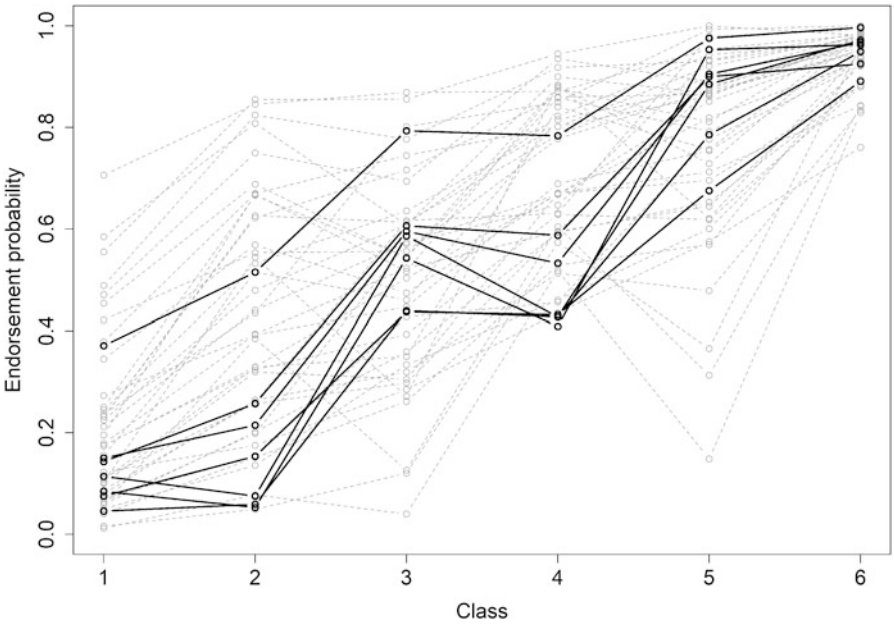


Fig. 2.5 *Endorsement probabilities for the six-class LCM. Black solid lines pertain to 17 items that have higher endorsement probabilities in Class 3 than in Class 4*

using the R package `NPCD` (Zheng & Chiu, 2016). First, the attribute profiles $\mathbf{A}$ were estimated using a nonparametric algorithm minimizing the plain Hamming distance (Chiu & Douglas, 2013), and given estimate $\widehat{\mathbf{A}}$, maximum likelihood estimates of the guessing parameters and slip parameters were obtained. Table 2.4 shows DINA results and Table 2.5 shows DINO results. Items printed in boldface had a high slipping or a high guessing parameter estimate, and the models fitted worse for these items. For global model fit, R package `NPCD` provides AIC and BIC but not AIC3. For the DINA model, AIC = 50,705 and BIC = 51,154, and for the DINO model, AIC = 50,296 and BIC = 50,745. One may notice that one cannot compare the AIC and BIC values of the these CDMs to the AIC and BIC values of the LCMs in Fig. 2.3. The reason is that for the CDMs, the likelihood is derived under the assumption that the Hamming distance-based class assignments ($\widehat{\mathbf{A}}$) are fixed, whereas for the LCMs, the class assignments are part of the likelihood (von Davier, personal communication). Comparing AIC and BIC of the CDMs and LCMs would be unfair and in favor of the DINO and the DINA given fixed $\widehat{\mathbf{A}}$.

For this example, $\Xi_{nk} = 1$ (DINA) means that respondent n suffers from all the disorders item k assesses, and $\Psi_{nk} = 1$ (DINO) means that respondent n suffers from at least one of the disorders item k assesses. Slipping parameter $s_k = P(X_{nk} = 0 | \Xi_{nk} = 1)$ (Eq. 2.11) is the probability that respondent n does not endorse item k, even though respondent n suffers from all the disorders related to item k; and guessing parameter $g_k = P(X_{nk} = 1 | \Xi_{nk} = 0)$ (Eq. 2.12) is the probability that respondent n endorses item k, even though respondent n does not

Table 2.4 Slipping and guessing parameters for the 44 items estimated from the DINA model

k	s_k	g_k	k	s_k	g_k	k	s_k	g_k	k	s_k	g_k
1	.08	.26	61	.03	.40	108	.34	.13	147	.05	.42
4	.10	.26	68	.17	.26	109	.16	.30	148	.15	.24
11	**.73**	**.03**	72	.15	.26	111	.28	.19	**149**	**.38**	**.13**
22	**.44**	**.11**	74	.22	.19	**117**	**.64**	**.08**	150	.38	.04
34	.11	.35	75	.20	.21	**124**	**.51**	**.05**	151	.33	.13
37	**.63**	**.03**	76	.18	.18	**128**	**.47**	**.06**	**154**	**.39**	**.22**
40	.19	.23	**78**	**.72**	**.02**	130	.16	.16	162	.19	.25
44	.07	.25	**83**	**.13**	**.38**	134	.33	.12	164	.31	.13
55	.15	.21	**102**	**.60**	**.07**	135	.24	.19	**168**	**.52**	**.12**
56	.12	.33	**104**	**.45**	**.11**	142	.12	.19	**170**	**.48**	**.08**
58	.19	.30	**107**	**.48**	**.09**	145	.08	.36	171	.35	.09

Note: k Item number, s_k slipping parameter, g_k guessing parameter. If $s_k + g_k > .5$, the values are printed in boldface

Table 2.5 Slipping and guessing parameters for the 44 items estimated from the DINO model

k	s_k	g_k	k	s_k	g_k	k	s_k	g_k	k	s_k	g_k
1	.09	.19	61	.07	.24	108	.34	.16	147	.05	.45
4	.13	.20	68	.14	.31	109	.15	.32	148	.28	.14
11	**.47**	**.06**	72	.14	.32	111	.32	.15	**149**	**.48**	**.07**
22	**.40**	**.14**	74	.26	.15	117	**.59**	**.09**	150	.29	.08
34	.16	.29	75	.32	.14	124	**.48**	**.06**	**151**	**.43**	**.10**
37	**.44**	**.09**	76	.29	.08	128	**.42**	**.10**	**154**	**.34**	**.25**
40	.18	.25	78	**.69**	**.03**	130	.21	.12	162	.14	.28
44	.04	.32	83	**.10**	**.42**	134	.27	.15	164	.30	.14
55	.19	.16	102	**.55**	**.08**	135	.25	.23	**168**	**.47**	**.14**
56	.11	.38	104	**.41**	**.15**	142	.19	.11	**170**	**.47**	**.09**
58	.17	.31	107	**.52**	**.07**	145	.14	.23	171	.28	.12

Note: k Item number, s_k slipping parameter, g_k guessing parameter. If $s_k + g_k > .5$, the values are printed in boldface

suffer from all the disorders related to item k. Because in the clinical context, one assumes that one endorses an item if one possesses at least one of the disorders, the DINO model seems more in line with this assumption than the DINA model. Based on the BIC, the DINO model fitted better than the DINA model, but for both models, proportions of slipping and guessing were high; see Tables 2.4 and 2.5.

For the DINO model, slipping parameter estimates were generally higher than guessing parameter estimates, and for 14 items, slipping parameter estimates exceeded .40 (Table 2.5). Hence, respondents suffering from a relevant disorder did not always endorse the item. An explanation could be that some items refer to rare circumstances. An example is item 78, "Even when I'm awake, I don't seem to notice people who are near me" ($s_{78} = .69$) that even respondents suffering from thought disorder may find too unlikely to endorse. Some other questions were

Table 2.6 Class sizes based on the Hamming distance-based attribute profiles for the DINO model in percentages

Class	Prevalence	Class	Prevalence
No disorder	45.3%	H and SS	1.6%
CC	1.9%	A and SS	4.7%
SS	2.4%	A and H	1.7%
H	2.1%	All but A	0.5%
A	8.7%	All but H	12.6%
CC and SS	1.3%	All but SS	1.1%
CC and H	0.2%	All but CC	3.4%
CC and A	6.5%	All disorders	6.0%

Note *A* anxiety, *H* somatoform, *SS* thought disorder, *CC* major depression

double barreled, which may explain low endorsement. An example is item 107, "I completely lost my appetite and have trouble sleeping most nights". Only for item 83, "My moods seem to change a great deal from one day to the next", the guessing parameter estimate exceeded .40 ($g_{83} = .42$). Item 83 relates to thought disorder, but given its high popularity, respondents not suffering from thought disorder also seemed to endorse the item.

Based on the estimated attribute profiles $\widehat{\mathbf{A}}$ of the DINO model, four attribute profiles had substantial size (Table 2.6): no disorder (45.3%), only A (8.7%), CC and A (6.5%), and SS, CC and A (12.6%). Approximately 73% of the respondents belonged to one of these four classes. If one adds the percentages in Table 2.6 that pertain to A, one finds that the DINO model identified anxiety (44,7%) as the most common disorder, followed by thought disorder (32,5%), major depression (30.2%), and somatoform (16,6%).

2.4 Discussion

This chapter discussed the relation between nonparametric IRT models and CDMs. The two approaches are related via the LCM, and both IRT models and CDMs may be viewed as restricted LCMs with a large number of classes. For IRT models, the number of classes equals the number of distinct θ values, but IRT models are mainly used for measuring individuals on a scale for the attribute of interest, and for this purpose the IRFs or ISRFs are nonparametrically or parametrically restricted. For CDMs, the number of classes equals 2^D attribute profiles, and a parametric functional form restricts the response probabilities within a class.

The nonparametric IRT models, LCMs, and CDMs are related, but researchers use the models in different situations. Nonparametric IRT models are useful for ordinal measurement and as a preliminary analysis for measurement using parametric IRT models, whereas LCMs are useful for nominal measurement; that is, identifying prototypes of respondents in the data, but also as a density estimation tool. CDMs are also used for nominal measurement. Yet by identifying the presence

or the absence of cognitive skills or clinical disorders, CDMs provide insight into the attribute of interest.

Comparing the fit of the nonparametric IRT models to an LCM or a CDM is not straightforward. Nonparametric IRT models have many methods to investigate the local fit, but a global goodness of fit statistic is unavailable. If one uses an ordinal LCM to investigate goodness of fit of a nonparametric IRT model, relative fit measures such as AIC or AIC3 are available. However, these measures suffer from the problem that they indicate which of the models the researcher compares fits best to the data, but not whether the best fitting of these models actually fits the data well. The data analysis using local fit methods showed that the nonparametric IRT model fitted well, and the nonparametric CMDs fitted well relative to the LCM.

The interpretation of the nonparametric IRT model and the CDMs was different. Using the IRT models, one uses a single continuous attribute to explain the responses to four comorbid disorders. The CDMs provide additional information. First, one could argue that the CDM analyses corroborated the conclusion from nonparametric IRT and LCA that the data were largely unidimensional, because classes showed a cumulative structure. That is, Class 0000 represents no disorders (45.3%), Class 1000 represents only an anxiety disorder (8.7%), Class 1010 represents anxiety and thought disorder (6,5%), Class 1110 represents all disorders but major depression (12.6%), and Class 1111 represents all disorders (6%). The classes can be considered ordered. Because over 73% of the sample belonged to these classes, a unidimensional scale may be appropriate for the majority of the sample (also, see von Davier & Habermann, 2014). However, to obtain a finer-grained picture, CDMs can retrieve information from the data that IRT models cannot.

References

Andrews, R. L., & Currim, I. S. (2003). A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research, 40*, 235–243. https://doi.org/10.1509/jmkr.40.2.235.19225

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573. https://doi.org/10.1007/BF02293814

Bouwmeester, S., Vermunt, J. K., & Sijtsma, K. (2007). Development and individual differences in transitive reasoning: A fuzzy trace theory approach. *Developmental Review, 27*, 41–74. https://doi.org/10.1016/j.dr.2006.08.001

Brusco, M. J., Köhn, H.-F., & Steinley, D. (2015). An exact method for partitioning dichotomous items within the framework of the monotone homogeneity model. *Psychometrika, 80*, 949–967. https://doi.org/10.1007/s11336-015-9459-8

Chiu, C. Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*, 225–250. https://doi.org/10.1007/s00357-013-9132-9

Croon, M. (1990). Latent class analysis with ordered latent classed. *British Journal of Mathematical and Statistical Psychology, 43*, 171–192. https://doi.org/10.1111/j.2044-8317.1990.tb00934.x

Croon, M. A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology, 44*, 315–331. https://doi.org/10.1111/j.2044-8317.1991.tb00964.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J, van der Ark, L. A., & Rossi, G. (2018). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development, 51*, 281–296. https://doi.org/10.1177/0748175615569110

Douglas, J. A. (2001). Asymptotic identifiability of nonparametric item response models. *Psychometrika, 66*, 531–540. https://doi.org/10.1007/BF02296194

Douglas, R., Fienberg, S. E., Lee, M.-L. T., Sampson, A. R., & Whitaker, L. R. (1991). Positive dependence concepts for ordinal contingency tables. In H. W. Block, A. R. Sampson, & T. H. Savits (Eds.), *Topics in statistical dependence* (pp. 189–202). Hayward, CA: Institute of Mathematical Statistics. Retrieved from http://www.jstor.org/stable/4355592

Ellis, J. L. (2014). An inequality for correlations in unidimensional monotone latent variable models for binary variables. *Psychometrika, 79*, 303–316. https://doi.org/10.1007/s11336-013-9341-5

Formann, A. K., & Kohlmann, T. (2002). Three-parameter linear logistic latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.),. *Applied latent class analysis* (pp. 183–210). Cambridge, UK: Cambridge University Press.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231. https://doi.org/10.1093/biomet/61.2.215

Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika, 53*, 383–392. https://doi.org/10.1007/BF02294219

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Heinen, T. (1996). *Latent class and discrete latent trait models. Similarities and differences*. Thousand Oaks, CA: Sage.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*, 331–347. https://doi.org/10.1007/BF02294555

Hemker, B. T., van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika, 66*, 487–506. https://doi.org/10.1007/BF02296191

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171–189. https://doi.org/10.1007/BF02295273

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics, 14*, 1523–1543.. Retrieved from https://projecteuclid.org/euclid.aos/1176350174

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis, 52*, 5186–5201. https://doi.org/10.1016/j.csda.2007.11.008

Junker, B. W. (1993). Conditional association, essential independence and monotone unidimensional item response models. *The Annals of Statistics, 21*, 1359–1378. Retrieved from http://www.jstor.org/stable/2242199

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*, 65–81. https://doi.org/10.1177/01466216000241004

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272. https://doi.org/10.1177/01466210122032064

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28*, 110–125. https://doi.org/10.1177/0146621603260678

Leighton, J. A., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education. Theory and applications*. Cambridge, UK: Cambridge University Press.

Ligtvoet, R., & Vermunt, J. K. (2012). Latent class models for testing monotonicity and invariant item ordering for polytomous items. *British Journal of Mathematical and Statistical Psychology, 65*, 237–250. https://doi.org/10.1111/j.2044-8317.2011.02019.x

Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis, 19*, 173–187. https://doi.org/10.1093/pan/mpr006

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–29. https://doi.org/10.18637/jss.v042.i10

MacReady, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99–120. https://doi.org/10.3102/10769986002002099

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. https://doi.org/10.1007/BF02296272

Michell, J. (1999). *Measurement in psychology. A critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Millon, T., Millon, C., Davis, R., & Grossman, S. (2009). *MCMI-III Manual* (4th ed.). Minneapolis, MN: Pearson Assessments.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands/Berlin, Germany: Mouton/De Gruyter.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. https://doi.org/10.1177/014662169201600206

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630. https://doi.org/10.1007/BF02294494

Ramsay, J. O. (2016). Functional approaches to modeling response data. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume one. Models* (pp. 337–350). Boca Raton, FL: Chapman & Hall/CRC.

Rossi, G., Elklit, A., & Simonsen, E. (2010). Empirical evidence for a four factor framework of personality disorder organization: Multigroup confirmatory factor analysis of the million clinical multiaxial inventory–III personality disorders scales across Belgian and Danish data samples. *Journal of Personality Disorders, 24*, 128–150. https://doi.org/10.1521/pedi.2010.24.1.128

Rossi, G., Sloore, H., & Derksen, J. (2008). The adaptation of the MCMI-III in two non-English-speaking countries: State of the art of the Dutch language version. In T. Millon & C. Bloom (Eds.), *The Millon inventories: A practitioner's guide to personalized clinical assessment* (2nd ed., pp. 369–386). New York, NY: The Guilford Press.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement. Theory, methods, and applications*. New York, NY: The Guilford Press.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. https://doi.org/10.2307/2958889

Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63*, 183–200. https://doi.org/10.1007/BF02294774

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sijtsma, K., & Molenaar, I. W. (2016). Mokken models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Volume one. Models* (pp. 303–321). Boca Raton, FL: Chapman & Hall/CRC.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*, 137–158. https://doi.org/10.1111/bmsp.12078

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*, 506–532. https://doi.org/10.1177/1094428116630065

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55*, 293–326. https://doi.org/10.1007/BF02295289

Stout, W. F. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518. https://doi.org/10.1007/BF02295128

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 75–99. https://doi.org/10.1007/s00357-013-9122-y

Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 12*, 117–123. https://doi.org/10.1027/1614-2241/a000115

Suppes, P., & Zanotti, M. (1981). When are probabilistic explanations possible? *Synthese, 48*, 191–199. https://doi.org/10.1007/BF01063886

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Tijmstra, J., Hessen, D. J., van der Heijden, P. G. M., & Sijtsma, K. (2013). Testing manifest monotonicity using order-constrained statistical inference. *Psychometrika, 78*, 83–97. https://doi.org/10.1007/s11336-012-9297-x

van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*, 283–304. https://doi.org/10.1007/s11336-000-0862-3

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*(11), 1–19. https://doi.org/10.18637/jss.v020.i11

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(5), 1–27. https://doi.org/10.18637/jss.v048.i05

van der Ark, L. A., & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272–279. https://doi.org/10.1007/S11336-010-9147-7

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*, 380–392.

van der Linden, W. J. (Ed.). (2016). *Handbook of item response theory. Volume one. Models*. Boca Raton, FL: Chapman & Hall/CRC.

van Onna, M. J. H. (2002). Bayesian estimation and model selection in ordered latent class models for polytomous items. *Psychometrika, 67*, 519–538. https://doi.org/10.1007/BF02295129

van Schuur, W. H. (2011). *Ordinal item response theory. Mokken scale analysis*. Thousand Oaks, CA: Sage.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing nonparametric and parametric item response theory models. *Applied Psychological Measurement, 25*, 283–294. https://doi.org/10.1177/01466210122032082

Vermunt, J. K., van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology, 38*, 369–397.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71. https://doi.org/10.1111/bmsp.12003

von Davier, M., & Haberman, S. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models – A commentary. *Psychometrika, 79*, 340–346. https://doi.org/10.1007/s11336-013-9363-z

Wetzel, E., Xu, X., & Von Davier, M. (2015). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement, 75*, 739–763.

Yang, C.-C., & Yang, C.-C. (2007). Separating latent classes by information criteria. *Journal of Classification, 24*, 183–203. https://doi.org/10.1007/s00357-007-0010-1

Yang, X., & Embretson, S. E. (2007). Construct validity and cognitive diagnostic assessment. In J. A. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education. Theory and applications* (pp. 119–145). Cambridge, UK: Cambridge University Press.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245–262. https://doi.org/10.1177/014662168100500212

Zheng, Y., & Chiu, C.-Y. (2016). *NPCD: Nonparametric methods for cognitive diagnosis*. R package version 1.0–10 [computer software]. Retrieved from https://CRAN.R-project.org/package=NPCD

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42*(3), 531–555.

Zijlmans, E. A. O., van der Ark, L. A., Tijmstra, J., & Sijtsma, K. (2018). Methods for estimating item-score reliability. *Applied Psychological Measurement, 42*, 553–570. https://doi.org/10.1177/0146621618758290.

Chapter 3
The Reparameterized Unified Model System: A Diagnostic Assessment Modeling Approach

William Stout, Robert Henson, Lou DiBello, and Benjamin Shear

Abstract This chapter considers the Reparameterized Unified Model (RUM). The RUM a refinement of the DINA where which particular required skills that are lacking influences the probability of a correct response: Hartz, A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practicality. Dissertation, University of Illinois at Urbana-Champaign, 2001; Roussos, DiBello, Stout, Hartz, Henson, and Templin. The fusion model skills diagnosis system. In: JP Leighton and MJ Gierl (eds) Cognitive diagnostic assessment for education: Theory and applications. New York, Cambridge University Press, pp 275–318, 2007). The RUM diagnostic classification models (DCM) models binary (right/wrong scoring) items as the basis for a stills diagnostic classification system for scoring quizzes or tests. Refined DCMs developed from the RUM are discussed in some detail. Specifically, the commonly used "Reduced" RUM and an extension of the RUM to option scored items referred to as the Extended RUM model (ERUM; DiBello, Henson, & Stout, Appl Psychol Measur 39:62–79, 2015) are also considered. For the ERUM, the latent skills space is augmented by the inclusion of misconceptions whose possession reduces the probability of a correct

This research is supported by IES Grant R305D140023.

This book chapter is dedicated to Lou DiBello, deceased in March 2017, who contributed seminally to the research and methodology reported herein.

W. Stout (✉) · L. DiBello (deceased)
Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA
e-mail: w-stout1@illinois.edu

R. Henson
Educational Research Methodology (ERM) Department, The University of North Carolina at Greensboro, Greensboro, NC, USA
e-mail: rahenson@uncg.edu

B. Shear
Research and Evaluation Methodology, University of Colorado, Boulder, CO, USA
e-mail: Benjamin.Shear@Colorado.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_3

response and increases the probability of certain incorrect responses, thus providing increasing classification accuracy. In addition to discussion of the foundational identifiability issue that occurs for option scored DCMs, available software using the SHINY package in R and including various appropriate "model checking" fit and discrimination indices is discussed and is available for users.

3.1 Introduction

Restricted latent class modeling (RLCM) is a major diagnostic classification modeling (DCM) approach. The parametrically lean DINA RLCM (Haertel, 1989; see also Chap. 1: von Davier & Lee, this volume) and the more nuanced parametrically rich Unified Model (DiBello, Stout, & Roussos, 1995), denoted UM, have helped initiate a strong renaissance of RLCM-based DCM research and applications. This chapter considers the Reparameterized Unified Model (Hartz, 2001) for binary data, denoted RUM, and the option scoring based Extended RUM model (DiBello et al., 2015), denoted ERUM. The chapter's central focus is how the RUM and its ERUM generalization provide a highly useful approach for statistically analyzing multiple choice (MC) test data to diagnose student knowledge when using diagnostic assessments (DAs). Herein, we define a *diagnostic assessment (DA*) to be any assessment that classifies students in multidimensional detail, such as classifying examinees as one of 2^D elements of a discrete 0/1 mastery/nonmastery D dimensional latent "skills" space.

This chapter summarizes the *RUM/ERUM model-based diagnostic test data analysis system,* which for simplicity is referred to as the *RUM diagnostic system*. Here "system" refers to the combination of DCM model, estimation procedure and classification algorithm, model-data-checking (namely, fit and discrimination), and software that carries out the complete RUM system. In addition to the RUM and its "Reduced RUM" (RRUM) reduction, which are introduced below, the RUM diagnostic system also allows selection of the ERUM model. The main goals of a RUM/RRUM or ERUM based analysis are model estimation (especially the item response functions, IRFs) and subsequent examinee classification. To adequately carry out these inferences, the RUM diagnostic system includes a user-available simulation option that uses the estimated DCM model, allowing the user to roughly predict the RUM diagnostic system's classification performance for her DA application. As explained in the Software section, the various components of the RUM diagnostic system are available in a SHINY (see Beeley, 2013) driven user interface. Of course, there are other software options available when using the RUM/RRUM DCM model (e.g., Chiu, Kohn, & Wu, 2016: uses an EM algorithm approach; Feng, Habing, & Huebner, 2014: also uses an EM approach; Chung & Johnson, 2017: uses MCMC; and Chap. 24: Zhang, Douglas, Wang, & Culpepper, this volume: also uses MCMC). That is, no claim is being made here that one *should* use our RUM system software, nor our MCMC approach!: it is merely available for interested users.

In addition, this chapter discusses the challenging foundational issue of the intrinsic non-identifiability that can occur for ERUM and similarly structured option-scoring-based DCM modeling of multiple choice (MC) items. Also, the chapter summarizes a large number of simulation studies (largely unpublished) and surveys various RUM-system-based real data applications. These simulation and real data studies combine to demonstrate the RUM system's DA potential, both when using right/wrong scoring RUM/RRUM and option scoring ERUM.

3.2 The RUM and RRUM

We first introduce the more often used and parametrically simpler *RRUM* and then present the less parsimonious RUM, RRUM's historical precursor. The RRUM (unlike the RUM, as clarified below) is a RLCM. In particular, its D dimensional latent space $\{\boldsymbol{a}\}$ of examinees with latent distribution (or prior distribution from the Bayesian viewpoint) $P(\boldsymbol{a})$ consists entirely of binary "skills", referred to sometimes as attributes (mastered or not: $\alpha_d = 1$ or 0 respectively; $d = 1, \ldots, D$), resulting in a RLCM with 2^D possible latent classes. Let P $(\boldsymbol{x}|\boldsymbol{\alpha})$ denote the RRUM RLCM where $\boldsymbol{x}$ denotes the test response vector of a K item test. Assuming local independence (LI) and examinees responding independently, specifying the RRUM model amounts to specifying the IRF P$(x_k|\boldsymbol{a})$ for each item k with x_k=1 denoting a correct response and 0 an incorrect response, thus producing the RLCM.

The RRUM, like many other DCMs, requires that an (*item* × *skill*) 0/1 incidence matrix $Q^{[K \times D]}$ be specified for the test. The test's design matrix Q specifies for each item which skills are required to produce an "item master": An examinee $\boldsymbol{a}$ is an *item master* of item k provided all the item's required skills (all d such that $q_{kd} = 1$) are mastered. Ideally, P$(x_k|$*item master*$) \approx 1$, at least roughly so. A critical requirement for effective RRUM, or other DCM, modeling is that the Q matrix be reasonably accurate, either as developed via expert assignment (via cognitive scientist, curriculum specialist, student think-alouds, classroom teacher, etc.) alone or, ideally, augmented by data-driven Q estimation and/or by Q model fit considerations, noting the issue of Q matrix determination for DCMs is discussed in Chap. 12 (Liu & Kang, this volume) and is briefly addressed for RUM/ERUM below. By "reasonably accurate" is meant that the specified latent space includes most of the important skills affecting item performance and further that Q captures for each item the particulars of which skills influence the item responses.

The RRUM IRF defines the probability of a correct response given an examinee's mastery profile as

$$P\left(X_k = 1 \;\middle|\; \mathbf{a}\right) = \pi_k \prod_{d=1}^{D} r_{kd}^{q_{kd}(1-\alpha_d)}, \tag{3.1}$$

where $0 < r_{kd} \leq 1$, $0 < \pi_k \leq 1$ are assumed. If $\boldsymbol{a}$ is an item master and thus has mastered all the required attributes for that item (d required if $q_{kd} = 1$), according to Eq. 3.1, $\pi_k = P(X_k = 1|\boldsymbol{a})$, thus providing the substantive interpretation of π_k. When $\boldsymbol{a}$ lacks one or more required skills, $\boldsymbol{a}$ is termed an item non-master. As Eq. 3.1 states, for each required skill α_d not mastered, $r_{kd} < 1$ supplies a skill-specific penalty, providing the interpretation of r_{kd}. Note from the conjunctive modeling viewpoint that the DINA is fully conjunctive (except for noise) in the sense that $P(X_k = 1 \mid \boldsymbol{a})$ for DINA is the same value for *all* levels of non-mastery. By contrast, the more nuanced RRUM distinguishes between the various levels of item non-mastery and thus can be thought of as a merely "semi (as opposed to fully) -conjunctive" model in the sense that when lacking one (or more) skill, mastery of some of the other skills increases the probability of a correct response. Thus the RRUM, in contrast to the DINA, has the characteristic that differing levels of item non-mastery $\boldsymbol{a} < \boldsymbol{a}'$ imply that $P(X_k = 1|\boldsymbol{a}) < P(X_k = 1|\boldsymbol{a}') < P(X_k = 1|\textit{item master})$, and thus the closer to item mastery the more nonmastery is compensated for.

In its original non-reduced form, RUM also includes a $P_c(\eta)$ multiplicative factor: according to

$$P\left(X_k = 1|\boldsymbol{a}\right) = \left[\pi_k \prod_{d=1}^{D} r_{kd}^{q_{kd}(1-\alpha_d)}\right] P_c\left(\eta\right), \tag{3.2}$$

where $P_c(\eta) \equiv Logit(\eta + c)$ is a Rasch IRF probability with the posited continuous N(0, 1)[1] latent η encapsulating the combined influence of any influential skills not built into the specified latent class space, with c defining a *completeness factor*. A large value, e.g., $c = 2$ would indicate that the specified discrete D dimensional latent skills space captures most of the systematic IRF variation for item k and hence that approximate completeness of the posited latent learning space holds. Interestingly, sometimes RUM-based applications analyses display better fit with this continuous latent trait present whereas other analyses do not benefit from it (see the RUM/ERUM Applications section).

The *RUM diagnostic system* (RUM model, MCMC Estimator, Model-Data Checks, Simulator, Classifier)

The RUM diagnostic system was originally developed in the pivotal Hartz thesis and is further refined and summarized in the Roussos, DiBello, Stout, Hartz, Henson, and Templin (2007) chapter. Its user-available software package is named *Arpeggio* (see the manual; DiBello & Stout, 2008). Hartz proposed the RUM as a reparameterized version of the Unified Model, with the RUM parameters identified

[1]The original RUM (DiBello et al., 1995) also allowed compensatorily for mixing in the possible influence of non-Q based alternative cognitive strategies, but this added complexity has not been implemented anywhere and is thus not discussed herein. Indeed this multi-strategy generalization seems unrealistically complex and further in most cases not necessary for effective use of the RUM. Further, the original RUM had a slightly richer nonidentifiable parameterization than the RUM does with π_k in 3.2 replaced by a product of D substantively meaningful but nonidentifiable multiplicands π_{kd}.

unlike for the original Unified Model by DiBello et al. (1995). An MCMC-based hierarchical Bayesian estimation system was developed by Hartz together with customized fit and discrimination indices (this in cooperation with Louis Roussos). The RUM system software, Arpeggio, classifies examinees using this estimated RUM, possibly reduced to RRUM, as detailed in the Software section below.

The RUM Versus Fusion Model (FM) Terminology Confound The original RUM diagnostic system built around Arpeggio was dubbed the "Fusion Model Skills Diagnostic System" in Roussos et al. (2007). Although appropriately stressing the need for and carefully describing the complete RUM diagnostic package, the paper's terminology choice of "Fusion Model" produced an unnecessary terminology confound in that FM and RUM are *exactly* the same family of DCM model IRFs, sometimes including the $P_c(\eta)$ completeness factor as in Eq. 3.2 but sometimes without it as in Eq. 3.1. Following the literature, it would be correct to call Eq. 3.1 either the FM or the RRUM IRF. However, going forward, in this paper and elsewhere, "RRUM" (or "RUM" if the η factor is present) is recommended for clarity. Also, for clarity note that below throughout "*the RUM diagnostic system*" indicates the *current* diagnostic system (see Software section below) that allows the user to *also* choose the option scoring ERUM extension in addition to RUM or its RRUM reduction.

3.3 ERUM, the Option Scoring and Skills/Misconceptions Based Latent Space RRUM Generalization

As an extension of the RRUM, the ERUM models an examinee's MC item response beyond only correct or incorrect. Specifically, the ERUM is a multinomial extension of the RRUM. Furthermore, the ERUM latent space includes "misconceptions", defined below, in addition to "skills" in its latent space. ERUM's purpose is to improve upon the effectiveness of the RRUM and other right/wrong scoring DCMs by capturing most of the diagnostic information available from a DA-focused well-designed MC test.

3.3.1 The Formative Assessment MC Challenge as Aided by ERUM

DAs have many uses, among which DAs are sometimes used to assist in conducting classroom formative assessments. We define a *formative assessment (FA)* to be any assessment whose goal is to enhance learning *while* instruction and learning are occurring. FAs can include many different item types of varying complexities and formats, noting that this includes MC items, the modeling focus of the RUM system. Indeed, the authors' belief is that well-designed MC assessments, especially when

modeled by ERUM, will prove to be effective enough to become a valuable FA tool in the classroom. To illustrate, suppose the broad goal of a 6 week 8th grade algebra unit is the mastery of 8 essential curriculum-specified skills that combine to approximately capture unit competence. Then, can a 15–20 item MC mid-unit quiz be designed that reasonably assesses each student's attribute profile, with respect to the 256 possible profiles, as an interim assessment to guide instruction during the final 3 weeks?

Currently, most DCM modeled MC DAs only assess skills mastery/nonmastery and indeed do so only via right/wrong item scoring. In addition, instructors using MC tests to do FAs are often forced to use a curriculum-provided summative MC quiz or test not designed as a FA, which is also suboptimal. The use of MC tests should *and can* be improved upon in several critically important ways, which also adds to the motivation for introducing the ERUM. Specifically, MC FAs can be designed to detect incorrect or incomplete modes of thinking via test MC (or coded short answer) items created with carefully designed distractors that are attractive to examinees possessing particular incorrect or incomplete modes of thinking, labeled "misconceptions" for convenience.

Incomplete or incorrect modes of understanding often occur during the learning process. So FAs, in addition to assessing desirable modes of thinking and understanding (deemed "skills"), need to assess such misconceptions with the goal of supplanting them with improved states of understanding. Thus, both "skills" and "misconceptions" are merely convenient generic labels to encompass a wide variety of cognitive/instructional constructs that, conditional on the student's latent state, control (via each item's IRF) the probabilistic attractiveness of each option of each item. Both skills and misconceptions are called "attributes". As an illustration, suppose that, for a MC question with stem $\frac{1}{3} + \frac{2}{5}$, the responded MC option is $\frac{3}{8}$. This response clearly provides evidence of a particular misconception about adding fractions, even though it is possible that in individual cases this option was chosen via guessing.

We argue that an essential step in having truly effective MC instruments used for DA purposes is the introduction of useful DCMs that model option scoring and hence can capture diagnostic information given by "distractors". Using information from distractors is one main motivation for the ERUM generalization of the RRUM. We note that there are other existing psychometric approaches besides ERUM that link MC options to both desirable and undesirable modes of thinking, including the Ordered Multiple Choice Model (Briggs, Alonso, & Wilson, 2006), the MC-DINA model (de la Torre, 2009, see also Kuo et al., 2017), and the Scaling Individuals and Classifying Misconceptions (SICM) model (Bradshaw & Templin, 2014).

It must be noted there is a "catch 22" aspect for achieving effective DCM modeled MC DAs: Well-designed option scored MC tests will likely fail to provide effective diagnostics, especially of misconceptions, if the DCMs used are not option scored with misconceptions as well as skills present in the latent space. On the other hand, effective option scoring based DCMs like ERUM could seem not more useful than right/wrong scoring DCMs because many of the existing MC DAs fail to have been constructed to provide valuable information about both skills and

misconceptions via carefully chosen incorrect option responses. That is, effective *option scoring* based DCMs and MC DAs with *much information in their incorrect options must coexist* for MC DA effectiveness to be realized.

3.3.2 ERUM Generalization of the RRUM

Modeling Details Let a given DA have D attributes as its latent space. Such attributes must be carefully selected and can include both skills and misconceptions. Ideally, these attributes were defined as a first step in a DA test design. However, the attributes could be defined as the result of an analysis of an existing assessment being retrofitted as a DA, which may still be useful even if less desirable. For simplicity, ERUM assumes dichotomously coded attributes of thinking that a student either possesses or lacks. That is, for ERUM we do not (yet) model polytomous attribute levels, nor a continuous residual completeness factor as in the RUM, noting 3.2.

The latent space for a given set of D attributes can be represented by the set of all 2^D vectors $\boldsymbol{a} = (\alpha_1, \ \ldots \ , \alpha_D)$, where $\alpha_d = 1/0$ according as the student possesses/lacks, respectively, attribute d. Note the shift in language from "master/non-master of a skill" to "possession/non-possession of an attribute". Whether $\alpha_d = 1$ is encoded as increasing or decreasing option attractiveness or not depends on whether α_d is a desirable or problematic component of thinking combined with whether the option is a correct or incorrect option. Importantly, it may be or likely will be the teacher's focus to identify misconceptions possessed as much as or more than detecting skills mastered. Clearly correct/incorrect scoring and construction of MC tests with incorrect options that are not designed from this FA perspective (i.e., the detection of misconceptions being instructionally desirable) are both seriously suboptimal.

3.3.3 Expanded Definition of Q

The required ERUM Q matrix is a generalization of the usual DCM Q in two vital ways. First, Q must have a row for each response option of each item, instead of just one row per item as in the dichotomous scoring case. The Q matrix *link vector* for option $h = 1, \ \ldots \ , H$ of item k is the (k, h) row $\boldsymbol{q}_{kh} = (q_{kh1}, \ \ldots \ , q_{khD})$ of Q. So, the Q matrix $\{\boldsymbol{q}_{kh}\}$ for a test with 30 items for which each item has 4 options will have $30 \times 4 = 120$ rows and D columns. Second, if incorrect options and misconceptions are to be well modeled, each Q matrix entry for the ERUM needs to be any of *three* possible values, chosen to be 0/1/N for convenience: The latent states $\boldsymbol{a}$ that are *cognitively most strongly attracted* to option h of item k must satisfy

for each $d = 1, \ldots, D$ *for which* $q_{khd} \neq N$ *that* α_d *must match* q_{khd} *(i.e.,* both $= 0$ or both $= 1$). In other words, to maximize probabilistic attractiveness:

- if $q_{khd} = 0$ then the student *must lack* attribute d, i.e., $\alpha_d = 0$,
- If $q_{khd} = 1$ then the student *must possess* attribute d, i.e., $\alpha_d = 1$, and
- If $q_{khd} = N$ then α_d regardless of its value does not affect the strength of attraction to option h.

The key difference between this enhanced Q and the usual DCM Q is the introduction of *lacking an attribute* (in addition to the usual possessing of an attribute) as increasing or decreasing the relative attractiveness of an option. For example, lacking a misconception can make the correct answer *more* attractive and lacking a skill can make a particular incorrect option *more* attractive. Similarly, lacking a misconception can make a particular incorrect option *less* attractive and lacking a skill can make a particular incorrect option *more* attractive.[2]

One clerical detail becomes important here: Notice that the value $q_{khd} = 0$ for ERUM (*must lack* attribute d) has a very different interpretation from that of $q_{kd} = 0$ (attribute d value has no influence) in the dichotomous DCM case, namely $q_{khd} = 0$ means the lacking of attribute d makes option h more attractive. $q_{khd} = 1$ has the same meaning in both cases. An ERUM Q value N has the same meaning that 0 has in the traditional dichotomous Q DCM case discussed in many other chapters in this handbook, namely $q_{ihk} = N$ means that neither possessing nor lacking attribute k matters for option h attractiveness.

3.3.4 The ERUM IRF

Given the forced choice (one option must be selected) imposed by the standard MC format, ERUM posits that a student's strategy for responding to an item can be categorized as one of two types: "cognitive" or "guessing". A *cognitive strategy* means the student's possession of or lack of the various components of her $\boldsymbol{a}$ guide the option selection via a reasoned pattern of thinking. A "guessing" strategy occurs when the student selects randomly from the available options with equal probability.[3] The ERUM is a mixture model of these two strategies, as follows. Suppressing item index k, the ERUM's *cognitive kernel* $F_{ERUM,\,h}(\boldsymbol{a})$ *is* defined by:

$$F_{\mathrm{ERUM},h}\,(\mathbf{a}) = \pi_h \prod\nolimits_{d:q_{hd}\neq N} r_{hd}^{|q_{hd}-\alpha_d|}. \tag{3.3}$$

[2]It is perhaps mathematically equivalent to let misconceptions be represented by the negation of skills. We avoid this because it seems cumbersome for users.

[3]Restricted options guessing has been modeled but is not discussed here.

Then the ERUM item mixture model IRF is given by

$$P(h|\mathbf{a}) = \frac{F_{\text{ERUM},h}(\mathbf{a})}{S_{\mathbf{a}}}\omega_{\mathbf{a}} + \frac{1}{H}(1-\omega_{\mathbf{a}}) = \begin{cases} F_{\text{ERUM},h}(\mathbf{a}) + (1-S_{\mathbf{a}})\dfrac{1}{H} & \text{if } S_{\mathbf{a}} < 1 \\ \dfrac{F_{\text{ERUM},h}(\mathbf{a})}{S_{\mathbf{a}}} & \text{if } S_{\mathbf{a}} \geq 1 \end{cases} \tag{3.4}$$

where

$$S_{\mathbf{a}} \equiv \sum_{h=1}^{H} F_{\text{ERUM},h}(\mathbf{a}); \qquad \omega_{\mathbf{a}} = \min\{1, S_{\mathbf{a}}.\}.$$

Note that both $F_{ERUM,\,h}(\boldsymbol{a})$ and $P(h|\,\boldsymbol{a})$ for each item form a matrix of dimensionality (# options) $\times 2^D$.

In some sense, the guessing strategy occurs when the options combine to be relatively unattractive. Then examinee responding is a mixture of a weak cognitive strategy overlaid with some guessing. For example, taking H = 3, if a column of F is $\left(\frac{1}{4}\ \frac{1}{8}\ \frac{1}{16}\right)'$ then $S = \frac{7}{16}$ and the student augments this weak cognitive signal to produce the P column $\left(\frac{1}{4}+\frac{9}{48}\ \frac{1}{8}+\frac{9}{48}\ \frac{1}{16}+\frac{9}{48}\right)$.The cognitive kernel captures how the examinee would respond to each option if there was no interference (guessing or competition) from the other options, that is if $S_{\boldsymbol{a}} = 1$ for every $\boldsymbol{a}$. Note $\boldsymbol{a}$ that the model assumes that the mixture *cognitive probability* $\omega_{\boldsymbol{a}}$ is controlled by the item specific cognitive kernel $F\ H \times 2^D$matrix's column sum $S_{\boldsymbol{a}}$. The intuition is that if the combined cognitive attractiveness of the various options is weak then guessing (G: $S_{\boldsymbol{a}} < 1$) is likely to occur and this guessing probability is then parsimoniously modeled by the RHS's second summand in the first line of Eq. 3.4. Note, of serious consequence, that if the opposite occurs and the combined attractiveness of the options is cognitively strong, "competition" (C: $S_{\boldsymbol{a}} > 1$) occurring among the options, then the natural attractiveness of each option h is multiplicatively diluted by the sum of their attractiveness, i.e., via $1/S_{\boldsymbol{a}}$ in the second line of 3.4. This dilution implies that excessive competition or excessive guessing both degrade the diagnostic power of the item to diagnose examinee $\boldsymbol{a}$**,** and it is a strength of the ERUM model that it validly captures this important C/G degradation aspect of MC testing.

3.3.5 Identifiability of the ERUM[4]

Non-identifiability ("*nonid*") for an item is defined as holding if there exist item parameterization values $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$ that both produce the same model IRF model P

[4]This somewhat technical section can be skipped over, but with the caveat that identifiability is endemic to DCM option scored MC modeling, and this section helps explain this perhaps surprising claim.

in Eq. 3.4. Whereas identifiability ("*id*") always holds for RRUM and, closely associated, the RRUM item parameters are nicely interpreted as discussed above, by contrast, nonid and hence non-interpretability may hold for some ERUM models, this in an algebraically deep way. There is not space, nor is it appropriate given the focus of this article, for a deep analysis of this nonid to be given herein (a thorough discussion of the issue is being submitted elsewhere in the near future). Certain aspects of this ERUM nonid illuminate and relate to (i) statistical inference of the ERUM and user interpretation of its estimated parameters and (ii) the intrinsic nature of nonid when well modeling MC examinee behavior at the option level. Both (i) and (ii) are important from the user perspective. First, given a reasonable sized dataset (that is, sufficient number of examines and of items to produce effective estimation of identifiable parameters), the ERUM model IRF probabilities $P_k(h|\boldsymbol{a})_{h,\,k,\,\boldsymbol{a}}$ *can always be* well estimated because *the item model Ps are always identifiable*. The hope would then be that the ERUM model's item parameters $\boldsymbol{\beta} \equiv \{\boldsymbol{\pi}, \boldsymbol{r}\}$ (see Eqs. 3.3 and 3.4) would also all be id and hence well estimated. But ERUM parameter id is simply not the case in general!

The reason for this is informative, not only statistically but substantively: First, for a fixed item we note from Eq. 3.4 it follows for every $h \neq h'$ pair and every latent $\boldsymbol{a}$ that one of the following two relationships holds between model P and the cognitive kernel F:

$$\frac{F_{ERUM,h}(\boldsymbol{a})}{F_{ERUM,h'}(\boldsymbol{a})} = \frac{P_{ERUM,h}(\boldsymbol{a})}{P_{ERUM,h'}(\boldsymbol{a})} \text{when } S_{\boldsymbol{a}} \geq 1 \text{ (competition, } C\text{, holds) and}$$

$$\begin{aligned} F_{ERUM,h}(\boldsymbol{a}) - F_{ERUM,h'}(\boldsymbol{a}) &= P_{ERUM,h}(\boldsymbol{a}) \\ &\quad - P_{ERUM,h'}(\boldsymbol{a}) \text{ when } S_{\boldsymbol{a}} < 1 \text{ (guessing, } G\text{, holds)}. \end{aligned} \tag{3.5}$$

Equation 3.3 shows how a given set of item parameter values $\boldsymbol{\beta}$ produces the cognitive kernel F appearing in Eq. 3.4. Each actual MC item has its own unique true F. We have mathematically proved elsewhere that $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$ translates to their corresponding pair of cognitive kernels satisfying $F \neq F'$. So nonid holding for some specified $\boldsymbol{\beta}$ parameterization is thus *equivalent to* nonid holding for the cognitive kernel F.That is, nonid holding implies one can have different items, thus each with their own unique different F, yet all having the same IRF P of Eq. 3.4. That is, nonid is intrinsically about the relationship of the cognitive kernel F to the overall item probability model P. As a consequence, informally put, the data, while telling us what the item's model P is at least approximately, sometimes cannot tell us anything about which item cognitive structure F has produced it.

More deeply, nonid is caused by the fact that in the $F \rightarrow P$ ERUM model specification process given by Eqs. 3.3 and 3.4, the true cognitive structure F gets distorted and hence masked by the intrinsic forced competition and/or forced guessing among options that occurs when an actual examinee responds to an actual MC item. Specifically, when for Examinee $\boldsymbol{a}$, $S_{\boldsymbol{a}} \neq 1$, then F gets modified (distorted) via Eq. 3.5 to produce P. All this suggests to the authors that other option

scoring well-fitting MC DCMs, such as EDINA (DiBello et al., 2015), will have nonid analogous to ERUM nonid occurring.

Finally, when parametric nonid is discovered, namely multiple $\boldsymbol{\beta}$ producing the same P for at least one of the items of the test, we then know that there are multiple sets of item parameters capable of producing the item's IRF P and we cannot know from the data which one is the *true* set of item parameters that produced the data. And, it thus follows that estimated item parameter interpretation becomes impossible for nonid parameters. To be absolutely clear, suppose we do a simulation of a nonid item using a particular "true" item kernel F with ERUM data produced via the simulation. Then when we MCMC estimate the item's ERUM model from this simulated data, we plausibly could infer that a substantively different item (a different kernel $F^{'}$ and thus a corresponding different $\boldsymbol{\beta}^{'}$ that also would produce the simulation's model P) produced the data, noting all such admissible Fs here seem equally plausible from the data-driven statistical inference perspective. [Statistically speaking, MCMC estimation of $\boldsymbol{\beta}$ and hence of F, may simply fail of course because of nonid, which the RUM system thus must deal with, as explained below.] This is sufficient background for the reader to understand how nonid comes about for ERUM. However, noting this discussion is quite abstract, we give a simple concrete numerical example of a nonid ERUM item next.

Example Let $H = 2$ and $D = 2$ (for simplicity and clarity) and let $\mathrm{Q} = \begin{pmatrix} 0 & 1 \\ N & 0 \end{pmatrix}$. Let the $\boldsymbol{a}$ column order for F, P be 11, 10, 01, 00. For convenience replace the double subscripted r_{hd} by r_d, s_d.Then

$$F = \begin{matrix} \pi_1 r_1 & \pi_1 r_1 r_2 & \pi_1 & \pi_1 r_2 \\ \pi_2 s_2 & \pi_2 & \pi_2 s_2 & \pi_2 \end{matrix}.$$

Suppose the 5 parameter values $\boldsymbol{\beta} = (\pi_1, \pi_2, r_1, r_2, s_2)$ are such that the column sums $S_{\boldsymbol{a}} > 1$ for all four $\boldsymbol{a}$, the CCCC column case for F. The id question is whether "knowing" the 8 $P_{hj} (h = 1, 2; j = 1, 2, 3, 4)$ values, either because P is known as in a simulation study or because P is well estimated, determines the five parameter values and thus id would hold. First, via Eq. 3.5, these eight P values only determine F via determining the four $\frac{F_{hj}}{F_{h'j}} = \frac{P_{hj}}{P_{h'j}}$ equations. Thus, there would appear to be four P-determined ratio-based constraints on F and hence on $\boldsymbol{\beta}$, leaving 1 *degrees of freedom* (*DF*). Not true!: First, the determined third column ratio and the first column ratio combine to determine $\frac{\pi_1 r_1}{\pi_2 s_2} / \frac{\pi_1}{\pi_2 s_2} = r_1$. Out of the four column ratios, we first find two algebraically independent determined ratios: $\frac{\pi_1}{\pi_2 s_2}$ and $\frac{\pi_1 r_2}{\pi_2}$. There are no more independent expressions! Because r_1 is known, the second and fourth column ratios contribute only one independent determined ratio. We thus have five parameter unknowns with three independent column ratios determined constraints, producing $DF = 5 - 3 = 2$. (DF being the dimensionality of the nonid space) Thus the nonid $\boldsymbol{\beta}$ space is two dimensional.

This is aptly illustrated by giving two distinct sets of $\boldsymbol{\beta} \rightarrow \boldsymbol{\beta}^{'}$ transformations that produce the same model P. First, let for some $c > 0$

$$\pi_1' = c\pi_1, \pi_2' = c\pi_2 \text{ with } r_1, r_2, s_2 \text{ held constant.}$$

We claim this transformation leaves the model's P values unchanged. To see this, first note by Eq. 3.5 that $P(h|\boldsymbol{a}) = \frac{F_{ERUM,h}(\boldsymbol{a})}{S_{\boldsymbol{a}}}$ for all four columns. Consider Column 2 (the 10 column). $S_{10} = \pi_1 r_1 r_2 + \pi_2$. Then, taking $h = 1$ for example, $P(1|10) = \frac{\pi_1 r_1 r_2}{\pi_1 r_1 r_2 + \pi_2}$. Thus, substituting $\pi'_1 = c\pi_1$, $\pi'_2 = c\pi_2$, (rescaling the π) with r_1, r_2, s_2 held constant,

$$P'(1|10) = \frac{c\pi_1 r_1 r_2}{c\pi_1 r_1 r_2 + c\pi_2} = \frac{\pi_1 r_1 r_2}{\pi_1 r_1 r_2 + \pi_2} = P(1|10).$$

In this manner, we see the π are scaling-wise nonid because $P = P'$. This is typical of course of nominal scoring models where division by column sum occurs and hence a resulting scaling nonid occurs.

But there is a second source of nonid that is demonstrated by a second $\boldsymbol{\beta} \to \boldsymbol{\beta}'$ distinctly different transformation that leaves the model's P values unchanged. Let for some $c' > 0$ that preserves the all columns C structure

$$\pi_2' = c'\pi_2, \quad r_2' = c'r_2, \quad s_2' = \frac{1}{c'}s_2, \pi_1' = \pi_1, r_1' = r_1$$

To see that nonid holds, consider $P'(1|11) = \frac{\pi_1' r_1}{\pi_1' r_1 + \pi_2' \frac{1}{c'} s_2} = \frac{1}{1 + \frac{\pi_2'}{\pi_1'} \frac{s_2}{r_1 c'}} = \frac{1}{1 + \frac{\pi_2}{\pi_1} \frac{s_2}{r_1}} = P(1|11)$, etc. Clearly this second transformation, without giving the justification that produced it, has a certain opaque quality about it. But it *is* nonetheless a simple consequence of our general ERUM nonid theory (applies for all H, D, Q; to be submitted soon). Let's do a numerical reification: Let $c' = \frac{3}{4}$. $\pi_1 = \pi_2 = 1$ *and* $r_1 = r_2 = s_2 = \frac{2}{3}$. *Then* $\mathrm{F} = \begin{matrix} \frac{2}{3} & \frac{4}{9} & 1 & \frac{2}{3} \\ \frac{2}{3} & 1 & \frac{2}{3} & 1 \end{matrix}$ and $F' = \begin{matrix} \frac{2}{3} & \frac{1}{3} & 1 & \frac{1}{2} \\ \frac{2}{3} & \frac{3}{4} & \frac{2}{3} & \frac{3}{4} \end{matrix}$, noting both have all C columns. Note, by direct substitution that $P = \frac{F}{S} = \frac{F'}{S'}$ and hence both F and F' produce the same P.

Because the interpretation of a model's parameters is always important, we consider the issue of the user interpretation of ERUM item parameters when nonid holds: In general, a model is *well parameterized* if distinct parameterizations $\boldsymbol{\beta} \neq \boldsymbol{\beta}'$ produce distinct substantive interpretations, or, simply put, the parameter values always have intrinsic (substantive) meaning. Interestingly**,** the ERUM *is* well parameterized by $\boldsymbol{\beta} = (\boldsymbol{\pi}, \boldsymbol{r})$ even though nonid can hold. Nonid simply means that the data does not reveal the unobservable, *but still substantively meaningful, cognitive structure* that would hold if there was no guessing or competition (if all the item's $S = 1$ actually, forcing $P = F$) among the options. So, this provides a clear interpretation of the ERUM $\boldsymbol{\beta}$ (even if we sometimes cannot estimate it because of nonid!). $\boldsymbol{\beta}$ simply quantifies the item cognitive structure F we would observe if there was no interaction among the item's options, which never happens in practice of course. But, the whole point, $\boldsymbol{\beta}$ still has substantive meaning! But, we cannot estimate $\boldsymbol{\beta}$ and moreover its virtual meaning is useless to the user.

Fortunately, when ERUM is assumed for the RUM diagnostic system, an alternative identifiable $\boldsymbol{\beta^0} = (\boldsymbol{\pi^0}, \boldsymbol{r^0})$, and hence estimable from data, parameterization has been derived. This alternative parametrization $\boldsymbol{\beta^0}$ has essentially the same useful interpretation as the 0/1 scored RRUM $\boldsymbol{\beta} = (\boldsymbol{\pi}, \boldsymbol{r})$ parametrization described in the paragraph following Eq. 3.1. Thus, importantly, evaluation of ERUM modeled item classification quality *is* possible via this alternative ERUM parameterization even though the RRUM $\boldsymbol{\beta} = (\boldsymbol{\pi}, \boldsymbol{r})$ parametrization is sometimes not identifiable and its estimation will in general fail. Space constraints preclude a detailed description of the definition of this user-interpretable identifiable transformation, which allows users to evaluate item quality via well estimated $(\boldsymbol{\pi^0}, \boldsymbol{r^0})$ ERUM parameter values.

Now we consider MCMC estimation of $\boldsymbol{\beta}$ in the face of nonid. Clearly, if nonid holds then to successfully estimate $\boldsymbol{\beta}$ we need to constrain one or more $\boldsymbol{\beta}$ components so $\boldsymbol{\beta}$, and hence F, then becomes *id*. The number of such components of $\boldsymbol{\beta}$ required to be constrained is labeled the *degrees of freedom* (*DF*). *DF*[5] is thus the dimensionality of the *nonid* indeterminacy. Thus nonid holds if and only if $DF > 0$. Our ERUM id producing MCMC item parameter estimation algorithm carefully constrains, as it must, *DF*[5] components of $\boldsymbol{\beta}$ for each nonid item, noting that which components are constrained matters to achieve id estimation success. Suppose for an item that the *P*s are determined (i.e., known), or at least well estimated from the data. The algebraic structure of the ERUM model produces an algebraically deep theory concerning the DF constraints needed to produce id, as the above example suggests. Here we simply state the two structural entities that control the *DF*, sometimes producing $DF = 0$, namely id, and sometimes $DF > 0$, namely nonid, first noting via Eq. 3.5 that knowing the item's $H \times 2^D$ values determines only $(H - 1) \times 2^D$ *of the* $H \times 2^D$ *F* values, opening up the possibility of nonid.

However,

- The columns of the matrix F that are G (i.e., $S_{\boldsymbol{a}} < 1$ for Column $\boldsymbol{a}$), as determined by the parameter values of $\boldsymbol{\beta}$, each have their H component values determined by the either known or well estimated P, thus reducing the DF, as desired. For example, for a determined model P (well estimated say) if πr is in a G column then πr is determined, but not necessarily its separate values.

Similarly, but in a more complex manner:

- *N*s in Q determine, via the "known" *P*s, certain parameter values of $\boldsymbol{\beta}$ in ways too complex to describe herein, thus reducing the DF, as desired.

Thus, guessing columns occurring in F and *Ns* occuring in Q each determine additional parameter values and thus reduce the *DF* and hence either reduce the

[5] *DF* has a mathematical meaning: the number of elements that need to be constrained to uniquely determine the full vector. That is how *DF* is used herein. Note here however that $DF > 0$ is a bad thing that has to be changed to $DF = 0$. This is contrary to many statistical applications where $DF > 0$ is beneficial, such as in linear models where $DF > 0$ allows the size of the residual variance to be estimated and estimated well when $DF > 0$ is large.

dimensionality of the nonid $\boldsymbol{\beta}$ space or even cause id to hold. Specifically, enough G columns and/or enough Ns in Q produce id, as desired.

3.4 The GDCM Generalization of the ERUM

Other cognitive kernels besides in Eq. 3.3 can be used in our option scored DCM framework, which we refer to as Generalized Diagnostic Classification Models (GDCM). For example, various compensatory models, such as a compensatory version of the RRUM can be introduced. Parametrically simpler models, such as the DINA could also be used, producing the Extended DINA (EDINA) form of the GDCM:

$$F_{\text{EDINA},kh}(\mathbf{a}) = \begin{cases} 1 - s_{kh} & \text{if } \alpha_d = q_{khd} \text{ for all } d \text{ such that } q_{khd} \neq N \\ g_{kh} & \text{if there exists a } d \text{ such that } q_{khd} \neq N \text{ and } \alpha_d \neq q_{khd} \end{cases}. \quad (3.6)$$

Here, the DINA-based cognitive kernel converts the DINA- matching of required skills to matching all non-N elements of $\{q_{khd}\}$, with "slip" parameters s_{kh} and "guessing parameters" g_{kh}. The non-negative kernel function $F_{EDINA,\,kh}(\boldsymbol{a})$ results in an "all-or-nothing matching of all non-N elements" for EDINA that is analogous to DINA's $P(X_k = 1|\,\boldsymbol{a})$ determined via the matching or non-matching in $\boldsymbol{a}$ of all the required item k skills. Note that nonid can hold for other GDCM mixture models similarly as it does for ERUM. Indeed, largely undeveloped herein, nonid seems a problem for non GDCM option scoring-based DCM models.

3.5 Model-Data Checking of the RUM/ERUM

Some Introductory Remarks "Model-data checking" refers to both model-data-fit and discrimination investigations, often involving specific indices. Both fit and discrimination are qualities of the model that should be examined in a thorough RUM system application. All the fit and discrimination approaches described in this section can be applied to both RRUM and ERUM. However, in each case, we explain via either RRUM or ERUM, with the understanding that each approach also transfers to the other model.

Fit A model can fit the data poorly for two main reasons, lack of theoretical model fit or poor model estimation due to insufficient data or a poor choice of the statistical estimation procedure (to illustrate the latter, even if a perfect fitting quadratic regression is chosen, failing to adjust for overly influential data points can for some data sets produce poor estimation of the true regression line). Thus, the performance of the statistical methodology (and even the quality of the data) is a component of model-data fit assessment. Simply put, a good fit requires a good model choice, a good choice of statistical methodology, and good data. Thus

fit indices and approaches address both model and methodology, and involve the data too. DCMs have several major components: the latent space $\{\boldsymbol{\alpha}\}$, as specified in the ERUM case by the listing of the major skills and misconceptions assumed to influence examinee responding, "design matrix" Q, the test's IRFs as specified by item parameters, and the latent space distribution $P(\boldsymbol{\alpha})$, possibly parameterized. Thus, in order to claim a good model fit *for the intended purpose of DCM analysis namely examinee diagnosis and/or test design and evaluation*, the fit of Q, of the IRFs, and even of $P(\boldsymbol{\alpha})$ all can come into play.

In this regard, the importance of the latent class distribution $P(\boldsymbol{\alpha})$ needs clarifying. First, a goal of the inference itself can be to know certain aspects of $P(\boldsymbol{\alpha})$. Such knowledge can be used to learn how the various skills and misconceptions are associated with each other in the population of test takers. Further, if the test-taking population is of user focus, then $P(\boldsymbol{\alpha})$ provides aggregated information on classroom learning. Second, even if inference about $P(\boldsymbol{\alpha})$ is not of importance to the user, estimating $P(\boldsymbol{\alpha})$ accurately will improve both IRF estimation and examinee classification. In this regard, the RUM system does estimate $P(\boldsymbol{\alpha})$. Related, our MCMC approach casts the RUM family of models in a Bayesian framework, which requires defensible specification of a prior for the model's parameters, which includes the parameterization of $P(\boldsymbol{\alpha})$.

Next some of the RUM system's fit methods are summarized. For more detailed descriptions, including some applications to real data settings, the reader is directed to Roussos et al. (2007), DiBello et al. (2015), Shear and Roussos (2017), and the DiBello and Stout (2008) RUM system Arpeggio software manual. Related, Chap. 13 (Han & Johnson, this volume) discusses fit for DCM models in general.

A MAD (Predicted − Observed) Based Index of Fit (MAD = Mean Absolute Deviation). We consider the ERUM setting, with the simplification to the RRUM setting obvious. Our $D_{k,\,h}$ fit index measures estimated model predicted IRF misfit with the data at the individual item, option level. Viewed over all items k and options h, if the fit is good, the estimated DCM IRF probabilities $P(X_k = h \mid \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})$ should typically deviate at most moderately from the corresponding observed proportions $\frac{\mathrm{N}_{\widehat{\boldsymbol{\alpha}},k}(h)}{\mathrm{N}_{\widehat{\boldsymbol{\alpha}}}}$, where $\mathrm{N}_{\widehat{\boldsymbol{\alpha}}}$ is the number of examinees who were classified as $\widehat{\boldsymbol{\alpha}}$ via the MMAP (Marginal Maximum a Posteriori) classifier (or other) using the fitted model. Here $\mathrm{N}_{\widehat{\boldsymbol{\alpha}},k}(h)$ denotes the number of examinees classified as $\widehat{\boldsymbol{\alpha}}$ who selected response option h for item k. Our $D_{k,\,h}$ index developed from these building blocks is a weighted average of MADs using the proportion of examinees in each $\widehat{\boldsymbol{\alpha}}$ category as weights,

$$D_{k,h} = \sum\nolimits_{\widehat{\boldsymbol{\alpha}}} \frac{\mathrm{N}_{\widehat{\boldsymbol{\alpha}}}}{N} \left| P(X_k = h \,|\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}) - \frac{\mathrm{N}_{\widehat{\boldsymbol{\alpha}},k}(h)}{\mathrm{N}_{\widehat{\boldsymbol{\alpha}}}} \right|. \tag{3.7}$$

Note that a particular $D_{k,\,h}$ may be large for any of three intertwining reasons: model bias (i.e., poor model choice), model estimation error, and/or classification error. Because we want to detect lack of fit caused by the presence of any combination of these, this item option specific set of indices seems to provide a good overall

assessment of fit. Note that via simulation studies, these three sources can be disentangled, for example via substituting the true $\boldsymbol{\beta}$ for $\widehat{\boldsymbol{\beta}}$ and thus isolating the performance of the MMAP classification procedure (our usual classification method choice).

AIC and BIC Standard indices of fit, particularly AIC and BIC, are available for both RUM and ERUM. For example, see Shear and Roussos (2017) geometry DA ERUM study.

MCMC Chain Convergence Assessment As remarked, poor statistical performance will produce lack of fit. MCMC is at the core of the RUM system's statistical approach. Thus, poor RUM system MCMC convergence will produce poor fit. Lack of MCMC chain convergence can have various causes including a poor MCMC design (such as having an inadequate burn-in period), a badly fitting model, or just poor data. The point is that good chain convergence is a necessary condition for a successful RUM system application and hence must be checked as part of the RUM System Fit protocol. An informal visual check of chain graphs is recommended, with lack of drift, sufficient noisiness (jiggle), etc., all indicators of appropriate convergence (Gilks, Richardson, & Spielgelhalter, 1996). Beyond visual inspection of the chains, Gelman and Rubin (1992) defined the widely used index R to measure the extent of convergence when using multiple chains, this multiple chain with index R computation being built-in option for the RUM system. It is recommended that both visual inspection and index computation(s) be used to assess convergence. In the software section of the chapter we note that specific approaches have been developed to assess convergence of MCMC chains. The reader is also referred to Chap. 31 on MCMC estimation of CDMs (Liu & Johnson, this volume)

Q Fit An important topic is the issue of the creation, estimation, and fit of the Q matrix, as discussed in Chap. 12 (Liu & Kang, this volume). The estimation of Q is becoming a topic of interest in the literature; see for example Chen, Culpepper, Chen, and Douglas (2018). One could develop fit indices tuned to our option scoring based Q matrix selection, but this has not been done yet. Of particular interest, competing substantive explanations for the latent space to be chosen can produce different Qs, and one could compare their relative fit to help assess the relative validities of competing explanations. The interested reader is referred to Shear and Roussos (2017), a geometry DA study for an ERUM based example of this approach.

Uninfluencial IRF Model Parameter Elimination One useful interaction between Q, model fit, and the item parameter estimated values is that large r estimates ($\widehat{r} \approx 1$) suggest for RRUM that the corresponding 1 in Q should be replaced by a 0 (or 0 or 1 replaced by N if doing an ERUM analysis), thus accepting the data driven conclusion that the skill is not affecting that item's responding and thus that the corresponding r be eliminated from the item's ("option" in the ERUM case) modeling. The reader is directed to Roussos et al. (2007) for a thorough discussion of this suggestion.

Even if an r does not satisfy $\widehat{r} \approx 1$ for an item (or option in the ERUM case), it is still possible that the corresponding attribute's influence on the option is small. Notice that this idea is analogous to eliminating variables in multiple regression that have little predictive influence even when their coefficients are statistically judged "different" from 0. Such a situation would suggest that changing the Q matrix to eliminate the parameter's role upon the option would not affect the fit of the model. Such model complexity reduction both can produce a more interpretable model and further make estimation better and classification easier. See Roussos et al. (2007) for details in the RRUM case.

Discrimination For the RUM diagnostic system, "discrimination" refers to indices and methodologies that assess various aspects the system's ability to classify examinees accurately. It has two distinct aspects: providing information to improve DAs at the design and development (pretest) stage and providing information to users about diagnostic strengths and weaknesses of DAs at the actual testing stage. Thus, importantly, in addition to evaluating an existing DA, discrimination indices can assist in test design, allowing the identification, with subsequent improvement or elimination, of weakly performing items, even at the option level where individual options can be modified to improve their capacity to detect attribute possession versus non-possession.

Note that discrimination indices function at the test level and the individual item level, the latter including sometimes at the individual option level. Similarly, discrimination indices function either at the overall latent space level (i.e., joint or vector level) or at the individual attribute component level. Also, these two foci of discrimination, attribute and item, can interact to assess discrimination. For example, an index could address how well an item k is functioning with respect to the attribute it is best discriminating about (see L'_k in Eq. 3.8 below). Correct classification rates, joint (CCR) and marginal (CCRd), are the most important indicators of test classification performance from the user perspective, in part because they directly measure classification performance rather than producing indices that correlates well with performance:

Notation CCR (i.e., $P(CC)$) denotes the correct classification (CC) rate for the entire attribute vector, that is, correctly classifying the examinee on all D components simultaneously. Importantly this is in contrast to CCRd (i.e., $P(CC_d)$), denoting the correct classification (CC_d) rate for attribute d.

$P(CC)$, $P(CC_d)$, $P_k(CC)$, $P_k(CC_d)$ as Discrimination Indices As with certain other discrimination indices, correct classification rates can be at the individual attribute component level or for the entire attribute vector and also can be used for the entire test or item by item. This produces four kinds of indices, as the notation captures. Computing item by item or attribute by attribute both can be especially useful in the design phase, for example improving a test by modifying or eliminating weakly performing items or by adding items that measure certain attributes well. For the user it can be helpful to know how well various attributes are being measured and by which items. Formulas have been derived that make these computations feasible

given the estimated model, especially for individual items (unpublished, available from authors). When easy to estimate, even approximately, these probability-scaled indices are highly useful.

Marginal vs. Joint Correct Classification Rates Clearly both CCR (joint) and CCR*d* (marginal) are of interest. However, from the instructional perspective, teachers usually want to know separately attribute by attribute which skills and misconceptions are possessed and not possessed. Thus, $P(CC_d)$ and $P_k(CC_d)$ are usually the more important discrimination indices from the classroom perspective.

Other Customized Item and Attribute Specific Indices of Discrimination One might wish to assess discrimination power without the computational challenge of estimating $P(CC_d)$s or $P(CC)$, which can sometimes be hard to estimate when test length K is large (something being worked on). Moreover, for a specific item, one may wish to target a single option. Or, one may wish to use a globally informative graphical approach. Thus even though correct classification forms the core of our discrimination approach, other indices and approaches are of value and are part of the RUM system discrimination approach. We discuss several of these:

IMSTATS When using the RRUM, a set of statistics, called Item Mastery Statistics (*IMSTATS)* by Roussos et al. (2007), is used to quantify various discriminatory aspects of an item. This customized and graphically aided approach to discrimination posits that a well-designed item will differentiate well between examinees who have mastered the item's required skills, named *item masters*, and those who have not, named *item non-masters* (for ERUM, the modified IMSTATS deals with individual option matching vs. non-matching as determined by Q) and other DCMs are natural extensions. For every item k, IMSTATS compares the item correct proportions for the item masters versus the item nonmasters. If a set of items collectively are effective (a substantive assessment goal), and if the estimated model fits the data at least moderately well (which indicates that fit plays a role in this item level index of discrimination), then for every k item k masters should be performing noticeably better than item k nonmasters.

Note that a test with little diagnostic information concerning the targeted latent attribute space could be fit well and yet the IMSTATS results would still be weak. Figure 3.1 below shows an IMSTATS graph for an English as a Second language (ESL) FA test analyzed using the RUM system (Jang, 2005). Two things are worth noting: The fit seems good with the average masters versus nonmasters item proportion correct difference being approximately $+0.4$ and further for most items the positive difference being substantial. Further, notice that this provides a valuable tool for finding poorly performing items, for example Items 5, 6, 11, 14, 16, 17, 21, 27 and 30 have small differences, some even reversals, due either to poor item-specific discrimination or, possibly, bad fit. Also, some items perform exceptionally well such as Items 24 and 26.

EMSTATS Examinee Statistics (called EMSTATS) can also be computed in an approach analogous to IMSTATS but the focus is shifted to examinees. It detects

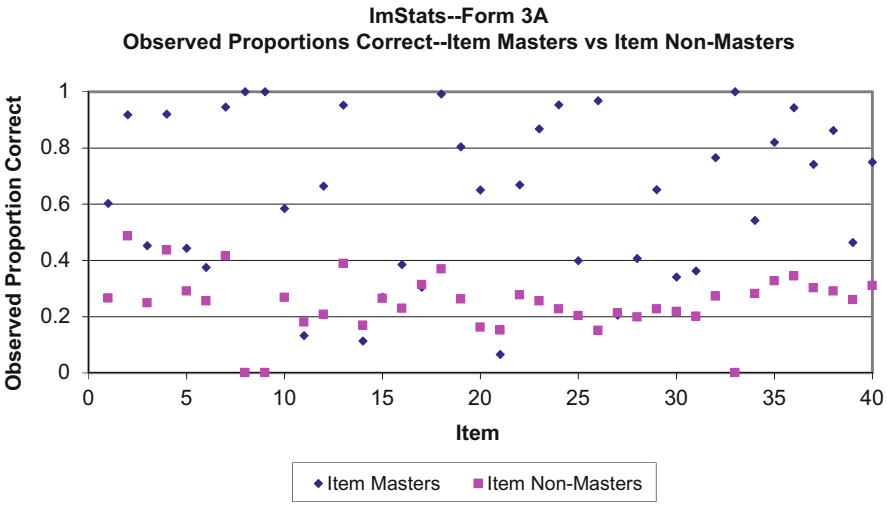


Fig. 3.1 IMSTATS for an ESL Test

examinees for whom discrimination is poor or the fit is poor. Details about the approach are given by Roussos et al. (2007).

The Henson Discrimination Index We consider the Henson, DiBello, and Stout (2018) ERUM discrimination index (as modified by Shear & Roussos, 2017) $L'_{k,d}$ that measures the discrimination power of an item k to measure an attribute d well. We also consider the index L'_k that measures the discrimination power of an item to well measure the attribute it measures best:

$$\begin{cases} L'_{k,d} = \max\limits_{h} \left\{ \dfrac{\Sigma_{\boldsymbol{\beta}_{-d}} \left[\left| \hat{P}\left(X_k = h | \alpha_d = 1, \boldsymbol{\beta}_{-d}\right) - \hat{P}\left(X_k = h | \alpha_d = 0, \boldsymbol{\beta}_{-d}\right) \right| \right]}{2^{D-1}} \right\}. \\ L'_k = Max_d L'_{k,d} \end{cases} \tag{3.8}$$

In the above equation, $\boldsymbol{\beta}_{-d}$ denotes a $D-1$ dimensional attribute vector via the d^{th} attribute of $\boldsymbol{\beta}$ removed. That is, via Eq. 3.8, for each item by attribute combination, one maximizes over all options h the average over $\boldsymbol{\beta}_{-d}$ of the absolute difference in model-predicted probabilities of choosing each option h for examinees who differ only in whether they possess attribute d or not. The larger the index the better the discrimination of item k in diagnosing attribute d. The above summand differences (removing the max operator via inserting the h_d that maximizes $L'_{k,d}$)

$$\widehat{P}\left(X_k = h_d | \alpha_d = 1, \boldsymbol{\beta}_{-d}\right) - \widehat{P}\left(X_k = h_d | \alpha_d = 0, \boldsymbol{\beta}_{-d}\right)$$

thus compares the model-predicted probability of selecting option h_d for two examinees who have the same attribute profile except for differing on attribute d. Because there are 2^{D-1} possible different profiles when we exclude attribute d, this formula in effect uses the option h_d with the largest average absolute difference.

Stout & DiBello Option Specific Discrimination Index Another discrimination index, proposed by Stout and DiBello (unpublished), and available in the RUM System Software, assesses how well each option discriminates with respect to the attribute it is most discriminating of and is given by

$$L''_{(h,k)} = \text{Max}\left\{\max_d \frac{P\left(X_k = h|\alpha_d = 1\right)}{P\left(X_k = h|\alpha_d = 0\right)},\ \max_d \frac{P\left(X_k = h|\alpha_d = 0\right)}{P\left(X_k = h|\alpha_d = 1\right)}\right\}. \tag{3.9}$$

Software note: These fit and discrimination indices are currently available and further will be available in a "user friendly" SHINY R based version of the RUM system in Spring 2019.

3.6 Rum/ERUM Based Estimation and Classification Software

We discuss three software choices for researchers and practitioners wishing to use the RUM/RRUM/ERUM approaches discussed in this chapter, two of which will be combined as callable functions in a single SHINY interface-based R package with a release date in Winter 2018. This package will allow for RUM/RRUM and ERUM to be estimated, and further a small set of GDCM cognitive kernels such as EDINA.

Arpeggio RUM System The first choice is the RUM/RRUM using Markov Chain Monte Carlo (MCMC) based estimation and classification procedure named the Arpeggio Suite (DiBello & Stout, 2008). As background, estimation of the right/wrong scored RUM model, with inclusion of the continuous residual $P_c(\eta)$ component optional, was accomplished by Hartz (2001) using MCMC estimation and Marginal Maximum a Posteriori (MMAP; doing MAP marginally at the individual attribute level recall) classification and programmed in FORTRAN, our basic computational language. It is easily proved that MMAP is the optimal classification method in the sense that for each d it maximizes (PCC_d) over all possible classification procedures. The resulting MCMC parameter estimation algorithm uses Metropolis Hastings within Gibbs sampling. Further, a Bayes approach is adopted that places (non-hierarchical) uniform priors on all item parameters with range [0,1] (thus, including the values 0 and 1), but which could easily be converted to a hierarchical empirical Bayes treatment when the test's item parameter distribution justifies doing so.

In addition, the joint distribution of the latent attribute patterns, $P(\boldsymbol{\alpha}, \eta)$, was modeled hierarchically using a prior that assumed the discrete skills were the results of a discretized D-dimensional multivariate normal with mean vector $\boldsymbol{\mu}$, which

controls population mastery proportions, and tetrachoric correlation theory inspired correlation matrix $\boldsymbol{\rho}$. The resulting $(D+1) \times (D+1)$ correlation matrix coefficients represent the estimated tetrachoric correlations between any given pair of skills or the estimated biserial correlation between any skill with the continuous η trait in the RUM. The probability of mastery for any given attribute A_d is thus equal to the probability of its associated random latent component of the multivariate normal distribution being ≥ 0.

Via MCMC, post burn-in posterior means are used to provide item parameter estimates and the proportion of times that an examinee is identified as a master, post burn-in, is used to estimate the probability of mastery for any given attribute, this trivially convertible to classification via MMAP. As remarked above, convergence of the chains can be examined either through visual inspection or by use of available software such as the CODA package in R (Plummer, Best, Cowles, & Vines, 2006). After inspecting convergence, as detailed in Roussos et al. (2007), various customized fit and discrimination indices are available in Arpeggio, including IMSTATS and EMSTATS, as discussed briefly above in the Model-data Checking section. Concerning the user need to assess a test's classification performance, simulation-based estimates of skill α_d correct classification rates (CCRds) are available for all the D skills. In this regard, note that, contingent upon their estimation accuracy, the simulation-based estimated CCRds seem a very direct and interpretable way to assess classification discrimination.

In summary, the Arpeggio Suite contains software that allows for (i) estimation of the chosen model; (ii) computation of various measures of fit that can help assess fit and moreover improve on it via reductions or modifications of the estimated model (importantly, including the fit of the Q-matrix); (iii) classification of examinees using the "known" (i.e., well estimated) item parameters, noting the items might only need to be calibrated and as such to be available in a test or test pool intended for future classification usage; (iv) the ability to simulate responses from the estimated RRUM model; and (v) the computation of various discrimination indices. We note that most, but not all, of the real data studies reported on herein used Arpeggio. A detailed Arpeggio user manual is available (DiBello & Stout, 2008) and the Roussos et al. (2007) paper is very accessible. The ERUM model is also a callable part of the RUM system and as such has almost all the capabilities described above in the callable Arpeggio suite.

MPlus RUM Availability Since the publication of Arpeggio, efforts to make user estimation of DCMs widely available have resulted in software packages that allow for the estimation of various RLCMs. For example, the software package Mplus has been used to obtain estimates for various diagnostic models from the literature including the RRUM (see Chiu et al., 2016; Templin & Hoffman, 2013). The use of Mplus for estimation of RLCMs is also discussed in Chap. 28 (Sullivan, Pace, & Templin, this volume). Although the inclusion of the residual continuous ability measure (in the RUM) is not available in these frameworks, Mplus based DCM treatments do provide some advantages. For instance, Mplus is a general statistical software package in which examinee covariates can be included to help predict item

performance, improve CCRs, or more easily address questions of applied interest. In addition, measures of fit can be easily invoked, like AIC and BIC. Further, not an advantage necessarily but a difference, the estimation approach is MMLE (M for marginal, in *d*) using the EM algorithm rather than MCMC. However, one challenge is that more demanding model estimation such as when the test length is at all long or the number of skills is sizable can be quite time consuming, when compared to MCMC estimation using the Arpeggio Suite.

The SHINY R RUM/ERUM Package. (Winter 2018 Release) The third choice, which is beyond the use of command-based software and Mplus, are a new set of RRUM and ERUM MCMC based procedures that have been programmed using the R language. R is a widely used statistics programming computation and graphics language that allows users to create customized packages that can be used for a wide range of specialized statistical analyses. In the DCM context, we note that R based packages have been created for the DINA and G-DINA, which are discussed in detail in Chap. 26 (Robitzsch & George, this volume) and Chap. 29 (Ma, this volume). In addition to the ability to create customized statistical packages, R now has a set of tools such as the R "SHINY" interface facilitator that allow for the creation of easy to use and flexible web interfaces for any given statistics computational package, noting SHINY will be our new medium for our Spring 2019 release. This release of all of our RUM and ERUM procedures will use SHINY. It will include as a callable function the 0/1 scored RUM/RRUM Arpeggio suite, as described above, and also a callable option scored ERUM model-based suite and a small callable set of other option scored GDCM-based cognitive kernels such as EDINA.

3.7 RUM and ERUM Simulation Studies

In addition to theoretical considerations discussed above and the real data studies reported below, a wide range of realistic simulation studies have been carried out to help practitioners and researchers understand and assess the performance capabilities of the RUM system. In this section we give a short incomplete summary of the various simulations studies conducted. These were conducted from the viewpoint of how much information typical, but instead sometimes carefully designed, DAs can provide RUM system using practitioners and researchers. Because it is anticipated that ERUM usage may subsume RUM/RRUM usage, and because RUM/RRUM simulations have been reported elsewhere in the literature (e.g., see Hartz, 2001; DiBello et al., 1995), the emphasis below is on ERUM simulation studies.

Parameter estimation accuracy can be assessed by mean absolute deviations (MADs) for model parameters, noting that unlike in real data situations, in simulation studies the model parameters are known and hence MADs can be computed directly. We report here also on the individual attribute CCR*d*s, denoted $P(CC_d)$ recall. This, as previously noted, is appropriate because teachers usually

want to know separately attribute by attribute which skills and misconceptions are possessed and not possessed.

In our simulation studies, the model assumed by the estimation procedure is the true simulation model generating the data, and hence there is no model misfit studied in the simulations. Such robustness studies are planned in the future. Sometimes, we have bypassed estimation via the MMAP classification procedure using the true model, thus eliminating model estimation error. When this is done, the simulation study isolates the classification accuracy potential of the test by removing model misfit and model estimation error. Thus, the quality of the test itself becomes the only source of classification error. We give below a summary of results from various simulation studies run (results available from authors and to be published elsewhere in the future).

Based on emulation of parameter estimation values obtained in real data studies, we simulated items as having "high" (H) and "moderate" (M) item parameter discrimination values. High parameter value ranges of $r \sim$U(0.05,0.2) and $\pi \sim$U(0.7,0.9) represent well designed items, both for the correct and the incorrect options. Moderate parameter ranges of $r \sim$U(0.1,0.3) and $\pi \sim$U(0.5,0.9) represent typical ranges of items not specially designed for DA usage. Omitting details, we note that nonid plays no role here in the interpretation of the discrimination level of item parameters.

The Population Sampling Was Done so that the Marginal Attribute Proportions Were Always Between 0.4 and 0.6 This was done for several reasons. First, in practice these proportions may vary considerably but of course we don't know what they might be in an application. Choosing proportions close to ½ seems like an excellent "baseline" approach. Further, it has the advantage that the Kruskal and Goodman λ (or other) adjustment needed to properly interpret resulting simulation CCR*d*s when population attribute proportions seriously deviate from ½ is minor and can be ignored for practical purposes.

Noting that the ERUM option scoring enhancement of right/wrong scoring RUM is analogous to the MC-DINA enhancement of DINA, further corroboration of the superiority of option scoring is provided by a comparative simulation study (de la Torre, 2009) that found that the option scored MC-DINA improved on the skill correct classification rates (CCR*d*s) over right wrong scored DINA on average over all students and attributes from 91% to 97%. This is a striking improvement, noting that the error rate was reduced from approximately 9% to approximately 3%.

Because of the improvement realized by replacing right/wrong scoring by option scoring, it is important to observe that right/wrong scoring RUM/RRUM has already performed well both in simulation studies and real data analyses, and hence ERUM analyses should only do noticeably better, especially for well-designed DAs. To illustrate such RUM/RRUM results, Hartz (2001) did a simulation study of a 40 item, 7 skills, H and M cognitive structure, 1500 examinee test administration using the RUM Arpeggio system. She showed that individual attribute correct classification occurred 93% of the time for H cognitive structure and 84% for the

Table 3.1 Summary of the factors varied in the ERUM Simulation Studies

# attributes (skills and misconceptions)	Test length	Cog. $\pi\, r$ structure	Sample size, including True model (T) bypassing estimation	Correlational Structure (0 or ρ) : 0 denoting independent attributes
2(skills) + 2 (misc)	8, 15, 20, 30	HM,HH,MH,MM	125,250,500, 1000, 2000, T	0, ρ
$4 + 3 = 7$	8, 15, 30	"	1000	0
$3 + 4 = 7$	8, 15, 30	"	1000	ρ
$4 + 6 = 10$	15, 20, 25, 30	"	T	0
$5 + 7 = 12$	23	"	T	0

M cognitive structure, thus displaying reasonable classification accuracy, especially for the H structure.

We now briefly summarize our various ERUM simulation studies, where an effort has gone into realism, both for randomly generated item parameters and for the randomly generated Q matrices. As with all DCMs, the two major ERUM performance issues are (i) how accurately the ERUM model parameters are estimated and (ii) how accurately students can be classified, which is studied as a function of test length, examinee calibration sample size, Q matrix structure, item discrimination power as delineated by the simulated π and r ranges, and sometimes the latent attribute space correlational structure.

The ERUM simulation studies reported below are of short to long length MC tests that thus have the potential to provide accurate classification information of significant usefulness for teachers without (hopefully) taking up disproportionate amounts of classroom time. For instance, one could imagine using half of a class period to administer a 15-item MC quiz or a full class period for a 30-item test.

Table 3.1 summarizes the simulation design, including the various (usually partially crossed) factors considered for the ERUM simulation studies conducted, with actual simulation results summarized following the table. It and the brief summaries reported below for many of the simulation studies done substitute for fully reported results (available from authors) (Table 3.1).

Baseline Simulation Studies, as Proof of Concept First, as a baseline, we studied $D = 2$ skills $+2$ misconceptions $=4$ attributes (too small a number of attributes for most –but not all– practical applications, hence "baseline") with varying test lengths $K = 15, 20, 30$ and sample sizes $N = 250, 500, 1000$ and always 4 options/item, usually uncorrelated (0) but occasionally correlated (ρ) attributes (always MCMC estimated, regardless of whether the assumed simulation model structure is uncorrelated or not), highly discriminating items with $\pi\ U$ [0.6,0.9] and $r\ U[0.05, 0.2]$; i.e., HM. We note that the MADs were suitably moderate to small depending on sample size, with both π *and* r MADs around 0.1 for $N = 250$, around 0.07 for $N = 500$ and around 0.04 for $N = 2000$. Further, the average marginal CCRd rates $\geq$0.95 for all the studies, even the extreme case of $N = 250$, $K = 15$.

This is strong evidence that the methodology works in ideal baseline cases and suggests that π, r MAD rates as high as 0.1 do not seem to be deleterious. One study with a "reasonable correlational structure was done ($\rho_{SS} = 0.6$, $\rho_{SM} = -0.4$, $\rho_{MM} = 0.3$, where S denotes skill and M denotes misconception) with $N = 1000$ and $K = 15$ and resulted in an extremely high average CCRd of 0.98 and an associated $\rho\ MAD = 0.03$, surprisingly small. Clearly, estimation of the correlational structure improves on CCR rates.

Realistic Simulation Settings $D = 7, D = 10, D = 12$ were chosen as representative numbers of attributes for instruction-driven latent spaces. For example, a 20 min. quiz of 15 items might be designed to assess 7 attributes and a full period test of 30 items designed to assess 10 or 12 attributes, each space a mixture of skills and misconceptions. Of course, there are realistic settings (not yet simulated) where a longer test could be used to assess a relatively large number of attributes, say a 2 hour $K = 60$ item test designed to assess 8 skills +12 misc. =20 attributes.

As previously discussed, most DA usage is "low stakes" in the sense that moderate misclassification rates occurring are not overly detrimental in classroom FA settings, as contrasted obviously with high stakes settings where one might want to guarantee high CCRds, say ≥0.9. Thus a 0.78 rate (for attributes with possession rates close to ½) could be considered "effective", even though likely not so for a high stakes test. In interpreting the *CCRd*s, it is essential to note if one were ignoring test data information and merely randomly guessing at possession and non-possession of attributes, noting that attribute proportions were approximately 0.5 in our simulations, that this chance *CCRd* rate for each attribute would be approximately 0.5. We thus propose the following table for interpretation of our simulation results *CCRd* rates, *merely stated to evocatively fix ideas*, noting that each classroom, or other, setting will induce its own acceptable versus unacceptable *CCRd* error rates. Bluntly put, Table 3.2 should not be in any way taken as suggesting acceptable DA classification standards!

To save space, we summarize broadly the results of the large number of realistic ERUM simulation settings conducted, referring to averaged CCRds (over options) and, where appropriate, averaged MADs. First, if the number of attributes is modest, say 2 + 2 or even the 3 + 4 or 4 + 3 cases, and if the length of the test is not short, say ≥15, then high cognitive structure (HM) always produced *very effective* to *highly effective* CCRds and moderate cognitive (MH) structure produced either close to or within the *very effective* category. For larger numbers of attributes, say 10 or 12, then if the length of the test was ≥15 then high cognitive structure produced

Table 3.2 Ranges for interpreting our *CCRd* simulation values

Interpretation of the given CCRd range	CCRd range
Not effective	<0.7
Effective	[0.7, 0.8)
Very effective	[0.8,. 0.9)
Highly effective	≥0.9

very effective CCR*d*s. Whereas, for the same larger numbers of attributes, moderate cognitive structure produced *effective to very effective* CCR*d*s. As one concrete example high cognitive structure (H) coupled with a moderately long test length $K = 23$ and a large number of attributes $D = 5 + 7 = 12$ still yielded the *very effective* CCR$d = 0.85$.

Concerning item parameter estimation, the results were also very encouraging. Even smaller sample sizes such as $N = 250$, produced reasonable MADs, especially as interpreted by how estimation errors influenced the CCR*d*s. As a side note, when a plausible latent space correlational structure was introduced ($\rho_{SS} = 0.6$, $\rho_{SM} = -0.4$, $\rho_{MM} = 0.3$, where S denotes skill and M denotes misconception), MCMC estimation recovered the latent correlational structure well. And, this improved, sometimes significantly, the resulting CCR*d* rates.

Even when challenging non-small D numbers of attributes or small test length settings were studied the results were usually surprisingly good. For example, for $D = 3 + 4$ and $K = 8$, this interpretable as a very short "quiz" and with a moderate number of 7 attributes and high cognitive structure items, yielded an average CCR*d* rate of 0.76, namely "effective".

The ERUM "Simulator" randomly generates the Q structure with the result that the number of options that *touch* (touch ≡[an attribute is influencing an option response]) an attribute can vary considerably from attribute to attribute, exactly as is true for real tests, especially if not carefully designed for attribute balance. This suggests that variation in the CCR*d* rates from attribute to attribute should correspondingly vary as a function of number of touches. Omitting details, this turned out to be strongly the case.

In summary, ERUM simulation studies seem to (cautiously) suggest that DAs taking up a moderate amount of class time can effectively classify a large enough number of skills and misconceptions to be worth the class time expended.

3.8 Real Data Applications *of RUM and ERUM*

To date the RUM/RRUM has been applied most extensively in the area of language assessment (e.g., Jang, 2009; A.-Y. (Alicia) Kim, 2015; Y.-H. Kim, 2011; Lee & Sawaki, 2009a; Li & Suen, 2013a, 2013b), including a special issue devoted to the use of DCMs published in the journal *Language Assessment Quarterly* (Lee & Sawaki, 2009b). As noted by Roussos et al. (2007), DCMs can be applied in two primary ways: to revisit or re-analyze existing tests (sometimes called "retrofitting") or as part of a test development process wherein assessments are developed with the explicit purpose of conducting skills (or, now, misconceptions too) diagnosis. Most applications of the RUM/RRUM (and DCMs in general) have been of the former type, with one recent latter instance described below (Ranjbaran & Alavi, forthcoming). Such DCM designed tests are becoming more frequent (see, for example, Kunina-Habenicht, Rupp, & Wilhelm, 2009). It is also useful to differentiate between two types of retrofitting: applying a DCM to a

test (likely with summative intentions) initially designed to yield unidimensionally based conclusions versus one intentionally designed to make multidimensionally based diagnostic inferences, but without RUM based or other DCM playing a role at the design level. Here we describe four applications: (1) use of the RRUM to provide validity evidence for a pre-existing test of English language reading comprehension test used formatively, (2) use of the RUM to inform development of a new diagnostic test of English language reading comprehension, (3) use of the RRUM to analyze a pre-existing multi-dimensional concept inventory, and (4) use of the ERUM to analyze a pre-existing diagnostic assessment of middle school geometry concepts.

Researchers vary in whether they use the full RUM or the RRUM, recalling that the interpretation of the continuous η_d latent trait is that it is a unidimensional summary of all the attributes influencing item performance other than those specified in the Q modeled latent space. Li and Suen (2013a) found that excluding the residual η_d parameter from the model resulted in poor convergence of the MCMC chains for many item parameters, and hence used the full RUM with better results. Their think-aloud studies indicated the η_d factor may reflect construct-irrelevant test-taking strategies. Jang (2009), however, found that including the additional η_d parameter reduced the interpretability of the r_{kd} item parameters and opted to use the RRUM. Jang and Roussos et al. (2007) hypothesize that when fitting the RUM to pre-existing and often essentially unidimensional tests, the primary dimension may tend to overwhelm other parameters through the η_d term. We include this to sensitize potential RUM/RRUM users that the decision to include or exclude the η_d factor can be consequential. When in doubt, one perhaps should use the RRUM, especially if the latent space is judged reasonably "complete".

A series of studies by Jang (2005, 2009) provides the most comprehensive application of the RUM. Jang applied the RUM to two forms of a pre-existing English language reading comprehension test developed at ETS. The RUM framework was used to provide validity evidence to support use of the test as a diagnostic formative assessment in English language courses. Jang used think-aloud protocols with a sample of test-takers as well as content experts' judgments to identify a set of skills required by the test and construct a Q matrix. The RUM was then used to refine the Q matrix, evaluate the adequacy of model fit, and determine the consistency of estimated skill profiles of test takers. Finally, students and teachers in TOEFL preparation courses provided feedback about the perceived utility of the "diagnostic score reports" generated from the RUM. Subsequent applications of the RUM have followed a similar methodology, relying on a combination of expert judgment, think-aloud protocols, and statistical modeling to create and refine Q matrices for pre-existing tests (e.g., A.-Y. (Alicia) Kim, 2015; Li & Suen, 2013a). While these studies tend to find evidence of adequate model-data fit and interpretable item parameter estimates, there has been little reporting on applied uses of the test results, either for instructional purposes or to inform the revision and improvement of the tests. It is hoped that future efforts will move in this direction.

Ranjbaran and Alavi (2017) used the RUM in the process of developing a new diagnostic second-language English reading comprehension test intended to be used formatively. They used an evidence-centered design (ECD) framework to create a

20-item MC test. As with prior studies, think-alouds and content expert judgments were used to create and refine an appropriate Q matrix for the test. Although revisions to the test have not yet been made, the results of fitting the RUM to the data have provided useful insights about the nature and quality of the items that could be used to guide future test and item revisions. These researchers used the full RUM, although they found that the η_d parameter only seemed to be relevant for a small number of items on the test. Nonetheless, this still suggests use of the full RUM improved the fit of the model.

Distractor-driven tests provide a promising but under-utilized avenue for applying DCMs, particularly the RUM and ERUM. Distractor-driven tests are, as the name suggests, MC tests with systematically written distractors (incorrect response options) intended to provide additional information beyond simple right/wrong answer scoring. Concept inventories (CI's) are a form of distractor-driven test designed to assess non-quantitative understanding of key disciplinary concepts, often in the sciences. The distractors are written based on common student misconceptions or errors and are intended to provide useful diagnostic information for instructors. CI's are thus inherently multi-dimensional and are intended to be used formatively, to help teachers diagnose student reasoning and plan instruction. Santiago-Román and colleagues (Santiago-Román et al., 2010a, b) used the RRUM to analyze data from a Concept Assessment Tool for Statics (CATS; Steif & Dantzler, 2005), a CI designed to assess students' understanding of nine important concepts in (engineering) statics. Through a combination of conceptual and statistical analyses based on the RRUM, they identified 10 attributes of understanding assessed by the CATS that could be identified with high levels of consistency. Jorion et al. (2015) provide an extensive discussion regarding the use of DCMs to provide important validity evidence for CI's.

More recently, Shear and Roussos (2017) used the ERUM model to analyze data from a distractor-driven test of middle school geometry concepts developed by the Diagnostic Geometry Assessment (DGA) project (Masters, 2012). The DGA project developed a series of MC tests, each written to assess student understanding of geometry concepts and identify students reasoning with systematic misconceptions. The incorrect response options on the DGA tests were written to correspond to common student misconceptions. Previously, DGA tests reported two separate scores – a "knowledge" score indicating the number of correct response options selected and a "misconception" score indicating the number of misconception response options selected. Because these scores required separate scoring keys, they could not be analyzed in a single model using traditional psychometric methods such as CTT or unidimensional IRT models. This made it difficult to a) model the internal structure of the test while taking into account the correct, incorrect and misconception responses simultaneously, and b) determine appropriate cut scores for the "knowledge" and "misconception" scores that could be used to classify examinees

Shear and Roussos (2017) used the ERUM to address these challenges, focusing on a single DGA test form. First, by comparing the fit of multiple Q matrices, they concluded that the test was likely assessing two misconceptions rather than

one, as was initially hypothesized. While earlier analyses had suggested this might be a possibility, the hypothesis had not been evaluated in a systematic modeling framework. Second, the ERUM provided a defensible, model-based method for identifying which students were likely to understand the targeted concept and that were likely to be reasoning with each of the targeted misconceptions. The results also provided information that could be used by test developers to identify items functioning best for diagnostic purposes. Although the DGA tests were not developed within a DCM framework, the multidimensional nature of the tests and the intended diagnostic uses made the ERUM a valuable tool for evaluating and refining the test. The applications by Santiago-Román et al. and Shear and Roussos suggest the RUM and ERUM appear to have significant potential for supporting and improving the use of multi-dimensional, distractor-driven tests, including those that are pre-existing. Given the potential value of such tests for formative assessment practice, this remains an important and exciting area for further research.

3.9 Discussion and Summary

The RUM diagnostic system holds the promise of providing accurate psychometrically assisted diagnostic classifications via MC (or finite response category coded short answer) diagnostic assessments (DAs), especially Formative Assessments and Interim Assessments (interior to the unit but not necessarily intended for formative purposes). Such DAs can assess learning progress and thus enhance instruction. And, the diagnostic system can be useful during DA development and refinement stage. Further, applied research questions concerning the nature of cognitive processing and the validity of MC assessments can be addressed via RUM/ERUM, including at the fine-grained individual attribute and individual item/option level, as suggested by some of the real data studies reported upon above.

The RUM system, even when the ERUM model with its nonid parameters is being employed, outputs identifiable and usefully interpretable model parameters and resulting examinee attribute classifications. Because of its estimable and interpretable parameters, evaluation of item and hence test quality becomes possible, even at the fine-grained individual attribute and individual item option level. Indeed, it is recommended that when possible future MC DAs be optimally designed with items functioning in concert to well classify examinees on all of the attributes in the latent attribute space, misconceptions as well as skills. Psychometrically informed DA MC test design, with item redesign or replacement, becomes possible. Importantly, via ERUM modeled individual option scoring, incorrect "distractor" options provide useful recoverable diagnostic information, especially about misconceptions. Thus, combining RUM system modeling with well-designed (possibly psychometrically aided) DAs, suggests the resulting classification accuracy should be good, a conclusion our simulation and real data studies reported upon above cautiously supports.

This chapter provides a user-oriented description of the RUM/RRUM and then introduces the ERUM GDCM generalization, noting that ERUM is a mixture model with a cognitive and a guessing component, as seems realistic for MC items, where guessing and competition are forced to occur. The issue of ERUM non-identifiability is briefly discussed, making clear that the intrinsic competition and guessing nature of MC items can sometimes produce DCM non-identifiability because cognitively distinct items (with varying amounts of guessing and competition) can produce the same ERUM or other realistic DCM MC probability model. This non-identifiability of ERUM is foundational and as such is not a superficial modeling flaw that might be eliminated by a better parameterization. When ERUM non-identifiability holds, the ERUM MCMC estimation software produces identifiability via adding in DF (in number) parameter constraints. Alternatively, and not reported on herein, a Bayesian approach under development removes ERUM non-identifiability via the introduction of an appropriately chosen prior on the F column sums S_α, a prior that favors item models that reduce guessing and competition among all the equivalent nonidentifiable ERUM modeling choices.

In the Model-data Checking Section, both fit and discrimination are considered. In addition to standard fit approaches such as AIC/BIC, several specialized indices of fit have been developed that are tuned to the special nature and purposes of DCMs and the RUM system. The fit approach also includes the fit of Q with some tools available in the RUM system that allow minor modifications of Q that can improve fit. However, more work on the fit of Q for the RUM system is needed, likely drawing on the work of researchers focusing on Q development and will be carried out in the future.

Use of correct classification rates (CCRs), marginal for each attribute separately but also available for all the attributes simultaneously, were identified as the major discrimination approach, noting that from the instructional perspective, marginal rates are the main interest usually. In addition, specialized, easily computed, and useful discrimination indices were presented as well.

The RUM/ERUM system has three distinct applications foci, namely existing test evaluation/new test construction, examinee classification, and test validity studies, including at the individual attribute and item levels: First, aided by the system's discrimination indices described above, the RUM/ERUM system can be a valuable tool in the design and evaluation of proposed diagnostic MC items and in the subsequent construction of DAs. Second, the system has the potential to capture most of the diagnostic attribute classification information in a DA, especially including the capture of classification information in incorrect options. Thus, misconceptions as well as skills can be well classified. One important use of such diagnostic assessments is that they can help drive remedial instruction, at both the classroom level and the individual student level. Another use is assessing the validity of existing tests: for example, is a studied test measuring the attributes the designer of the test intended and/or does it measure other attributes, especially non-construct-valid attributes that could contribute to DIF?

The large number of simulation studies we have carried out have been only briefly summarized, noting these studies combine to suggest the levels of perfor-

mance one can expect from RRUM/ERUM associated DAs. The studies suggest that MC tests, even relatively short tests or quizzes, can in fact assess a reasonable number of attributes (skills and misconceptions) with good to very good classification accuracy, as quantified by reported marginal CCR*d*s. Further, a number of real data studies have been reported upon above, both for RUM and for ERUM, for existing tests, some of which were "retrofitted" and some actually designed for DA usage. In the Ranjbaran and Alavi ESL study, the RUM model was used to enhance DA test construction, as this chapter highly recommends for optimal DA development. And, in the Shear and Roussos ERUM study reported upon, interesting attribute level validity conclusions resulted. These real data studies display a variety of successful applications of RUM or ERUM, and further have helped suggest realistic ranges for IRF parameters to be used in current and future ERUM simulation studies, noting defensible realism is a major goal.

The authors believe that the RUM (includes ERUM recall) system can be a valuable tool in bringing psychometrically supported DAs heavily into the instructional process and further can aid cognitive and test validity research. Relatively easy to use RUM system software will be available in Winter 2018, driven by the SHINY interface embedded within R, with the heavy estimation, simulation and classification computation done by our existing Fortran embedded code.

References

Beeley, C. (2013). *Web application development with R using SHINY*. Birmingham, UK: PACKT Publishing.

Bradshaw, L., & Templin, J. (2014). Combining scaling and classification: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*, 403–425.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–63.

Chen, Y., Culpepper, S., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika, 83*, 89–108.

Chiu, C., Kohn, H., & Wu, H. (2016). Fitting the reduced RUM with Mplus: A tutorial. *International Journal of Testing, 16*(4), 331–351.

Chung, M., & Johnson, M. (2017). *An MCMC algorithm for estimating the Reduced RUM*. https://arxiv.org/abs/1710.08412

de la Torre, J. (2009). A cognitive diagnosis model for cognitively-based multiple choice options. *Applied Psychological Measurement, 33*, 163–183.

DiBello, L., Stout, W., & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Lawrence Erlbaum Associates.

DiBello, L. V., Henson, R., & Stout, W. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement, 39*, 62–79.

DiBello, L. V., & Stout, W. (2008). *Arpeggio documentation and analyst manual.* Department of Statistics, University of Illinois, Champaign-Urbana IL (Contact W. Stout).

Feng, Y., Habing, B., & Huebner, A. (2014). Parameter estimation of the reduced RUM using the EM algorithm. *Applied Psychological Measurement, 38*, 137–150.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–511.

Gilks, W. R., Richardson, S., & Spielgelhalter, D. J. (1996). *Markov chain Monte Carlo*. London, UK: Chapman and Hall/CRC.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.

Han, Z., & Johnson, M. (this volume). Global model and item-level fit indices. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Hartz, S. M. (2001). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality.* Dissertation. University of Illinois at Urbana-Champaign.

Henson, R., DiBello, L., & Stout, W. (2018). A generalized approach to defining item discrimination for DCMs. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 18–29. https://doi.org/10.1080/15366367.2018.1436855

Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* Dissertation. University of Illinois at Urbana-Champaign.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to language assessment. *Language Testing, 26*(1), 031–073. https://doi.org/10.1177/0265532208097336

Jorion, N., Gane, B. D., James, K., Schroeder, L., DiBello, L. V., & Pellegrino, J. W. (2015). An analytic framework for evaluating the validity of concept inventory claims: Framework for evaluating validity of concept inventories. *Journal of Engineering Education, 104*(4), 454–496.. https://doi.org/10.1002/jee.20104

Kim, A.-Y. (Alicia). (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing, 32*(2), 227–258. https://doi.org/10.1177/0265532214558457

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28*(4), 509–541. https://doi.org/10.1177/0265532211400860

Kunina-Habenicht, O., Rupp, A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*, 64–70.

Kuo, B., Chen, C., & de la Torre, J. (2017). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement, 42*(3), 179–191.

Lee, Y.-W., & Sawaki, Y. (2009a). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239–263. https://doi.org/10.1080/15434300903079562

Lee, Y.-W., & Sawaki, Y. (2009b). Cognitive diagnosis and Q-matrices in language assessment. *Language Assessment Quarterly, 6*(3), 169–171. https://doi.org/10.1080/15434300903059598

Li, H., & Suen, H. K. (2013a). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test. *Educational Assessment, 18*(1), 1–25. https://doi.org/10.1080/10627197.2013.761522

Li, H., & Suen, H. K. (2013b). Detecting native language group differences at the subskills level of reading: A differential skill functioning approach. *Language Testing, 30*(2), 273–298. https://doi.org/10.1177/0265532212459031

Liu, J., & Johnson, M. (this volume). Estimating CDMs Using MCMC. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Liu, X., & Kang, H. (this volume). Q matrix learning via latent variable selection and identifiability. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Ma, W. (this volume). The GDINA R-package. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Masters, J. (2012). *Diagnostic geometry assessment project: Validity evidence* (Technical Report). Measured Progress Innovation Laboratory.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*, 7–11.

Ranjbaran, F., & Alavi, M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation 55* 167–179.

Robitzsch, A., & George, A. (this volume). The R package CDM for diagnostic modeling. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). New York, NY: Cambridge University Press.

Santiago-Román, A. I., Streveler, R. A., & DiBello, L. V. (2010a). *The development of estimated cognitive attribute profiles for the concept assessment tool for statics*. Presented at the 40th ASEE/IEEE Frontiers in Education Conference, Washington, DC.

Santiago-Román, A. I., Streveler, R. A., Steif, P. S., & DiBello, L. V. (2010b). *The development of a Q-matrix for the concept assessment tool for statics*. Presented at the ERM division of the ASEE annual conference and exposition, Louisville, KY.

Shear, B. R., & Roussos, L. A. (2017). Validating a distractor-driven geometry test using a generalized diagnostic classification model. In I. B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (Vol. 69, pp. 277–304). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-56129-5_15

Steif, P. S., & Dantzler, J. A. (2005). A statics concept inventory: Development and psychometric analysis. *Journal of Engineering Education, 94*(4), 363–371. https://doi.org/10.1002/j.2168-9830.2005.tb00864.x

Sullivan, M., Pace, J., & Templin, J. (this volume). Using *Mplus* to estimate the Log-Linear Cognitive Diagnosis model. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50.

von Davier, M., & Lee, Y.-S. (this volume). Introduction: From latent class analysis to DINA and beyond. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Zhang, S., Douglas, J., Wang, S., & Culpepper, S. (this volume). Reduced Reparameterized Unified Model applied to learning system spacial reasoning. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Chapter 4
Bayesian Networks

Russell G. Almond and Juan-Diego Zapata-Rivera

Abstract Bayesian networks (or Bayes nets) are a notation for expressing the joint distribution of probabilities over a number of variables. Variables in a Bayesian network can be continuous or discrete (Lauritzen SL, Graphical models. Oxford University Press, New York, 1996), however, when all variables are discrete, all calculations can be represented as a series of sums and products. As such, Bayes nets provide a notation for expressing a wide variety of cognitive diagnostic models, including ones described in other chapters of this book. Several commercial software packages are available for supporting Bayesian networks including HUGIN (Andersen SK, Olesen KG, Jensen FV, Jensen F, Hugin—a shell for building Bayesian belief universes for expert systems. In: IJCAI'89, Detroit, MI, 1989. Reprinted in Shafer and Pearl 1990), Netica (Norsys, Inc., Netica [Computer software manual], 2004. Retrieved from http://www.norsys.com), Genie (BayesFusion, LLC, Genie, 2017. Retrieved from http://bayesfusion.com (Bayesian network Computer Software)) and BayesiaLab (Bayesia, Bayesialab, 2017. Retrieved from http://www.bayesia.com (Bayesian Network Computer Software)) as well as a number of free software packages.

4.1 Introduction to Bayesian Networks

Bayesian networks (or Bayes nets) are a notation for expressing the joint distribution of probabilities over a number of variables. As such, they provide a notation for expressing a wide variety of cognitive diagnostic models (CDMs), including ones

R. G. Almond (✉)
Department of Educational Psychology and Learning Systems, Florida State University, Tallahassee, FL, USA
e-mail: ralmond@fsu.edu

J.-D. Zapata-Rivera
Educational Testing Service, Princeton, NJ, USA
e-mail: dzapata@ets.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_4

described in other chapters of this book. In a Bayesian network, the variables (both latent proficiency variables and observable outcome variables) are expressed as nodes of an acyclic directed graph. In this respect, Bayes nets and structural equation models (SEMs) are similar. However, unlike SEMs, separation in a Bayesian network always implies conditional independence: this allows the network structure to be used to design efficient computational algorithms for a particular model.

Bayes nets have been well studied in computer science since Pearl (1988) with several volumes describing both the representations and algorithms for computing with them (e.g., Jensen, 1996; Cowell, Dawid, Lauritzen, & Spiegelhalter, 1999; Neapolitan, 2004). Almond, Mislevy, Steinberg, Yan, and Williamson (2015) describe the specific application to educational models, and this article follows the notation used in that book. Variables in a Bayesian network can be continuous or discrete (Lauritzen, 1996), however, when all variables are discrete, all of the calculations can be represented as a series of sums and products. Several commercial software packages are available for supporting Bayesian networks including HUGIN (Andersen et al., 1989), Netica (Norsys, 2004), Genie (BayesFusion, 2017) and BayesiaLab (Bayesia, 2017) as well as a number of free software packages.

The existence of this software makes it straightforward to score a single student. The variables corresponding to the observed performance of a student on a collection of tasks or items are instantiated to their observed values. The software propagates the information about the observed variables to all of the unobserved variables, drawing inferences about both latent proficiency variables and unobserved task outcomes. This makes Bayes nets an attractive model for embedding in an intelligent tutoring system which would make use of the inferences about student performance to select appropriate instruction and tasks.

4.1.1 *Graphical Representation of Joint Probability Distribution*

Consider a collection of variables, $X_1, \ldots, X_K$. This collection includes both the observable outcome variables from the tasks and items and the latent proficiency variables. Consider an ordering of the variables $1, \ldots, K$. Then the joint distribution can be written as $\Pr(X_1, \ldots, X_K) = \Pr(X_1) \prod_{k=2}^{K} \Pr(X_k | X_{k-1}, \ldots, X_1)$. Often some of the conditioning variables drop out of the factor $\Pr(X_k | X_{k-1}, \ldots, X_1)$. Let $\text{pa}(X_k)$ be the *parents of* X_k, the set of remaining variables. Drawing a directed edge from each parent variable to its child produces an acyclic directed graph (it is acyclic because parents are always earlier in the ordering), $\mathcal{G}$. The equation for the joint probability then becomes

$$\Pr(X_1, \ldots, X_K) = \prod_{k=1}^{K} \Pr(X_k | \, \text{pa}(X_k)) \, , \tag{4.1}$$

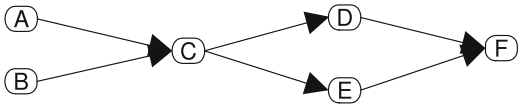


Fig. 4.1 A simple acyclic directed graph. (Reprinted from Almond, Mislevy, Williamson, & Yan 2007 with permission from ETS)

where $\Pr(X_k|\,\mathrm{pa}(X_k))$ is understood to be an unconditional probability if $\mathrm{pa}(X_k)$ is empty. If the variables are all discrete, then the factors $\Pr(X_k|\,\mathrm{pa}(X_k))$ are *conditional probability tables* (CPTs).

Figure 4.1 shows a simple directed graph. Terms related to ancestry are used to express the relationship among variables. Node A is a *parent* of Node C; Node C is a *child* of Nodes A and B; Nodes C, D, E, and F are *descendants* of Node A; Nodes C, B and A are *ancestors* of Node D; Nodes D and E are *siblings* with common parent Node C. Nodes which are connected by a single edge regardless of direction, for example Nodes C and D, are called *neighbors* and the set of nodes directly connected to a given node is called its *neighborhood*. For example, the neighborhood of Node E is Nodes C and F. A set of nodes which are all connected to each other is called a *complete set* and a maximal complete set (i.e., one to which another node connected to all of the others in the set cannot be found) is called a *clique*.

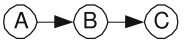
Associating a conditional probability table with each node in the graph turns it into a Bayes net. For Fig. 4.1 the Bayesian network is represented by the factors:

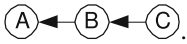
$$\Pr(A, B, C, D, E, F) = \Pr(A)\Pr(B)\Pr(C|A, B)\Pr(D|C)\Pr(E|C)\Pr(F|D, E)\ .$$

Note that there could be more than one graph which can be used to represent a given joint probability distribution (Pearl, 1988). In particular, the saturated graph (the one with every pair of variables connected) can always be used; however, it is usually inefficient. The computational cost for most Bayesian network algorithms is driven by the size of the largest clique (before calculating the largest cliques, the common parents of nodes need to be connected, see Cowell et al., 1999 or Almond et al., 2015 for details). In the Cowell et al. algorithm, computational cost is linear in the total number of nodes, but exponential in the size of the largest clique. Finding an optimal graphical representation can be difficult; however, graphs in which the edges are oriented according to a causal theory are generally more efficient. In applications in educational testing, the edges usually point from latent proficiency variable to observable outcome variables (although edges between proficiency variables and between observable variables are also allowed).

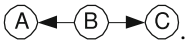
4.1.2 Conditional Independence

When a Bayes net is factored according to Eq. 4.1, then a number of conditional independence conditions hold. Basically, nodes which are separated in the graph are conditionally independent given the separating set. This is called a *Markov* property. With directed graphs, a more complicated version of separation called *d-separation* (Pearl, 1988) is needed. The rules for d-separation can be complex, but can be understood through three simple examples.

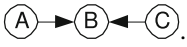
Evidence Chain 

or 

. If the variable in the middle (Node B) is known, then it makes the outer variables (Nodes A and C) independent; otherwise, they are dependent. For example, if A is that the student has good study habits, B is whether or not the student has mastered the class material, and C is that the student performs well on the test, then knowing whether or not the student knows the material renders the student's performance independent of the student's study habits.

Common Cause 

. If the common parent (Node B) is known, then it makes the two child variables (Nodes A and C) independent; otherwise, they are dependent. For example, if B is mastery of the material, and A and C are performance on two different tasks or items on a test, then knowing B renders A and C independent (This is the local independence assumption of item response theory, IRT).

Competing Explanations 

. When Nodes A and C have a common child (or common descendant), then A and C are independent when the common descendant is *unknown*. Let B be performance on a mathematics word problem and A and C be skill in English and mathematics respectively. If A and C are assumed to be independent a priori, learning that the student did poorly on the problem would change that independence. In particular, if the student did poorly then learning that the student had good command of English would indicate that the student has difficulty with mathematics. Similarly, learning that the student has good mathematical skills would point towards difficulty in English. So learning that the student was unable to solve the problem induces a negative correlation between the two explanations.

Together, these three rules make up the rules for d-separation. Two nodes A and C are conditionally independent given the values of a set of nodes **B** if every path from A to C is blocked by at least one node in **B**, for every route from A to C which travels through a common ancestor, at least one common ancestor common ancestor is in **B**, and no common descendant of A and C is in **B**.

Establishing these conditional independence properties is the hard part of building a Bayesian network. However, it is also these conditional independence relationships that make possible the message passing algorithms described in Sect. 4.2.

For educational testing models, the set of nodes in the Bayes net are often partitioned into a set of proficiency variables, $\{X_1, \ldots, X_k\}$, and observable outcome variables, $\{Y_1, \ldots, Y_J\}$. Let $q_{jk} = 1$ if and only if there is a directed edge from X_k to Y_j. (In educational models it is usually more efficient to parameterize Bayesian networks with edges flowing from proficiency to observable variables.) This is the same as the Q-matrix used in many other CDMs (see other chapters in this volume); the Q-matrix is the incidence matrix for edges between proficiency and observable variables. (Note that this is not a complete description of a Bayesian network, a complete specification also requires the analyst to specify a graphical structure for the proficiency variables, Almond, 2010).

If there are no edges between the observable variables, the graphical structure implied by the Q-matrix gives that the observable outcomes are independent given the proficiency variables. This is the multivariate extension of the local independence property of IRT models. Note that the Bayes net notation is not restricted to strict local independence; the analyst can model additional dependence among observables by adding edges between observables.

4.1.3 Parameterization of Conditional Probability Tables

When all of the variables in the Bayes net are discrete, the parameters of the Bayes net are the conditional probability tables (CPTs), $\Pr(X_k | \text{pa}(X_k))$. Let s be an index for the state of X_k and c be an index of the configuration (assignment of the variables to values) of the parent variables. Then p_{cs} is the conditional probability that $X_k = s$ given that $\text{pa}(X_k)$ are in configuration c. In particular, each row corresponds to a conditional probability distribution for the child given a specific parent configuration and each column corresponds to a state of the child variable. Note that this implies that $\sum_{s=1}^{S} p_{cs} = 1$.

When learning the CPTs from data, it is helpful to have a prior distribution over the CPT. A standard parameterization is to give every row of the CPT an independent Dirichlet distribution; this is called a *hyper-Dirichlet* distribution. The parameters of the hyper-Dirichlet distribution are a table of pseudo-counts, where a_{cs} represents the number of times X_k was observed in State s when its parents where in Configuration c. The row sums $A_c = \sum_{s=1}^{S} a_{cs}$ represent the total amount of information about the row and are inversely related to the precision. The expected value for the cells of the CPT are $E[p_{cs}] = a_{cs}/A_c$.

The hyper-Dirichlet model has two drawbacks. First, the number of parameters that must be specified or learned rises exponentially with the number of parents. Second, in many educational applications, the parent and child variables are ordered categories, and there is a certain implied monotonicity: better skills should lead to better expected performance. To overcome these problems a number of alternative parameterizations have been suggested.

The *noisy-or* (or more commonly in education *noisy-and*) was one of the first parameterizations proposed (Pearl, 1988; Díez, 1993; Srinivas, 1993). Let Y_j be the

child variable, and $X_1, \ldots, X_K$ be the parents of Y_j. With each parent variable, associate a *guessing parameter*, g_k, and let s_j be a *slipping parameter* for the item. Assume that all of the parents and children are binary. Then for the noisy-and model,

$$\Pr(Y_j = 1|X_1, \ldots, X_k) = s_j \prod_k g_k^{1-X_k} . \tag{4.2}$$

For certain restrictions of the parameters, these correspond to the noisy-input deterministic-and (NIDA) and deterministic-input noisy-and (DINA) models of Junker and Sijtsma (2001) (see also Rupp, Templin, & Henson, 2010). If the parents are ordered categorical variables, the noisy-and and noisy-or models extend into the noisy-min and noisy-max models. Certain Bayes net packages have special algorithms can exploit the noisy-and and noisy-or parameterizations (Li & D'Ambrosio, 1994).

DiBello introduced another paradigm for creating CPTs based on multivariate IRT models (Almond et al., 2001). The key idea is that if each state of each parent variable can be mapped onto a point on an IRT theta scale, then existing multivariate IRT models can be used to calculate the conditional probability tables. The method has three steps:

1. *Effective* θ_s. For each state x_{ks} of each parent variable X_k, let $\tilde{\theta}_{ks}$ be a corresponding real number. Almond et al. (2015) suggests assigning these numbers based on quantiles of the normal distribution (corresponding to the number of states of X_k).
2. *Combination Rule*. Each row of the CPT corresponds to an assignment of parent variables to states and hence effective θ_s. These are combined using a combination rule, $Z(\cdot)$ to produces a single effective θ for the row $\tilde{\theta}_c = Z(\tilde{\theta}_{1c}, \ldots, \tilde{\theta}_{Kc})$.
3. *Link Function*. The effective θ(s) for each row are placed in a link function, $g(\cdot)$ which converts the effective θ_s into conditional probabilities.

Although presumably an infinite number of combination rules and link functions is possible, in practice, only a small set is regularly used. The ones coded in the R (R Development Core Team, 2007) package CPTtools (Almond, 2015, 2017a) are listed below.

First the combination rules:

Compensatory This is appropriate for situations in which more of one skill compensates for less of another. The combination rule is a weighted average of the inputs: $Z(\tilde{\theta}_{1c}, \ldots, \tilde{\theta}_{Kc}) = \frac{1}{\sqrt{K}} \sum_k a_k \tilde{\theta}_{kc} - b$. The $\frac{1}{\sqrt{K}}$ is a variance stabilization term which keeps the variance of $Z(\cdot)$ from growing as the number of parents grows.

Conjunctive This is appropriate for situations in which all skills are necessary to solve the problem. The combination rule is a weighted minimum: $Z(\tilde{\theta}_{1c}, \ldots, \tilde{\theta}_{Kc}) = \min_k(a_k \tilde{\theta}_{kc}) - b$.

Offset Conjunctive In many applications, it is not that the input skills have different slopes, but rather different offsets, so should have different intercepts. This rule swaps the numbers of slope and intercept parameters for the conjunctive rule: $Z(\tilde{\theta}_{1c}, \ldots, \tilde{\theta}_{Kc}) = a \min_k(\tilde{\theta}_{kc} - b_k)$.

Disjunctive and Offset Disjunctive These are appropriate in situations for which the skills represent different solution paths so that performance is dominated by the strongest skill. The rules are formed by substituting a max for the min in the equations for the conjunctive and offset conjunctive rules respectively.

There are three structure functions in common use.

Normal Link (Almond et al., 2015). This link function represents a discretized regression. It uses a single structure function representing the expected position on the scale given the parent inputs, and a link scale function giving the residual standard deviation around that expected value. It is most useful for representing the relationship between proficiency variables.

Graded Response (Almond et al., 2001). This link function is based on the graded response model (Samejima, 1969). There is a different structure function for each state of the child variable, Y, and $Z_s(\cdot) = \text{logit}[\Pr(Y \geq y_s | \text{pa}(Y) = c)]$. The entries in the CPT are found by differencing these curves. Some care must be taken so that curves do not cross, which could cause negative probabilities. This is generally achieved by using the same form for the structure function at each step, using the same discrimination (slope) parameters and ensuring that the difficulties (intercepts) are in increasing order.

Partial Credit This link function is based on the generalized partial credit model (Muraki, 1992). There is a different structure function for each state of the child variable, Y, and $Z_s(\cdot) = \text{logit}[\Pr(Y \geq y_s | Y \geq y_{s-1}, \text{pa}(Y) = c)]$. The probabilities for each level can be calculated by multiplying the conditional probabilities together and then normalizing. Note that this link function is more flexible than the graded response link function in that the structure functions do not need to have the same discriminations or even be of the same structural form.

Detailed formulas are given in Almond (2015) and code is available in the CPTtools package (Almond, 2017a).

4.1.4 Evidence-Centered Assessment Design and Bayesian Networks

Bayesian networks can describe a large universe of possible models with detailed information needed to specify both the model graph and the CPT for each node in the graph. Evidence-centered assessment design (ECD; Mislevy, Steinberg, & Almond, 2003) was developed in part to support the knowledge engineering

efforts required to build Bayes nets for assessment. For this reason, it is common to use the terminology of ECD when building Bayes nets for education (Almond et al., 2015).

In ECD, the complete Bayesian network is split between the central *proficiency model*—models which describe the relationship among *proficiency variables* (often called attributes in other forms of CDMs) in the target population—and the statistical part of the *evidence models*—models which link the observable outcome variables (often called item responses) to the relevant proficiency models—for each task in the assessment. This makes a hub-and-spoke model of the assessment, with the proficiency model forming the central hub and the evidence models plugging into the central model. Note that in a particular form of the assessment, only certain tasks need to be used, in which case only the evidence models associated with those tasks are plugged into the central hub.

The proficiency model is a complete Bayesian network. The evidence models, however, are only fragments: they contain references to the proficiency variables (stubs) and not the complete variable definitions. The probability distribution over the proficiency variables is given in the proficiency model. The set of proficiency variables referenced in the evidence model is known as the *footprint* of the evidence model. The probability distribution for the observable outcome variables is given in the evidence model. If enough data are available, evidence models' parameters can be calibrated, giving task-specific values for the parameters.

In many assessments, each task is associated with a single outcome variable. In this case, specifying the evidence model involves mostly specifying which proficiency variables are relevant (the footprint), and the form and the parameters for the CPT. Almond (2010) suggests doing this with an augmented Q-matrix, in which additional columns in the Q-matrix are used to select the combination rule for the CPT for the evidence model. (Almond, 2017d, extends this notation to work with the more complex CPTs available with the partial credit link function.) In these situations (single observable per task) the Bayes net model looks much like many other CDMs.

There are two differences between a Bayes net expressed through a Q-matrix and other CDMs. First, the decision about whether to use a compensatory, conjunctive, or disjunctive model is made at the level of the observable outcome and not the whole model. (This is more an issue of implementation than a limitation of theory. There is no reason why multiple forms of a CDM could not be supported at once, it is just that most software packages do not offer a mechanism for doing customization at the item level.) Almond (2010) suggested augmenting the Q-matrix with additional information about which structure function should be used (as well as parameter values) for each observable.

Second, the Bayes net requires the relationship among the proficiency variables (attributes) to be explicitly stated and modeled. In particular, the proficiency model defines the distribution of proficiency profiles in the target population. Almond (2010) suggests this can be done through an inverse correlation matrix (possibly found through factor analysis). Most other forms of CDMs also define a population

distribution for the attributes, but it is often done through a default (such as assuming that the attributes are a priori uncorrelated and then learning the CPT from data).

One distinct difference between Bayesian networks and most other forms of assessment is how they handle complex tasks with multiple observed outcome variables. In this case, the psychometrician is free to develop a more complex Bayesian network which describes the relationship among proficiencies and observables. The local independence assumptions for a Bayesian network is slightly different for that of an item-based models. Observables associated with different tasks are conditionally independent given the footprint variables for the tasks. Within tasks, the limits are left to the modeler's imagination.

4.2 Single Student Inference

A completely specified Bayesian network (one whose conditional probability tables are all known) is a description of the joint probability of the latent and observable variables for the population of interest. Scoring a single student is straightforward and the required operations are supported by almost all Bayesian network software. First, a student specific copy of the network is made. Next, all of the observed variables are *instantiated* (set) to their observed values. Then a simple message passing algorithm is used to update the marginal probabilities for all nodes in the network. Statistics of these marginal probability distributions can be reported as scores.

The Bayesian network inference engine is available a shared code library in both free and commercially supported versions; this makes Bayesian networks an attractive model for embedding in an intelligent tutoring system, adaptive testing system, or other system that needs real-time scoring. As many other CDMs can be represented as Bayes nets, first translating the estimated models to a Bayes net and then using the Bayes net for scoring is an attractive method for using those models in embedded applications.

4.2.1 Calculating Posterior Proficiency Profiles

Kim and Pearl (1983) described the first version of the message passing algorithm. Messages going in the direction of the arrows would sum across states of the parent variables, giving the prior distribution of the child variable. Messages going in the opposite direction of the arrows would apply Bayes' theorem, passing the likelihood of observing the evidence below in the graph given the various states of the parent variable. Combining the message coming from the parents with the message coming from the children produces the marginal distribution of the target node.

The Kim and Pearl (1983) algorithm was restricted to polytrees, a special kind of directed graph with no undirected cycles, but Pearl (1988) noted that this limitation could be overcome by clustering nodes together. Eventually, it was determined that the optimal clusters of nodes were related to the cliques of the original graph (after the common parents of any node were connected) and the *tree of cliques* (the cliques arrayed in a tree structure) could be used for message passing. Adding nodes representing the intersection of the cliques between the clique nodes produces the *junction tree*, and the message passing algorithm is known as the *junction tree algorithm* (Cowell et al., 1999). The cost of the junction tree algorithm is related to the size of the largest clique in the graph (the *treewidth*). While alternative algorithms exist for cases with large treewidths, for most educational applications the junction tree algorithm works with an acceptable computational cost.

Although the junction tree algorithm was described using repeated applications of Bayes' rule to combine prior probabilities and likelihoods, the denominator of Bayes theorem, the normalization constant, does not need to be calculated until probability is interpreted. Saving the normalization step for the end has two consequences. First, it improves the numerical stability of the calculation. Second, the normalization constant is the prior probability of all of the evidence entered into the network. This quantity is useful in a number of model-fit and person-fit calculations.

Almond and Mislevy (1999) propose a variation on the junction tree algorithm for the situation where the Bayes net is distributed according to a hub-and-spoke model. The complete model of the assessment consists of the core proficiency model (calibrated to the population of interest) and a collection of task-specific evidence models (evidence models whose parameters have been calibrated to a particular task). The system then supports the following operations:

- When a new student starts the assessment, the proficiency model is copied for that student. The student-specific copy is the *student model.*
- When the scoring engine receives a message from other parts of the system that a particular student has completed a particular task along with the values for the observables for that task:
  1. The scoring engine fetches the student model for that student and the evidence model for that task, respectively.
  2. The scoring engine adjoins the evidence model, which is a fragment, to the student model, connecting the footprint variables.
  3. The observable variables are instantiated and the evidence is propagated into the student model variables.
  4. Once the evidence has been absorbed, the variables unique to the evidence model can be discarded.
- The student model for a particular student can be queried at any time to make score reports or decisions about what content to provide next.

– The scoring engine can combine a student model and a task-specific evidence model to predict how a student might perform on a task. This can be useful for adaptive selection or sequencing of items.
– A student model can be saved and restored either by writing its current information out to a database or by saving the sequence of tasks and observables and then replaying them to recalculate the score.

This algorithm takes advantage of the equivalence under Bayes theorem of incorporating all evidence at once or in batches (here corresponding to specific tasks). As a consequence, at any time the student model for a particular student contains the joint probability distribution of proficiency profiles (assignments of proficiency variables to states) based on the observed evidence for that student. Initially, this value is based on the population distribution, but as evidence for more tasks are observed it should become more student specific. A *score* for a Bayes net can be any statistic of that joint distribution, although there are a few commonly used choices.

The most commonly used score is the marginal probability of a single proficiency variable. This is usually a vector of numbers between 0 and 1 that sum to 1, with one number corresponding to each state of the proficiency variable. For example $\{Low : .05, Medium : .80, High : .15\}$ corresponds to a student who is thought to be in the middle category with probability .8. This probability of mastery is a combination of the *direct evidence* of mastery for that skill (from nodes that are descendants of the target proficiency), the probability of mastery in the population (from the proficiency model) and *indirect evidence* of mastery from other proficiency variables which are correlated with the target variables. In this sense, it resembles the augmented scores of Wainer et al. (2001).

Typically the proficiency variables are ordered categorical variables. One way to summarize the posterior distribution over those scores is to give the modal value (in the example above, this would be 'M'). This is sometimes called the MAP (maximum a posteriori) estimate. The probability that the person is in the middle category (.8 in the example) is then a measure of certainty of the MAP estimate. If one is willing to assign a real value to each category, it is also possible to calculate an expected value (EAP or expected a posteriori) and a variance. This is often useful for rank ordering students on a particular attribute.

Sometimes the most likely probability profile is of interest. This is not necessarily the same as the MAP estimate for each individual proficiency. Fortunately, a variation of the junction tree algorithm (that uses maximization in place of summation) is available that produces the most likely configuration. This is supported by most Bayes net engines. In general, the engines can calculate the joint distribution over any set of proficiency variables, correctly taking into account the correlation among the variables.

It is also straightforward to sample from a Bayesian network (either using the prior distribution or the posterior distribution given partial evidence). First, use the message passing algorithm to calculate the marginal distribution for a single node. Sample from that distribution, and instantiate the node to the sampled value.

Propagate the new information and sample the next node. The result is a sample from the joint distribution of all variables. Many software packages support this operation, which can be used for simulation studies. Again, this may be useful for other CDMs. If they can first be converted into a Bayes net, then, if more specialized software is not available, Bayes net software can be used to draw samples for simulation studies.

4.2.2 Weight of Evidence

At the student level, it is also possible to provide an explanation of how the model arrived at a particular score. Madigan, Mosurski, and Almond (1997) suggest a statistic called the weight of evidence. Let H be a binary hypothesis, and let $\overline{H}$ be its complement. This can be any partition of the space of possible proficiency profiles; however, the most frequent example is corresponds to $X_k \geq x$ for some proficiency variable X_k. Let E be a particular observation (which corresponds to a set of observable values $Y_1, \ldots, Y_J$ taking on values $y_1, \ldots, y_j$). The *weight of evidence that E provides for H*, $W(H{:}E)$ is

$$W(H{:}E) = \log \frac{\Pr(E|H)}{\Pr(E|\overline{H})} = \log \frac{\Pr(H|E)}{\Pr(\overline{H}|E)} - \log \frac{\Pr(H)}{\Pr(\overline{H})} \,. \tag{4.3}$$

Note that the right hand size of Eq. 4.3 is Bayes' theorem using log-odds. The posterior log-odds is equal to the prior log-odds plus the log of the likelihood ratio. Similar to the way that Bayes theorem applications can be chained so can the weight of evidence. However, this requires defining a conditional weight of evidence,

$$W(H{:}E_2|E_1) = \log \frac{\Pr(E_2|H, E_1)}{\Pr(E_2|\overline{H}, E_1)} \,. \tag{4.4}$$

It follows that $W(H{:}E_1, E_2) = W(H{:}E_1) + W(H{:}E_2|E_1)$, and this calculation can be extended for any number of observations. Using this model, a contribution can be assigned to each piece of evidence $E_1, \ldots, E_J$ (that is the observations from Tasks 1 through J). Note that the conditional weights of evidence are order sensitive (pieces of evidence observed earlier in the sequence tend to provide more weight) but that the total evidence is always the same no matter the sequence it is observed in.

Madigan et al. (1997) suggest a graphical display showing the contribution of each piece of evidence in sequence. Almond, Kim, Shute, and Ventura (2013) apply this to the game *Physics Playground*, which was scored using a Bayes net. In Fig. 4.2, the large bars in the third column represent game levels where the estimate of the student's physics ability had a large shift. Looking at replays of those levels helped the researchers understand how the student was approaching the problems and indicated game levels that might need additional work.

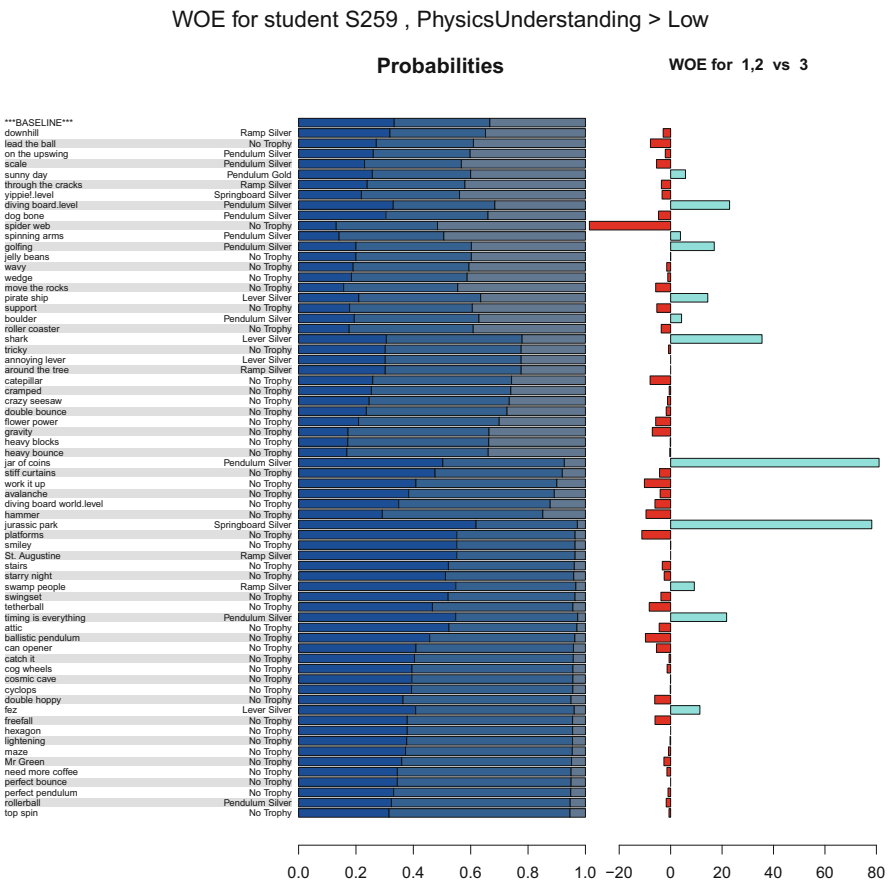


Fig. 4.2 Weight of evidence balance sheet for student S259 (Copyright Almond et al., 2013. Used by permission of the authors.) The leftmost column shows the observations in the order processed. The middle column shows the running probability that the target node (Physics Understanding) is in the high, medium or low state. The rightmost column shows the weight of evidence, cyan bars for positive evidence and red for negative evidence

Another use for weight of evidence is in item or task selection. Let $\{y_{jm}, m = 1, \ldots, M\}$ represent the possible outcomes of the observation Y_j. Then the *expected weight of evidence* for Y_j for the hypothesis H is

$$EW(H{:}E) = \sum_{m=1}^{M} W(H{:}y_{jm}) \Pr(y_{jm} \mid H) \, . \tag{4.5}$$

The joint expected weight of evidence can also be calculated across several observables. A simple task selection algorithm would select the task at any time that has the greatest EW. Note that this is a greedy search algorithm and does not always yield an optimal form of the assessment.

4.3 Estimation of Model Parameters

Algorithms for estimating Bayesian networks from data is an active topic of research in computer science (Neapolitan, 2004). These algorithms can be roughly divided into ones which take the structure of the Bayesian network as fixed (this section), and ones which attempt to learn the structure from data (Sect. 4.4). The difference is similar to the difference between exploratory and confirmatory factor analysis. Both classes of algorithm can take advantage of the graphical structure of the model.

Educational models have two features which are only sometimes present in the broader class of Bayesian network models. First, the proficiency variables (attributes) are usually latent variables. (Some of the learning algorithms require fully observed data). Second, the variables (both proficiency variables and observable outcomes) are usually ordinal in educational applications, and the CPTs are monotonic (greater skill should lead to better performance). Most parameter estimation algorithms do not force monotonicity constraints on the CPTs.

Rather than exhaustively survey all of the available parameter estimation techniques, this chapter focuses on two. First, it describes the basic expectation maximization (EM) algorithm with Bayes nets (Sect. 4.3.1), which is supported by many Bayes net software packages. Second, it describes some considerations necessary when doing Markov chain Monte Carlo (MCMC) with Bayes nets (Sect. 4.3.2).

4.3.1 EM Algorithms

If all nodes of a discrete Bayesian network are fully observed in a sample of size N, the CPTs can be updated via a simple counting algorithm. Consider first node X_k which has no parents and which can take on S states. Let $\mathbf{p}_k = (p_{k1}, \ldots, p_{kS})$ be probabilities for the S values X_k can take. The data will follow a multinomial distribution with parameters N and $\mathbf{p}$. The natural conjugate prior for the multinomial family is the Dirichlet distribution, with parameters $\mathbf{a}_k = (a_{k1}, \ldots, a_{kS})$ (which must all be non-negative). If $\mathbf{y}_k = (y_{k1}, \ldots, y_{kS})$ are the observed counts, and $\tilde{\mathbf{a}}_k$ are the prior parameters, then $\mathbf{a}_k^* = (\tilde{a}_{k1} + y_{k1}, \ldots, \tilde{a}_{kS} + y_{kS})$ are the posterior parameters. Note that $N = \sum_{s=1}^{S} y_{ks}$, so that the prior distribution can be expressed as a expected probability, $\tilde{\mathbf{p}}_k$ and a pseudo-count $\tilde{N}$, with higher counts expressing more certainty in the prior.

Next consider an arbitrary node, X_k with parents $\text{pa}(X_k)$. Let c be an index which runs over the configurations of the parent variables. The observed data are now classified two ways, so that $\mathbf{Y}_k$ is now a matrix, with y_{kcs} the number of cases where $X_k = x_{ks}$ when $\text{pa}(X_k)$ are in configuration c. Here $\mathbf{Y}_k$ follows a *conditional multinomial distribution* where $N_{kc} = \sum_{s=1}^{S} y_{kcs}$. Under the *local parameter*

independence assumption[1] the rows of the CPT are independent. As before, assign each row a Dirichlet distribution with parameters $\tilde{\mathbf{a}}_{kc}$ and stack these vectors to make a matrix $\tilde{\mathbf{A}}_k$ the prior distribution for the CPT is a *hyper-Dirichlet* distribution with parameter $\tilde{\mathbf{A}}_k$. The posterior distribution is also a *hyper-Dirichlet* distribution with parameter $\mathbf{A}_k^* = \tilde{\mathbf{A}}_k + \mathbf{Y}_k$. As in the no parent case, this can be expressed as an expected conditional probability table $\mathbf{P}_k^*$ and vector of pseudo counts for each row, $N_{kc}^* = \sum_{s=1}^{S} a_{kcs}^*$. The pseudo-counts for the rows need not be the same, and the posterior pseudo-counts will depend on the pattern of observations.

Spiegelhalter and Lauritzen (1990) show that under the *global parameter independence assumption* (parameters from different CPTs are independent given the observed data) and the local parameter independence assumption, that the global hyper-Dirichlet distribution (the one where every CPT has a hyper-Dirichlet prior) is the natural conjugate of the conditional multinomial distribution of the Bayes net. Thus, when the data are fully observed a simple counting algorithm can be used to calculate either the maximum likelihood or Bayesian estimate of the CPTs. The maximum likelihood estimate can have difficulties if one of the cell counts is zero, so generally the Bayesian estimates are preferred. A Bayesian prior can be created by putting expert values into the CPTs and then assigning them an effective samples size weight $\tilde{N}_k$. A noninformative prior can be produced by setting all of the cells to $1/S$, however, when informative priors are available they can help keep the estimates from violating the monotonicity assumptions when the sample size is small. In our experience, giving the priors an effective sample size between 10 and 100 strikes a good balance between making them responsive to data, and stabilizing the estimates when sample sizes are small.

If there are missing values, or latent variables, then the CPTs can be calculated using the expectation maximization (EM) algorithm. Under the global parameter independence assumption, the observed table of counts, $\mathbf{Y}_k$ is a sufficient statistic for the CPT for node X_k. Let $\mathbf{A}_k^{(0)}$ be the initial parameters for the distribution. Dropping these into the Bayes net produces initial values for all of the conditional probability tables. The E-step (expectation step) of the EM algorithm calculates the expected value of the sufficient statistics, $\hat{\mathbf{Y}}_k^{(r)}$ from the current values of the CPTs at Step (r). This is done variable-by-variable. For each variable create a table filled with zeros. Now for each observation, look at the X_k and $\text{pa}(X_k)$. If these are fully observed, add one to the appropriate cell. If not, calculate the joint probability distribution over the variables using the junction tree algorithm. Then add that probability distribution to the table. The M-step (maximization step) now just uses the counting rule to produce new posterior parameters. These are used to calculate the new CPTs which are dropped into the Bayes net. Both the counting algorithm and the EM algorithm are built into many Bayesian network software packages.

[1]This is different from the usual psychometric *local independence assumption* which assumes that the observable outcomes from different items or tasks are independent given the proficiency variables.

There are two problems with using the EM algorithm with hyper-Dirichlet priors to estimate the CPTs of a Bayes net in educational applications. First, the hyper-Dirichlet does not support the monotonicity constraint, so there is no guarantee that increasing skill will lead to increasing probability of success in the estimated CPTs. Second, if the parent variables are moderately correlated (often true in many educational settings) then the amount of information in certain configurations of the parent variables (rows of the table) will be small. In particular, if the there will be very few cases in which the test taker is very high on one skill and very low on another skill. This increases the chance that the estimated CPT will not satisfy the monotonicity constraint.

To get around this problem, Almond (2015) proposed an extension to the usual EM algorithm for CPTs that uses a parameterization other than the hyper-Dirichlet distribution. The E-step is the same as the one for the hyper-Dirichlet case, as the sufficient statistics are still the table of expected counts $\hat{\mathbf{Y}}_k^{(r)}$. The M-step then applies an optimizer to get the parameters for the CPT that are most likely to produce the observed data. The optimizer only needs to run for a few steps to produce a generalized EM algorithm. The `Peanut` (Almond, 2017b) package in `R` (R Development Core Team, 2007) is available to use this algorithm.

As many other CDMs can be expressed as Bayesian networks, parameter estimation software designed for those models can be used to estimate Bayes net CPTs. A particularly important case is structural equation models (SEMs). If the model graph for the SEM can be turned into a directed graph (by orienting bidirectional edges), then the biggest difference between the discrete Bayes net and the SEM is that in the latter, the latent variables are continuous. In this case, it is straightforward to convert the SEM to a Bayes net (Almond, 2010).

4.3.2 Markov Chain Monte Carlo

An alternative to the EM algorithm is Markov chain Monte Carlo (MCMC). Once again the graphical structure can provide guidance for the algorithm. In particular, a Gibbs sampler only needs to look at the Markov blanket of a particular node (the parents, children and certain siblings) when sampling the value of a particular node. Also, if the global independence assumption holds, the conditional probability tables for each node can be considered independently. If all of the conditional probability tables have a hyper-Dirichlet distribution, then the posterior distribution of a discrete Bayesian network can be sampled using a Gibbs sampler (Spiegelhalter & Lauritzen, 1990).

There are two model identification issues that need to be addressed. First, unless the conditional probability tables are constrained to be monotonic, there exists a possibility for label switching (Frühwirth-Schnatter, 2001). Suppose a Bayes net is constructed with binary skills in which 1 represents mastery and 0 lack of mastery.

Then the model with 0 representing mastery will have equal likelihood. Frühwirth-Schnatter (2001) suggests letting the chains mix over the states with inverted labels and then sorting the labels out after the sampling is complete.

The second issue is related to the scale invariance in IRT models. In an IRT model if a constant is added to the difficulty of all items and added to the ability of all students, then the likelihood is unchanged. The same happens in the Bayes nets through shifts to the conditional probabilities. Bafumi, Gelman, Park, and Kaplan (2005) recommend letting the model range over the possible values and then centering scale after the MCMC run. Note that each proficiency variable must have its scale anchored. Almond, Mislevy, and Yan (2007) recommend constructing a set of items which have an average of 0 difficulty to identify each scale.

There is no Bayes net package currently available for doing MCMC with educational models. (StatShop; Almond, Yan, Matukhin, & Chang, 2006; was developed for this purpose but is only available as a research release.) Building the sampler in a general purpose MCMC package such as OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009) or JAGS (Plummer, 2012) requires a great deal of coding. Another problem is that because of the discrete variables the deviance is not automatically calculated for model fit measures such as DIC (Spiegelhalter, Best, Carlin, & van der Linde, 2002) or WAIC (Gelman et al., 2013, Chapter 7).

4.4 Assessment of Fit

The question of model fit to data encompasses several specific questions: (1) How well does a given individual fit the assessment models? (Sect. 4.4.1) (2) Are the modeling assumptions for a particular item or task correct? (Sect. 4.4.2) (3) Is the current assessment design adequate for estimating the proficiency of candidates? (Sect. 4.4.3) (4) What is the best Bayes net structure for fitting a given set of data? (Sect. 4.4.4) Almond et al. (2015) provide fairly extensive discussion of methods for the first three. The last topic is the subject of fairly extensive research in the artificial intelligence community under the names *structural learning* and *causal modeling* (Neapolitan, 2004). These developments are only briefly surveyed here and this survey is almost certainly out of date at the time of printing.

4.4.1 Person Fit

Because the normalization constant calculated through the junction-tree algorithm is the likelihood for the observed response pattern (for a randomly selected person from the target population of the assessment), it is relatively simple to create fit metrics based on likelihood. In particular, calculating the likelihood of all observed

response patterns provides a reference distribution for the person likelihood statistic. Looking for outliers in this distribution provides a way of flagging unusual response patterns.

The weight of evidence balance sheet shown in Sect. 4.2.2 provides another mechanism for spotting unusual response patterns. Unusually large positive or negative evidence indicates that the candidate made an unusual response to a particular task. This has proved useful for both identifying unusual response vectors and identifying problematic tasks (Almond et al., 2013).

4.4.2 Item and Task Fit Measures

Just as the junction-tree algorithm can rapidly calculate marginal distributions for proficiency variables, it can rapidly calculate predictive distributions for observable variables given other observations. These can be used for a wide variety of item fit and task fit measures (Almond et al., 2015). The *observable characteristic plot* which gives the discrete analog of the item characteristic curve is particularly promising (Sinharay, Almond, & Yan, 2004; Sinharay & Almond, 2007).

The display focuses on a single CPT; for simplicity assume that the child variable is binary. Let $c \in \{1, \ldots, C\}$ be an index for the parent configurations. (Note that in some test designs, certain configurations may be indistinguishable. In these cases, replace the parent configurations with equivalence classes of configurations.) The CPT gives a predictive probability p_c for the observable under each configuration.

If the parent variables are fully observed, a simple counting algorithm can be used to update the conditional probability tables. Let c be a configuration of the parent variables, let N_c be the total number of cases that have that configuration, and let X_c be the number of successful cases with that configuration. The value X_c/N_c provides an estimate for p_c. As it is possible that $X_c = 0$ or $X_c = N_c$, it is better to use a Bayesian estimate using either a uniform prior or using p_c as a prior. This latter gives a beta distribution with parameters $a_c = X_c + p_c$ and $b_c = N_c - X_c + 1 - p_c$ (1/2 can be substituted for p_c in those expressions). Use the beta distribution to calculate a 95% interval for p_c and plot that against p_c for each parent configuration. Figure 4.3 shows two examples.

Figure 4.3a is a conjunctive CPT that works fairly well. The conditional probability values which should be low are low, and the value which should be high is high. Figure 4.3b shows a conjunctive CPT that does not work well. The conditional probability when the parents are in configuration $(1, 0)$ should be low, but the data are indicating it is moderate. The conjunctive combination rule is not appropriate for this task. These plots were made with the R package `CPTtools` (Almond, 2017a).

When the parent variables are not fully observed (the usual case), then expected configurations of the parent distributions can be used. Let Y_{ik} be the value of observ-

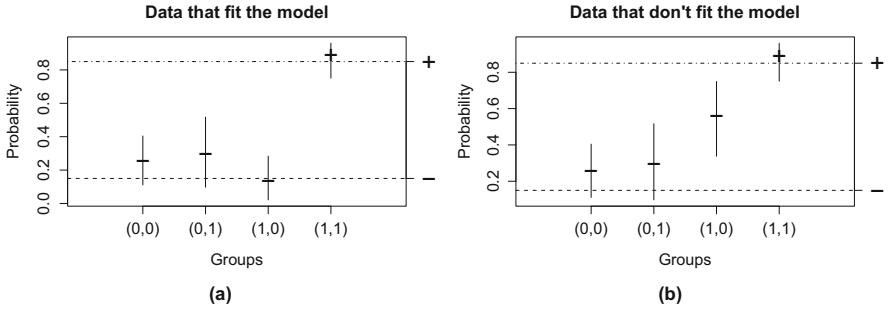


Fig. 4.3 Observable characteristic plot. (Reprinted with permission from ETS)

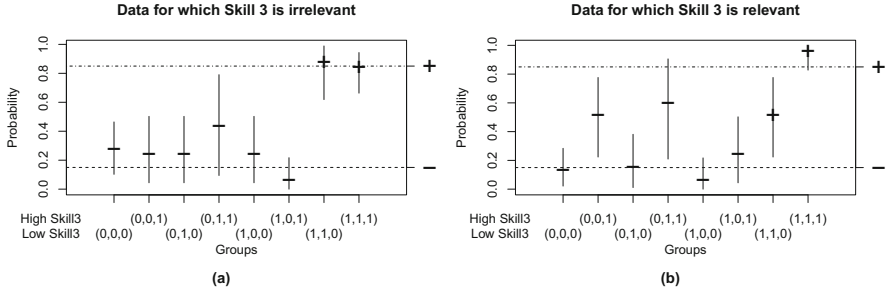


Fig. 4.4 Observable characteristic plot for additional skill. (Reprinted with permission from ETS)

able Y_k for test taker i, and let $q_{ic} = \Pr(\text{pa}(Y_k) = c | Y_{i1}, \ldots, Y_{i(k-1)}, Y_{i(k+1)}, Y_{iK})$, that is the probability that test taker i is in configuration c given all of the other observations in the model. This can be calculated using the Bayes net to find the joint distribution over $\text{pa}(Y_k)$ or via MCMC. Then $N_c = \sum_i q_{ic}$ and $X_c = \sum_i Y_{ik} q_{ic}$. This can be used to form beta credibility intervals as above.

Note that a_c and b_c form a $2 \times C$ contingency table, with expected values $N_c p_c$ and $N_c(1 - p_c)$, respectively. This can be used to calculate a χ^2 statistic with C degrees of freedom. This statistic may not have a χ^2 distribution if the expected values are for the parent variables, but it can be used as a heuristic for screening which observables to examine in detail.

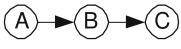
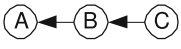
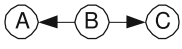
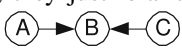
A simple extension of this procedure can be used to test for missing parent variables (Fig. 4.4). Simply create an augmented parent set $\text{pa}^*(Y_k) = \text{pa}(Y_k) \cup \{Y'\}$ and produce the observable characteristic plot using this augmented set. Figure 4.4a shows a plot which works fairly well. The conditional probabilities seem to have the same value for all values of the extra variable (the fastest moving one). Figure 4.4b shows an example that does not work as well. If the extra variable is in the higher state, the task appears to be easier. In this case, Y' is a possible missing parent.

4.4.3 Simulation Studies for Profile Recovery

Another important question in diagnosing problems with an assessment is whether or not the mix of tasks and items provides adequate coverage for the intended purpose. As most Bayes net engines support simulating from the joint distribution, it is straightforward to do a simulation experiment. First simulate both the proficiency variables and observable outcome variables for the target test form. Next, mask the proficiency variables, and estimate them using the observed data. If MAP estimates are used, this produces a 2×2 table for each proficiency variable comparing the actual and estimated proficiencies. Also, using the marginal distributions for the proficiency variables, an expected table can be constructed (Almond et al., 2015).

Two statistics are useful for evaluating how close the actual and estimated proficiencies are. The first is Cohen's kappa. The second is Goodman and Kruskal's lambda. The latter is similar to kappa, except it takes as its reference point not random agreement, but assigning everybody to the most prevalent category. Therefore, it provides an estimate of the improvement in classification accuracy over not testing at all.

4.4.4 Learning Model Structure from Data

As mentioned previously, there is a substantial literature on inferring the graphical structure from a set of data (e.g., Neapolitan, 2004). Much of this literature is devoted to discovering causal structure, assuming that the arrows of the directed graph point in a causal direction. While the algorithms often reveal that two nodes are connected, often they cannot determine the direction of the edge. In particular, the three graphs  ,  and  all express the same conditional independence relationships, they just re-arrange the ordering of the variables. On the other hand, the graph  has different conditional independence conditions, so can be distinguished from the others.

Another problem is that latent variables can be added as full or partial mediators between any two observed variables yielding the same probability distribution over the observed variables. As educational models usually center around latent proficiency variables, additional information is needed. Often an exploratory factor analysis, followed by discretizing the latent variables works well. Martin and VanLehn (1994) develop a discrete factor analysis for use with Bayesian networks.

4.5 Exemplary Applications

This section looks at two applications. ACED (Sect. 4.5.1) is an assessment and tutoring system that uses many simple tasks, each with a single observable outcome. NetPASS (Sect. 4.5.2) is a simulation-based assessment that has multiple observable outcomes per task.

4.5.1 ACED

Adaptive content with evidence-based diagnosis (ACED) is an adaptive, diagnostic assessment of mathematics sequences that makes use of a Bayesian network scoring model (Shute, Graf, & Hansen, 2005; Shute, Hansen, & Almond, 2007). The ACED prototype (a) uses expected weight of evidence to select the next task (Madigan & Almond, 1995), (b) implements targeted diagnostic feedback (Shute, Hansen, & Almond, 2008), and (c) uses technology to make it accessible to students with visual disabilities (Shute et al., 2007).

The ACED proficiency model is represented as a tree-shaped Bayesian network with an overall sequences proficiency node at the top and nodes for arithmetic, geometric, and other recursive sequences as its immediate children (Graf, 2003). The proficiency model consisted of 42 nodes. A total of 174 items were developed, 63 of them were connected to nodes in the geometric sequences branch of the proficiency model. Human experts provided information for the evidence model (i.e., indicators of the relative difficulty of the tasks associated with a particular node, and a correlation that indicated the strength of the relationship between the node and its parent). Evidence rules for each proficiency were defined at three levels of proficiency (high, medium and low). Prior probabilities for top-level proficiencies were elicited based on the probability that a student is at each of the three proficiency levels while prior probabilities for the other proficiency nodes was specified based on relative difficulty of the tasks associated with each of them. A task selection algorithm based on the expected weight of evidence was employed (Madigan & Almond, 1995).

The assessment-instruction cycle implemented in ACED included the following steps: (1) selecting an instructional area for testing based on heuristics, (2) identifying tasks associated to the current instructional area, (3) calculating the expected weight of evidence for each task, (4) selecting the task with highest expected weight of evidence, (5) administering the task, (6) updating the student's proficiency model based on the response, and (7) stop or iterate depending on a predefined criteria.

A study comparing the effects of adaptive sequencing and feedback showed that students benefited the most (greater pre-post learning gains) from elaborated feedback in the adaptive sequencing condition (Shute et al., 2008) and found out that despite the presence of feedback, the ACED system showed strong psychometric properties (i.e., split-half reliability of 63 ACED tasks associated with Geometric

Sequences was high, 0.88, and the top parent proficiency reliability was 0.88). The complete data for ACED, as well as a complete model description (including the Q-matrix) are available at https://ecd.ralmond.net/ecdwiki/ACED/ACED/.

4.5.2 NetPass

The Networking Performance Skill System (NetPASS) project is a performance-based assessment for designing and troubleshooting computer networks (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004; Williamson, Bauer, Steinberg, Mislevy, & DeMark, 2004; Levy & Mislevy, 2004). The main purpose of NetPASS is to provide students with opportunities to practice their computer networking skills and receive diagnostic feedback based on their performance. The NetPASS prototype informed the development of other simulation system and game systems in this domain that have been used by many students around the world (Scalise et al., 2007).

The NetPASS student model includes variables such as *Networking Disciplinary Knowledge* (top node), *Network Modeling* and *Networking Proficiency* (top nodes' children) and *Designing*, *Implement/Configure*, and *Troubleshoot* (Children of Network Proficiency) (Levy & Mislevy, 2004).

The design team and a group of subject matter experts worked on identifying features of work products that provide evidence for particular claims. For example, aspects of the network that students should verify during troubleshooting and the evidence that could be elicited from students' troubleshooting processes. Cognitive Task Analysis (CTA) (Newell & Simon, 1972) was used to identify task features and situations for eliciting student behaviors of interest to measure the intended construct (Williamson et al., 2004). Data from 24 students at three ability levels (8 lower, 8 average, 8 high) of the Cisco Networking Academy Program curriculum were used to inform the CTA. Students took a pretest and solved four scenarios following a think-aloud protocol. Data collected included transcripts of think-aloud solutions, log files, diagrams, and calculations. These data were analyzed and discussed by the researchers and subject matter experts. Reusable observables and patterns of behaviors associated with claims were identified.

In another study, Levy and Mislevy (2004) used data from 216 test takers to estimate the parameters of a Bayesian model for NetPASS using MCMC. Data included an average of 28 values for each observable. Three chains were run in parallel for 100,000 iterations using different starting values and convergence diagnostics were collected. Results showed an increase of precision on the posterior distributions for the parameters that define the conditional probability distributions for most of the parameters. Higher precision was observed at the observable level while mild increases in precision were observed for latent variables (i.e., student and evidence model variables). The authors elaborated on the broad application of Bayesian approaches for modeling educational assessments and described future

work involving the comparison of the existing model to models that require fewer parameters and evaluating the need for adding context variables to the evidence models.

4.6 Bayes Net Software Packages

A number of software packages are available for basic manipulation of Bayesian networks. These packages include academic and commercial packages. A list of available software can be found here (http://www.cs.ubc.ca/~murphyk/Software/bnsoft.html). There is also a Wikipedia article on Bayesian networks that includes general information about software (https://en.wikipedia.org/wiki/Bayesiannetwork).

A set of tools that allows domain experts to quickly build Bayesian networks from tabular representations (e.g., using a spreadsheet program) and make use of software such as Netica (Norsys, 2004) to make them functional is now available (Almond, 2017c). These tools include:

- `RNetica`: serves as the "glue" layer between the open source statistical programming language R and the proprietary Bayesian network engine Netica. RNetica uses the functionality of the Netica API inside of R.
- `CPTtools`: includes R functions for constructing and manipulating conditional probability tables. It also contains tools for displaying and analyzing the output of Bayesian network analyses.
- `Peanut`: an object oriented layer designed to rest on top of CPTtools. Peanut was designed as a high level interface that is compatible with multiple Bayesian network engines.

These tools are available at https://pluto.coe.fsu.edu/RNetica/.

4.7 Discussion

This chapter describes processes for authoring, calibrating and evaluating Bayesian networks that can provide diagnostic information. As Bayesian networks are a notation for describing models, many popular CDMs can be represented as Bayes nets. This in turn, allows the users of these models to exploit existing Bayes net software for building systems with embedded scoring for use in simulation experiments.

The use of Bayesian networks for designing and implementing diagnostic models has long been recognized. Among the ten reasons for considering Bayesian networks cited in Almond et al. (2015), are capabilities such as: reporting scores in terms of "Probability of Claim;" using a graphical representation for the proficiency model; incorporating expert knowledge about the cognitive domain; learning from

data; handling complex models and tasks; being fast; providing profile scores and real-time diagnosis; easily employed in the context of evidence-centered design; and their models can be considered "useful."

The examples provided show that Bayesian networks can be used to implement diagnostic models for adaptive assessment and learning systems that provide teachers and students with relevant feedback (Shute et al., 2008; Almond, Shute, Underwood, & Zapata-Rivera, 2009).

References

Almond, R. G. (2010). I can name that Bayesian network in two matrixes. *International Journal of Approximate Reasoning, 51*, 167–178. Retrieved from https://doi.org/10.1016/j.ijar.2009.04.005

Almond, R. G. (2015). An IRT-based parameterization for conditional probability tables. In J. M. Agosta & R. N. Carvalho (Eds.), *Bayesian Modelling Application Workshop at the Uncertainty in Artificial Intelligence (UAI) Conference*, Amsterdam, The Netherlands. Additional material available at http://pluto.coe.fsu.edu/RNetica/

Almond, R. G. (2017a). CPTtools: R code for constructing Bayesian networks (0–4.3 ed.) [Computer software manual]. Retrieved from http://pluto.coe.fsu.edu/RNetica/CPTtools.html (Open source software package).

Almond, R. G. (2017b). Peanut: An object-oriented framekwork for parameterized Bayesian networks (0–3.2 ed.) [Computer software manual]. Retrieved from http://pluto.coe.fsu.edu/RNetica/Peanut.html (Open source software package).

Almond, R. G. (2017c). RNetica: Binding the Netica API in R (0–5.1 ed.) [Computer software manual]. Retrieved from http://pluto.coe.fsu.edu/RNetica/RNetica.html (Open source software package).

Almond, R. G. (2017d). Tabular views of Bayesian networks. In J. M. Agosta & T. Singliar (Eds.), *Bayesian Modeling Application Workshop at the Uncertainty in Artificial Intelligence (UAI) Conference*, Sydney, Australia.

Almond, R. G., DiBello, L. V., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S., et al. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola & T. Richardson (Eds.), *Artificial Intelligence and Statistics 2001* (pp. 137–143). San Francisco, CA: Morgan Kaufmann.

Almond, R. G., Kim, Y. J., Shute, V. J., & Ventura, M. (2013). Debugging the evidence chain. In R. G. Almond & O. Mengshoel (Eds.), *Proceedings of the 2013 UAI Application Workshops: Big Data Meet Complex Models and Models for Spatial, Temporal and Network Data (UAI2013AW)*, Aachen, Germany (pp. 1–10). Retrieved from http://ceur-ws.org/Vol-XXX/paper-01.pdf

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–238.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

Almond, R. G., Mislevy, R. J., Williamson, D. M., & Yan, D. (2007). *Bayesian networks in educational assessment*. Paper presented at Annual Meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Almond, R. G., Mislevy, R. J., & Yan, D. (2007). *Using anchor sets to identify scale and location of latent variables.* Paper Presented at Annual Meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J.-D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning, 50*, 450–460. https://doi.org/10.1016/j.ijar.2008.04.011

Almond, R. G., Yan, D., Matukhin, A., & Chang, D. (2006). *StatShop testing* (Research Memorandum No. RM-06-04). Princeton, NJ: Educational Testing Service.

Andersen, S. K., Olesen, K. G., Jensen, F. V., & Jensen, F. (1989). Hugin—A shell for building Bayesian belief universes for expert systems. In *IJCAI'89*, Detroit, MI. Reprinted in Shafer and Pearl (1990).

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practial issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis, 13*, 171–187. https://doi.org/10.1093/pan/mpi010

BayesFusion, LLC. (2017). *Genie.* Retrieved from http://bayesfusion.com (Bayesian network Computer Software).

Bayesia, S. A. S. (2017). *BayesiaLab*. Retrieved from http://www.bayesia.com (Bayesian Network Computer Software).

Behrens, J. T., Mislevy, R. J., Bauer, M. I., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Measurement*, *4*, 295–301.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. New York, NY: Springer.

Díez, F. J. (1993). Parameter adjustment in Bayes networks. The generalized noisy or-gate. In D. Heckerman & A. Mamdani (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the 9th Conference* (pp. 99–105). San Francisco, CA: Morgan-Kaufmann.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association, 96*(453), 194–209.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall. The third edition is revised and expanded and has material that the earlier editions lack.

Graf, E. A. (2003). *Designing a proficiency model and associated item models for a mathematics unit on sequences*. Paper Presented at the Cross Division Math Forum, Princeton, NJ.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. New York, NY: Springer.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kim, J. H., & Pearl, J. (1983). A computational model for causal and diagnostic reasoning in inference systems. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence* (pp. 190–193). Karlsruhe, Germany: William Kaufmann.

Lauritzen, S. L. (1996). *Graphical models*. New York, NY: Oxford University Press.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a simulation-based assessment. *International Journal of Measurement, 4*, 333–369.

Li, Z., & D'Ambrosio, B. (1994). Efficient inference in Bayes nets as a combinatorial optimization problem. *International Journal of Approximate Reasoning, 11*, 55–81.

Lunn, D. J., Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2009). The BUGS project: Evolution, critique and future directions (with discussion). *Statistics in Medicine, 28*, 3049–3082.

Madigan, D., & Almond, R. G. (1995). Test selection strategies for belief networks. In D. Fisher & H. J. Lenz (Eds.), *Learning from data: AI and statistics V* (pp. 89–98). New York, NY: Springer.

Madigan, D., Mosurski, K., & Almond, R. G. (1997). Graphical explanation in belief networks. *Journal of Computational Graphics and Statistics, 6*(2), 160–181. Retrieved from http://www.amstat.org/publications/jcgs/index.cfm?fuseaction=madiganjun

Martin, J., & VanLehn, K. (1994). *Discrete factor analysis: Learning hidden variables in Bayesian networks* Technical report No. LRDC-ONR-94-1. LRDC, University of Pittsburgh.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective, 1*(1), 3–62.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. *Applied Psychological Measurement, 16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Englewood Cliffs, NJ: Prentice Hall.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Norsys, Inc. (2004). Netica [Computer software manual]. Retrieved from http://www.norsys.com

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Plummer, M. (2012). JAGS version 3.2.0 user manual (3.2.0 ed.) [Computer software manual]. Retrieved from http://mcmc-jags.sourceforge.net/

R Development Core Team. (2007). R: A language and environment for statistical computing [Computer software manual], Vienna. Retrieved from http://www.R-project.org

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34*(4), (Part 2).

Scalise, K., Bernbaum, D. J., Timms, M., Harrell, S. V., Burmester, K., Kennedy, C. A., et al. (2007). Adaptive technology for e-learning: Principles and case studies of an emerging field. *Journal of the American Society for Information Science and Technology, 58*(14), 2295–2309. Retrieved from https://doi.org/10.1002/asi.20701

Shute, V. J., Graf, E. A., & Hansen, E. G. (2005). Designing adaptive, diagnostic math assessments for individuals with and without visual disabilities. In L. M. Pytlikzillig, R. H. Bruning, & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 169–202). Charlotte, NC: Information Age Publishing.

Shute, V. J., Hansen, E. G., & Almond, R. G. (2007). *An assessment for learning system called ACED: The impact of feedback and adaptivity on learning.* Research report No. RR-07-26. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/research/researcher/RR-07-26.html

Shute, V. J., Hansen, E. G., & Almond, R. G. (2008). You can't fatten a hog by weighing it—Or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education, 18*(4), 289–316. Retrieved from http://www.ijaied.org/iaied/ijaied/abstract/Vol_18/Shute08.html

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitively diagnostic models—A case study. *Educational and Psychological Measurement, 67*(2), 239–257.

Sinharay, S., Almond, R. G., & Yan, D. (2004). *Assessing fit of models with discrete proficiency variables in educational assessment*. Research Report No. RR-04-07. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/research/researcher/RR-04-07.html

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society (Series B), 64*, 583–639.

Spiegelhalter, D. J., & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks, 20*, 579–605.

Srinivas, S. (1993). A generalization of the noisy-or model, the generalized noisy or-gate. In D. Heckerman & A. Mamdani (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the 9th Conference* (pp. 208–215). San Mateo, CA: Morgan Kaufmann.

Wainer, H., Veva, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., et al. (2001). Augmented scores—"borrowing strength" to compute scores based on a small number of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–388). Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Bauer, M. I., Steinberg, L. S., Mislevy, R. J., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. *International Journal of Testing, 4*, 303–332.

Chapter 5
Nonparametric Methods in Cognitively Diagnostic Assessment

Chia-Yi Chiu and Hans-Friedrich Köhn

Abstract Parametric estimation is the prevailing method for fitting diagnostic classification models. In the early days of cognitively diagnostic modeling, publicly available implementations of parametric estimation methods were scarce and often encountered technical difficulties in practice. In response to these difficulties, a number of researchers explored the potential of methods that do not rely on a parametric statistical model—nonparametric methods for short—as alternatives to, for example, MLE for assigning examinees to proficiency classes. Of particular interest were clustering methods because efficient implementations were readily available in the major statistical software packages. This article provides a review of nonparametric concepts and methods, as they have been developed and adopted for cognitive diagnosis: clustering methods and the Asymptotic Classification Theory of Cognitive Diagnosis (ACTCD), the Nonparametric Classification (NPC) method, and its generalization, the General NPC method. Also included in this review are two methods that employ the NPC method as a computational device: joint MLE for cognitive diagnosis and the nonparametric Q-matrix refinement and reconstruction method.

5.1 Introduction

Cognitive diagnosis (CD), a relatively recent development in educational measurement (DiBello, Roussos, & Stout, 2007; Haberman & von Davier, 2007; Leighton & Gierl, 2007; Nichols, Chipman & Brennan, 1995; Rupp, Templin, & Henson, 2010) explicitly targets mastery of the instructional content and seeks to provide

C.-Y. Chiu (✉)
Department of Educational Psychology, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA
e-mail: chia-yi.chiu@gse.rutgers.edu

H.-F. Köhn
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: hkoehn@illinois.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_5

immediate feedback to students about their strengths and weaknesses in terms of skills learned and skills needing study. CD terminology refers to skills, specific knowledge, aptitudes—any cognitive characteristic required to perform tasks—collectively as "attributes." CD models—or "Diagnostic Classification Models" (DCMs), as they are called here—describe an examinee's ability as a composite of these attributes, each of which an examinee may or may not have mastered. Mastery of attributes is recorded as a binary vector; different zero-one combinations define attribute vectors of distinct proficiency classes to which examinees are to be assigned based on their test performance (i.e., examinees' individual attribute vectors must be estimated).

The prevailing method for fitting DCMs uses either marginal maximum likelihood estimation relying on the Expectation Maximization algorithm (MMLE-EM) or Markov chain Monte Carlo (MCMC) techniques (de la Torre, 2009, 2011; DiBello et al., 2007; von Davier, 2008). In the early days of CD modeling, publicly available implementations of these parametric estimation methods were scarce and often encountered technical difficulties in practice (e.g., excessive CPU times, computational feasibility). In response to these difficulties, a number of researchers (Ayers, Nugent, & Dean, 2008, 2009; Chiu, 2008; Chiu & Douglas, 2013; Chiu, Douglas, & Li, 2009; Park & Lee, 2011; Willse, Henson, & Templin, 2007) explored the potential of nonparametric methods—that is, methods that do not rely on a parametric statistical model—as alternatives to MMLE-EM and MCMC for assigning examinees to proficiency classes. Of particular interest were clustering methods because efficient implementations were readily available in the major statistical software packages. Today, efficient implementations of MMLE-EM algorithms for fitting DCMs are available; for example, through the `R` packages `CDM` (Robitzsch, Kiefer, George, & Uenlue, 2016) and `GDINA` (Ma & de la Torre, 2017) (for further software options for fitting DCMs, consult "Part IV" in this book). They work well for large-scale assessments, where the data of hundreds or thousands of examinees are available. However, nonparametric methods are still useful for analyzing assessment data from educational micro-environments, say, for monitoring the instruction and learning process at the classroom level, where CD-based methods would be most useful and needed, but sample sizes are simply too small for maximum likelihood estimation to guarantee reliable estimates of item parameters and examinees' proficiency classes.

This article provides a review of nonparametric concepts and methods, as they have been developed and adopted for CD: clustering methods and the **A**symptotic **C**lassification **T**heory of **C**ognitive **Di**agnosis (ACTCD) (Chiu, 2008; Chiu et al., 2009; Chiu & Köhn, 2015a,b, 2016), the **N**on**P**arametric **C**lassification (NPC) method (Chiu & Douglas, 2013) and its generalization, the **G**eneral NPC method (Chiu, Sun, & Bian, 2018). Further exploration of the potential of nonparametric methods for CD also revealed the particular usefulness of the nonparametric classification methods for implementing joint maximum likelihood estimation for CD (Chiu, Köhn, Zheng, & Henson, 2016) and for the effective refinement and reconstruction of Q-matrices (Chiu, 2013), which are, therefore, included in this review.

5.2 Review of Technical Key Concepts: Cognitive Diagnosis and Diagnostic Classification Models

DCMs are constrained latent class models equivalent to a certain form of finite mixture models (Fraley & Raftery, 2002; Grim, 2006; McLachlan & Basford, 1988; McLachlan & Peel, 2000; von Davier, 2009). Let Y_{ij} denote the response to binary test item j, $j = 1, 2, \ldots, J$, obtained for examinee i, $i = 1, 2, \ldots, N$; the J-dimensional item-score vector of examinee i is written as the row vector $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$. Consider N examinees who belong to M distinct latent proficiency classes. For the general latent class model (Bartholomew, 1987; Bartholomew & Knott, 1999; Heinen, 1996; Langeheine & Rost, 1988; Lazarsfeld & Henry, 1968; Vermunt, 1997, see also Chap. 1 of this book for more details), the (conditional) probability of examinee i in proficiency class $\mathcal{C}_m$, $m = 1, \ldots, M$, answering correctly binary item j is defined by the item response function (IRF) $P(Y_{ij} = 1 \mid i \in \mathcal{C}_m) = \pi_{mj}$, where π_{mj} is constant for item j across all examinees i in proficiency class $\mathcal{C}_m$. The Y_{ij} are assumed independent conditional on proficiency-class membership (local independence); no further restrictions are imposed on the relation between the latent variable—proficiency-class membership—and the observed item response. DCMs, in contrast, constrain the relation between proficiency-class membership and item response such that the probability of a correct response is a function of attribute mastery, as it is determined by an examinee's proficiency class.

Suppose ability in a given domain is modeled as a composite of K latent binary attributes $\alpha_1, \alpha_2, \ldots, \alpha_K$. The K-dimensional binary vector $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \ldots, \alpha_{mK})'$ denotes the attribute vector of proficiency class $\mathcal{C}_m$, where the kth entry, $\alpha_{mk} \in \{0, 1\}$, indicates (non-)mastery of the corresponding attribute. If the attributes do not have a hierarchical structure, then there are $2^K = M$ distinct proficiency classes. The entire set of realizable attribute vectors, given a set of K attributes, defines the latent attribute space (Tatsuoka, 2009). The attribute vector of examinee $i \in \mathcal{C}_m$, $\boldsymbol{\alpha}_{i \in \mathcal{C}_m}$, is written as $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$.

The individual items of a test are also characterized by K-dimensional attribute vectors $\mathbf{q}_j$ that determine which attributes are required to respond correctly to an item ($q_{jk} = 1$, if a correct answer requires mastery of the kth attribute, and 0 otherwise). Given K attributes, there are at most $2^K - 1$ distinct item-attribute vectors (item-attribute vectors that consist entirely of zeroes are considered inadmissible). The J item-attribute vectors of a test constitute its Q-matrix, $\mathbf{Q} = \{q_{jk}\}_{(J \times K)}$, (Tatsuoka, 1985) that summarizes the specific item-attribute associations. The Q-matrix must be known (or the data cannot be analyzed within the CD framework) and complete. A Q-matrix is said to be complete if it guarantees the identifiability of all realizable proficiency classes among examinees (Chiu et al., 2009; Köhn & Chiu, 2017). An incomplete Q-matrix may cause examinees to be assigned to proficiency classes to which they do not belong. Formally, a Q-matrix is complete if the equality

of two expected item response vectors, $\boldsymbol{S}(\boldsymbol{\alpha})$ and $\boldsymbol{S}(\boldsymbol{\alpha}^*)$, implies that the underlying attribute vectors, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, are also identical: $\boldsymbol{S}(\boldsymbol{\alpha}) = \boldsymbol{S}(\boldsymbol{\alpha}^*) \Rightarrow \boldsymbol{\alpha} = \boldsymbol{\alpha}^*$, where $\boldsymbol{S}(\boldsymbol{\alpha}) = E(\boldsymbol{Y} \mid \boldsymbol{\alpha})$ denotes the conditional expectation of the item response vector $\boldsymbol{Y}$, given attribute vector $\boldsymbol{\alpha}$. Completeness of the Q-matrix is a general requirement for any diagnostic classification regardless of whether MMLE-EM, MCMC, or nonparametric methods are used to assign examinees to proficiency classes.

A plethora of DCMs has been proposed in the literature (e.g., Fu & Li, 2007; Rupp & Templin, 2008). They differ in how the functional relation between mastery of attributes and the probability of a correct item response is modeled. DCMs have been distinguished based on criteria like compensatory versus non-compensatory (can lacking certain attributes be compensated for by possessing other attributes or not), or conjunctive (all attributes specified for an item in $\mathbf{q}$ are required; mastering only a subset of them results in a success probability equal to that of an examinee mastering none of the attributes) versus disjunctive (mastery of a subset of the required attributes is a sufficient condition for maximizing the probability of a correct item response) (de la Torre & Douglas, 2004, Henson et al., 2009; Maris, 1999) (for a different perspective on these criteria, see von Davier, 2014a,b). The **D**eterministic **I**nput **N**oisy "**A**ND" Gate (DINA) Model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) is the standard example of a conjunctive DCM. Its IRF is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{(1-\eta_{ij})}$$

subject to $0 < g_j < 1 - s_j < 1$ for each item j. The conjunction parameter η_{ij} is defined as $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ indicating whether examinee i has mastered all the attributes needed to answer item j correctly. The item-related parameters $s_j = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$ and $g_j = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$ refer to the probabilities of slipping (failing to answer item j correctly despite mastery of the required attributes) and guessing (answering item j correctly despite lacking the attributes required to do so), respectively. The DINA model is conjunctive because an examinee must master all required attributes for maximal probability of answering an item correctly. (Thus, the conjunction parameter η_{ij} can be interpreted as the ideal item response when neither slipping nor guessing occur.) The **D**eterministic **I**nput **N**oisy "**O**R" Gate (DINO) Model (Templin & Henson, 2006) is the prototypical disjunctive DCM. The disjunction parameter, $\omega_{ij} = 1 - \prod_{k=1}^{K}(1 - \alpha_{ik})^{q_{jk}}$, indicates whether at least one of the attributes associated with item j has been mastered. (Like η_{ij} in the DINA model, ω_{ij} corresponds to the ideal item response when neither slipping nor guessing occur.) The IRF of the DINO model is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = (1 - s_j)^{\omega_{ij}} g_j^{(1-\omega_{ij})}$$

subject to $0 < g_j < 1 - s_j < 1$ for each item j.

The DINA model and the DINO model are rather limited in their flexibility to model the relation between response probabilities and attribute mastery. The Reduced **R**eparameterized **U**nified **M**odel (Reduced RUM; Hartz, 2002; Hartz & Roussos, 2008) offers far greater flexibility in modeling the probability of correct item responses for different attribute vectors. Specifically, the DINA model cannot distinguish between examinees who master none and those who master a subset of the attributes required for an item. Only if all required attributes are mastered can an examinee realize a high probability of answering the item correctly. This restriction has been relaxed in case of the Reduced RUM, as it allows for incremental probabilities of a correct response along with an increasing number of required attributes mastered.

General DCMs have been proposed with an even more flexible parameterization such that they can be used as meta-models for expressing the IRFs of specific DCMs in unified mathematical form and parameterization (de la Torre, 2011; Henson et al., 2009; Rupp et al., 2010; von Davier, 2005, 2008, 2014b). von Davier's General Diagnostic Model (GDM; 2005, 2008) is the archetypal general DCM. The IRF of (presumably) the most popular version of the GDM is formed by the logistic function of the linear combination of all K attribute main effects. Henson et al. (2009) proposed to use the linear combination of the K attribute main effects and all their two-way, three-way, ..., K-way interactions

$$v_{ij}=\beta_{j0}+\sum_{k=1}^{K}\beta_{jk}q_{jk}\alpha_{ik}+\sum_{k'=k+1}^{K}\sum_{k=1}^{K-1}\beta_{j(kk')}q_{jk}q_{jk'}\alpha_{ik}\alpha_{ik'}+\cdots+\beta_{j12\ldots K}\prod_{k=1}^{K}q_{jk}\alpha_{ik} \tag{5.1}$$

for constructing the IRF of a general DCM called the Loglinear Cognitive Diagnosis Model (LCDM)

$$P(Y_{ij}=1\mid\boldsymbol{\alpha}_i)=\frac{\exp\left(v_{ij}\right)}{1+\exp\left(v_{ij}\right)} \tag{5.2}$$

(see Equation 11 in Henson et al., 2009). de la Torre (2011) presented the Generalized DINA (G-DINA) model that, in addition to the logit link, allows for constructing the IRF based on the identity link, $P(Y_{ij}=1\mid\boldsymbol{\alpha}_i)=v_{ij}$, and the log link, $P(Y_{ij}=1\mid\boldsymbol{\alpha}_i)=\exp\left(v_{ij}\right)$. By imposing appropriate constraints on the β-coefficients in v_{ij}, the IRFs of specific DCMs can be reparameterized as general DCMs. For example, the IRF of the DINA model becomes

$$P(Y_{ij}=1\mid\boldsymbol{\alpha}_i)=\frac{\exp\left(\beta_{j0}+\beta_{j(\forall k\in\mathcal{L}_j)}\prod_{k\in\mathcal{L}_j}\alpha_{ik}\right)}{1+\exp\left(\beta_{j0}+\beta_{j(\forall k\in\mathcal{L}_j)}\prod_{k\in\mathcal{L}_j}\alpha_{ik}\right)}$$

subject to $\beta_{j(\forall k\in\mathcal{L}_j)}>0$. The set $\mathcal{L}_j=\{k\mid q_{jk}=1\}$ contains the non-zero elements in $\mathbf{q}_j$. (If $k\in\mathcal{L}_j$, then $q_{jk}=1$ is always true; hence, q_{jk} has been dropped from the IRF.)

5.3 Clustering Adapted to Cognitive Diagnosis

As mentioned earlier, clustering algorithms were among the first nonparametric methods that researchers studied for use in CD as an approximation to the computationally expensive parametric methods for assigning examinees to proficiency classes (Ayers et al., 2008, 2009; Chiu, 2008; Chiu et al., 2009; Willse et al., 2007). Efficient implementations of various techniques for clustering a set of objects were readily available in the major statistical software packages. The principal objective shared by all these techniques is the identification of maximally homogeneous groups ("clusters") that are maximally separated. Back then, researchers focused on hierarchical clustering methods (HACA) and K-means clustering.

Independent of the particular algorithm used, as a specific feature of the application of clustering in CD, examinees' raw score item vectors $\boldsymbol{Y}_i$ were aggregated into a K-dimensional vector of attribute sum-scores $\boldsymbol{W}_i$ that served as input to clustering. They are defined as $\boldsymbol{W}_i = (W_{i1}, \ldots, W_{iK})' = \boldsymbol{Y}_i\mathbf{Q}$. Because each cell entry of $\mathbf{Q} = \{q_{jk}\}$ indicates the association between item j and attribute α_k, each element of $\boldsymbol{W}_i$ is the sum of the correct answers of examinee i to all items requiring mastery of the kth attribute. (Items that require mastery of more than one attribute for their solution contribute to multiple elements of $\boldsymbol{W}_i$.) Across examinees, the attribute sum-score vectors $\boldsymbol{W}_i$ form the rows of a rectangular $N \times K$ matrix $\mathbf{W}$.

Among HACA algorithms, of particular interest were complete-link, average-link HACA (Johnson, 1967), and Ward's (1963) method. All three algorithms require as input an $N \times N$ square-symmetric matrix of Euclidean inter-examinee distances computed from examinees' item score vector matrix $\mathbf{W}$. The HACA link algorithms sequentially group ("agglomerate") examinees—or groups of examinees—closest to each other at each step into an inverted tree-shaped hierarchy of nested classes that represents the relationship between examinees. After each agglomeration, the inter-examinee distances are recalculated to reflect the latest status of cluster cohesion as input for the next agglomeration step. The specific method of updating distances distinguishes the link algorithms. Ward's method uses a different strategy that does not rely upon inter-examinee distances but instead attempts to minimize the increase in total within-cluster variance after merging. (As an aside, one should note that the complete-link algorithm amounts to minimizing the within-cluster diameter; for further technical details, consult, for example, Arabie, Hubert, & De Soete 1996; Everitt, Landau, & Leese, 2001; Gordon, 1999; and the classic reference, Hartigan, 1975; or Chapter 14 in Hastie, Tibshirani, & Friedman, 2009)

K-means clustering is presumably the most popular technique for identifying an exhaustive disjoint (i.e., non-hierarchical) grouping of a data set (called a partition) (Bock, 2007; Forgy, 1965; Hartigan & Wong, 1979; MacQueen, 1967; Steinhaus, 1956; Steinley, 2006). The number of clusters, K, to be extracted must be specified in advance. Different from the HACA algorithms, K-means uses the $N \times K$ matrix of attribute sum scores $\mathbf{W}$ directly as input. The grouping process attempts to minimize the loss function of within-cluster heterogeneity (which is equivalent to

maximizing between-cluster heterogeneity). A collection of M mutually exclusive and exhaustive subsets of the entire set of N examinees, $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M$, is sought so that the overall sum of squared within-cluster deviations of examinees from the K-vector of their cluster centroids, $\text{WCSS}(\mathbf{W}) = \sum_{m=1}^{M} \sum_{i \in \mathcal{C}_m} \|\boldsymbol{W}_i - \overline{\boldsymbol{W}}_m\|^2$, is minimized. $\overline{\boldsymbol{W}}_m$ denotes the centroid (mean) vector of cluster $\mathcal{C}_m$ obtained by averaging the observed attribute sum-score vectors, $\boldsymbol{W}_{i \in \mathcal{C}_m}$, where the K elements, $\overline{W}_{mk}$, are defined as $\overline{W}_{mk} = \frac{1}{N_m} \sum_{i \in \mathcal{C}_m} W_{ik}$, with N_m indicating the number of examinees in $\mathcal{C}_m$. The typical K-means clustering algorithm starts by selecting an initial set of examinees as cluster centers ("seeds"). The distances of the remaining examinees to these seeds determine the initial value of the loss function. The algorithm follows an iterative improvement strategy by repeatedly relocating examinees to clusters according to minimum distance; cluster centroids are re-calculated and examinees re-grouped until no further decreases in the loss function can be realized (that is, until each examinee is located closest to the centroid of the cluster to which he or she is assigned). To avoid an only locally-optimal solution, K-means is recommended to be used with a large number of random re-starts (Steinley, 2003).

The clustering methods described in the previous paragraphs as well as related computational procedures and algorithms discussed in Chiu et al. (2009) are implemented as the software package `ACTCD` in `R` (Chiu & Ma, 2016).[1]

5.3.1 *The Asymptotic Classification Theory of Cognitive Diagnosis*

The Asymptotic Classification Theory of Cognitive Diagnosis (ACTCD) provides the theoretical justification for using HACA for assigning examinees to proficiency classes in CD.[2] The original version of the ACTCD, developed by Chiu (2008) in her dissertation, consisted of three lemmas, each of which specified a necessary condition for a consistency theorem of classification to hold. (Chiu et al., 2009, provided a detailed presentation of the ACTCD, as it applies to the DINA model). Lemma 1 stated that the Q-matrix of a test is guaranteed to be complete if each

[1] `R` is an open source statistical computing language available through the Comprehensive `R` Archive Network (CRAN) for free public use.

[2] Recall that "classification" typically refers to supervised learning—that is, the groups are known a priori—and "clustering" to unsupervised learning, where the groups are to be discovered in the analysis. Thus, strictly speaking, neither classification nor clustering seem accurate descriptions of the use of HACA with CD because (a) the number of realizable proficiency classes is known in advance and used to "cut" the HACA tree accordingly so that assigning examinees to clusters might be legitimately addressed as "classification" and (b) HACA produces unlabeled groups (i.e., not identified in terms of the underlying attribute vectors $\boldsymbol{\alpha}$) that require additional steps to determine the underlying $\boldsymbol{\alpha}$ so that "clustering" might also appear as a fairly accurate characterization of the use of HACA in CD.

attribute is represented by at least one single-attribute item—so, $\mathbf{Q}$ has rows, $\mathbf{e}_1, \ldots, \mathbf{e}_K$, among its J rows, where $\mathbf{e}_k$ is a $1 \times K$ vector, with the kth element, e_k, equal 1, and all other entries equal 0. Chiu et al. (2009) proved for the DINA model that Lemma 1 describes a necessary condition for completeness: $\mathbf{Q}$ is complete if and only if it contains a $K \times K$ identity matrix as a submatrix. Lemma 2 described the condition under which the different proficiency classes are well-separated, given $\mathbf{Q}$ is complete. (The center of the proficiency class with attribute vector $\boldsymbol{\alpha}$ is defined as the conditional expectation of the attribute sum-score vector $E(\boldsymbol{W} \mid \boldsymbol{\alpha}) = \boldsymbol{T}(\boldsymbol{\alpha})$, where the kth entry, $T_k(\boldsymbol{\alpha})$, is $E(W_k \mid \boldsymbol{\alpha}) = \sum_{j=1}^{J} E(Y_j \mid \boldsymbol{\alpha}) q_{jk}$. Chiu et al. (2009) proved for the DINA model that $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{T}(\boldsymbol{\alpha}) \neq \boldsymbol{T}(\boldsymbol{\alpha}^*)$ is always true, given $\mathbf{Q}$ is complete. Thus, Lemma 2 provided the theoretical justification to use $\boldsymbol{W}$ as a statistic for $\boldsymbol{\alpha}$ because the centers of different proficiency classes are guaranteed to be distinct if the DINA model holds.) Lemma 3 established that complete-link HACA assigns examinees to their true proficiency classes provided the data conform to a finite mixture model with M latent classes (recall that DCMs are constrained finite mixture models). Building on these three lemmas, the consistency theorem of classification maintained that the probability of HACA assigning examinees correctly to their true proficiency classes using $\boldsymbol{W}$ as input approaches 1 as the length of a test (i.e., the number of items J) increases.

In a series of papers, Chiu and Köhn (2015a,b, 2016) proved the theoretical propositions of the ACTCD for the DINO model, the Reduced RUM, and, finally, for general DCMs. The general version of the ACTCD required modifications of Lemma 1 (Q-completeness) and Lemma 2 (separation of proficiency-class centers) so that the regularity conditions required for the consistency theorem of classification to hold would suit any DCM.

Lemma 1 *Completeness is not an intrinsic property of the Q-matrix, but can only be assessed in relation to a specific DCM supposed to underlie the data—that is, the Q-matrix of a given test can be complete for one model and incomplete for another. An even more complicated situation arises if the test items do not conform to a single DCM, but to a mix of several DCMs. In addition, completeness of the Q-matrix is often difficult to establish, especially, for tests with a large number of items involving many attributes. Recently, Köhn and Chiu (2017) investigated the technical requirements and conditions of Q-completeness, and how to use them to determine whether a given Q-matrix is complete. One of the key results is that any Q-matrix containing a $K \times K$ identity matrix as a submatrix—that is, each attribute is represented by at least one single-attribute item—is guaranteed to be complete for any DCM. For the DINA model and the DINO model, this is a necessary condition. For all other DCMs, inclusion of the K different single-attribute items is a sufficient, but not a necessary condition for Q-completeness—said differently, alternative compositions of the Q-matrix not including the K different single-attribute items also guarantee completeness. Köhn and Chiu (2017) showed that having full rank K is a necessary condition for Q-completeness. In most practical instances, Q-completeness can be assumed if this condition is fulfilled. However, there are certain rare constellations of the values of the item parameters where a*

full-rank-K Q-matrix may not be complete. Specifically, if for $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$, $\boldsymbol{S}(\boldsymbol{\alpha})$ and $\boldsymbol{S}(\boldsymbol{\alpha}^*)$ are not nested—and thus, are not guaranteed to be distinct—then $\mathbf{Q}$ might be incomplete because ambiguous constellations of the item parameters can occur that may prevent a clear distinction of proficiency classes.[3] *Thus, any Q-matrix that is of full rank must be further inspected for completeness. But how?—Recall the definition of completeness: $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{S}(\boldsymbol{\alpha}) \neq \boldsymbol{S}(\boldsymbol{\alpha}^*)$. The right-hand inequality implies the existence of at least one item j such that $\sum_{k=1}^{K} \beta_{jk} q_{jk} (\alpha_k - \alpha_k^*) \neq 0$.*[4] *This inequality is always true for at least one item j if the Q-matrix contains all single-attribute items. However, this inequality is also always true if $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are nested within each other. Thus, if a Q-matrix has rank K, and $\boldsymbol{\alpha} \succ \boldsymbol{\alpha}^*$, then $\boldsymbol{S}(\boldsymbol{\alpha}) \neq \boldsymbol{S}(\boldsymbol{\alpha}^*)$ is always true. (see Proposition 2 in Köhn & Chiu, 2017). Hence, instead of inspecting all $\binom{2^K-2}{2}$ pairs of $\boldsymbol{\alpha}$-vectors—note that pairs involving $\boldsymbol{\alpha}_1 = (00\ldots0)^T$ and $\boldsymbol{\alpha}_M = (11\ldots1)^T$ need not be inspected—only the non-nested pairs of $\boldsymbol{\alpha}$-vectors and their associated $\boldsymbol{S}(\boldsymbol{\alpha})$ and $\boldsymbol{S}(\boldsymbol{\alpha}^*)$ need to be evaluated—that is, for all j, check if all the coefficients in $S_j(\boldsymbol{\alpha}^*)$ also appear in $S_j(\boldsymbol{\alpha})$, or vice versa.*

Lemma 2 *Chiu et al. (2009) proposed the K-dimensional vector of attribute-related sum-scores $\boldsymbol{W} = \boldsymbol{Y}\mathbf{Q}$ as a statistic for $\boldsymbol{\alpha}$. The conditional expectation of $\boldsymbol{W}$, $\boldsymbol{T}(\boldsymbol{\alpha}) = E(\boldsymbol{W}|\boldsymbol{\alpha})$, corresponds to the center of the proficiency class characterized by $\boldsymbol{\alpha}$. If $\mathbf{Q}$ is complete, then, as Chiu et al. (2009) proved, $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{T}(\boldsymbol{\alpha}) \neq \boldsymbol{T}(\boldsymbol{\alpha}^*)$ is true for the DINA model, which implies that the centers of distinct proficiency classes are well-separated (Chiu & Köhn, 2015a, proved the equivalent claim for the DINO model). For DCMs other than the DINA model and the DINO model, however, the sum-score vector $\boldsymbol{W}$ is not a legitimate statistic for $\boldsymbol{\alpha}$ because, as Chiu and Köhn (2016) showed, distinct proficiency classes can have identical conditional expectations of $\boldsymbol{W}$—that is, $\boldsymbol{W}$ cannot guarantee well-separated proficiency-class centers—which invalidates Lemma 2: $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \not\Rightarrow \boldsymbol{T}(\boldsymbol{\alpha}) \neq \boldsymbol{T}(\boldsymbol{\alpha}^*)$.*

The inability of the attribute sum-score vector $\boldsymbol{W}$ to guarantee well-separated proficiency-class centers is resolved by an augmented attribute sum-score statistic $\boldsymbol{W}_{aug}$ that restores the separation guarantee for the centers of distinct proficiency classes (Chiu & Köhn, 2015b, 2016). Specifically, the $J \times K$ Q-matrix is augmented by a matrix of the same dimensionality denoted by $\mathbf{Q_e}$, which is constructed by retaining from the original Q-matrix the item-attribute vectors of all single-attribute items—that is, $\mathbf{q}_j = \mathbf{e}_k$ in $\mathbf{Q}$—whereas all other rows consist of zero vectors, $(0, 0, \ldots, 0)$. Said differently, only the q-entries of the single-attribute items are repeated. The augmented Q-matrix has dimensionality $J \times 2K$ and is written as

$$\mathbf{Q}_{aug} = [\mathbf{Q} \mid \mathbf{Q_e}].$$

[3] A K-dimensional vector $\boldsymbol{\alpha}^* \neq \boldsymbol{\alpha}$ is said to be *nested* within the vector $\boldsymbol{\alpha}$—written as $\boldsymbol{\alpha} \succ \boldsymbol{\alpha}^*$—if $\alpha_k^* \leq \alpha_k$, for all elements k, and $\alpha_k^* < \alpha_k$ for at least one k.

[4] Parameterization and notation refer to a general DCM as defined in Eqs. 5.1 and 5.2.

Here is an example for $J = 7$ items and $K = 3$ attributes:

$$\mathbf{Q} = \begin{pmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \\ 1\,1\,0 \\ 1\,0\,1 \\ 0\,1\,1 \\ 1\,1\,1 \end{pmatrix} \Rightarrow \mathbf{Q}_{aug} = [\mathbf{Q} \mid \mathbf{Q}_{\mathbf{e}}] = \left(\begin{array}{c|c} 1\,0\,0 & 1\,0\,0 \\ 0\,1\,0 & 0\,1\,0 \\ 0\,0\,1 & 0\,0\,1 \\ 1\,1\,0 & 0\,0\,0 \\ 1\,0\,1 & 0\,0\,0 \\ 0\,1\,1 & 0\,0\,0 \\ 1\,1\,1 & 0\,0\,0 \end{array}\right)$$

Thus, $\mathbf{Q}_{aug}$ preserves the information in the original Q-matrix but enhances the effect of the single-attribute items, resulting in well-separated proficiency-class centers. The definitions of the augmented sum-score statistic for $\boldsymbol{\alpha}$, $\boldsymbol{W}_{aug}$, and its expectation are readily derived from $\mathbf{Q}_{aug}$

$$\begin{aligned} \boldsymbol{W}_{aug} &= \boldsymbol{Y}\mathbf{Q}_{aug} = \boldsymbol{Y}[\mathbf{Q} \mid \mathbf{Q}_{\mathbf{e}}] = [\boldsymbol{W} \mid \boldsymbol{W}_{\mathbf{e}}], \\ \boldsymbol{T}_{aug}(\boldsymbol{\alpha}) &= E(\boldsymbol{W}_{aug} \mid \boldsymbol{\alpha}) = E\big([\boldsymbol{W} \mid \boldsymbol{W}_{\mathbf{e}}] \mid \boldsymbol{\alpha}\big) = \big[\boldsymbol{T}(\boldsymbol{\alpha}) \mid \boldsymbol{T}_{\mathbf{e}}(\boldsymbol{\alpha})\big]. \end{aligned}$$

In conclusion, the augmented sum-score statistic $\boldsymbol{W}_{aug}$ guarantees well-separated proficiency-class centers for distinct $\boldsymbol{\alpha}$—that is, $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^* \Rightarrow \boldsymbol{T}_{aug}(\boldsymbol{\alpha}) \neq \boldsymbol{T}_{aug}(\boldsymbol{\alpha}^*)$.

5.3.2 Clustering in Cognitive Diagnosis: Issues and Limitations

5.3.2.1 Only HACA Is Covered by the ACTCD

Generally, it has been shown that nonhierarchical clustering obtains better classification results than HACA. Consider Ward's method as an example: it has been demonstrated repeatedly that K-means clustering can always improve (at least it does not deteriorate) if it is initialized with the clustering solution obtained from Ward's method (e.g., Steinley & Brusco, 2007). Similarly, Chiu et al. (2009), could demonstrate in their simulation studies that K-means clustering outperformed HACA when used for assigning examinees to proficiency classes. So, why does the ACTCD not cover K-means clustering as well?

Unfortunately, no theoretical justification currently exists that would legitimize using K-means clustering as a method for assigning examinees to proficiency classes within the context of diagnostic classification. Recall that Lemma 3 of the ACTCD states that if a finite mixture model with M latent classes underlies the data, then HACA accurately assigns examinees to their true proficiency classes. Chiu et al. (2009) proved Lemma 3 for complete-link HACA. At present, such a proof seems not available for K-means clustering because the asymptotic theory for K-means clustering is an unresolved problem. Specifically, as Chiu et al. (2009)

remarked, the asymptotic theory for K-means has been worked out in some general cases, showing that estimates of the cluster centers converge with certain rates (see, for example, Hartigan, 1978; Pollard, 1981, 1982); however, these centers do not necessarily correspond to the expected values of different sum-score statistics as they define the centers of distinct proficiency classes, which is one of the key ideas of the ACTCD.

5.3.2.2 HACA—and Cluster Analysis in General—Have a Labeling Problem

Examinee clusters obtained from clustering methods serve as proxies for the proficiency classes. But, different from parametric techniques for classifying examinees, clustering methods cannot estimate the attribute vectors underlying the clusters, and so do not provide labels of the groups of examinees in terms of the attribute vectors as they characterize distinct proficiency classes. Hence, the clusters must be interpreted—labeled—that is, their underlying attribute vectors must be reconstructed from the chosen input data, which can be tedious if the number of examinees is large. Thus, as cluster analysis cannot inform on examinees' specific strengths and weaknesses regrading attribute mastery, does using clustering for classifying examinees not take away one of the main advantages of cognitive diagnosis?

Through the Q-matrix, the attribute sum-score vectors, $\boldsymbol{W}$ and $\boldsymbol{W}_{aug}$, are directly related to the attribute vectors of the proficiency classes, which allows for a relatively straightforward rationale for interpreting the clusters obtained from HACA because their underlying attribute vectors, $\boldsymbol{\alpha}$, can be deduced from the cluster members' attribute sum-score vectors, as they are of the same dimensionality as $\boldsymbol{\alpha}$. In fact, for $\boldsymbol{W}$ as input to clustering, Chiu et al. (2009) presented an automatic cluster labeling algorithm that seeks to identify an optimal match between examinees' within-cluster sum-score vectors and candidate attribute vectors potentially underlying this cluster. This algorithm has been further developed so that now also $\boldsymbol{W}_{aug}$ as input to clustering can be accommodated.

5.3.2.3 Alternative Statistics for Estimating $\boldsymbol{\alpha}$

The two attribute sum-score statistics for $\boldsymbol{\alpha}$, $\boldsymbol{W}$ and $\boldsymbol{W}_{aug}$, are theoretically well-supported by the ACTCD: the Consistency Theorem of Classification states that the probability of assigning examinees to their true proficiency classes using HACA with $\boldsymbol{W}$ or $\boldsymbol{W}_{aug}$ as input approaches one as the length of a test (i.e., the number of items) approaches infinity. However, $\boldsymbol{W}$ and $\boldsymbol{W}_{aug}$ require that the true Q-matrix underlying a given test be known. Unfortunately, in practice, the Q-matrix for most tests is unknown and must be estimated to determine the associations between items and attributes, risking a misspecified Q-matrix that may result in the incorrect classification of examinees. Another difficulty, as Köhn, Chiu and Brusco (2015)

demonstrated, is that aggregating the observed item scores of examinees who may belong to different proficiency classes can result in their having identical attribute sum-score vectors and therefore risks misclassification of those examinees.

Hence, Köhn et al. (2015) considered clustering examinees into proficiency classes using their item-score vectors $\boldsymbol{Y}$ rather than their attribute sum-score vectors, as the former do not require knowledge of the Q-matrix. Of course, the crucial question was whether $\boldsymbol{Y}$ is also consistent—does the Consistency Theorem of Classification also hold for $\boldsymbol{Y}$? As a necessary condition for consistency, $\boldsymbol{Y}$ must satisfy Lemma 2 of the ACTCD. In fact, Köhn et al. (2015) proved that, given a complete Q-matrix, $\boldsymbol{Y}$ guarantees well-separated centers of the different proficiency classes; that is, $\boldsymbol{Y}$ is a legitimate statistic for $\boldsymbol{\alpha}$ and covered by Lemma 2. But, unfortunately, so far at least, the Consistency Theorem of Classification cannot be proven for $\boldsymbol{Y}$ because the dimensionality of $\boldsymbol{Y}$ depends on J: if J goes to infinity, then $\boldsymbol{Y}$ contradicts the fundamental assumption of any classification algorithm that its input be finite. This difficulty is elegantly avoided by the attribute sum-score vectors because their dimensionality depends on K and not on J. Finally, one might raise the question whether assigning examinees to proficiency classes based on $\boldsymbol{Y}$, without having to know the Q-matrix of the test in question, does not essentially mean the elimination of the theoretical connection between individual items and the attribute vectors, $\boldsymbol{\alpha}$, that define the different proficiency classes. More to the point, does using $\boldsymbol{Y}$ as input to clustering not abandon the theoretical framework of cognitive diagnosis?

5.4 Nonparametric Classification of Examinees

5.4.1 The Nonparametric Classification Method

The Nonparametric Classification (NPC) method developed by Chiu and Douglas (2013) does not rely on parametric estimation of examinees' proficiency class membership, but uses a distance-based algorithm on the observed item responses for classifying examinees. Different from the use of clustering in CD, the NPC method is a genuine classification method because the $2^K = M$ proficiency classes to which to assign examinees are known a priori.

Proficiency class membership is determined by comparing an examinee's observed item response vector $\boldsymbol{Y}$ with each of the ideal item response vectors of the realizable $2^K = M$ proficiency classes. The ideal item responses are a function of the Q-matrix and the attribute vectors characteristic of the different proficiency classes. Hence, an examinee's proficiency class is identified by the attribute vector $\boldsymbol{\alpha}_m$ underlying that ideal item response vector which is closest—or most similar—to an examinee's observed item response vector. The ideal response to item j is the score that would be obtained by an examinee if no perturbation (e.g., slipping or guessing) occurred.

Let $\boldsymbol{\eta}_i$ denote the J-dimensional ideal item response vector of examinee i. (Recall that $\boldsymbol{\eta}_i$ is a function of the Q-matrix of a test and $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i \in \mathcal{C}_m} = \boldsymbol{\alpha}_m$, an examinee's attribute vector, as it is determined by his or her proficiency class because all examinees in proficiency class $\mathcal{C}_m$ share the same attribute vector; hence, $\boldsymbol{\eta}_i = \boldsymbol{\eta}_{i \in \mathcal{C}_m} = \boldsymbol{\eta}_m$.) As the Q-matrix and the M realizable proficiency classes are known, the construction of all possible ideal item response vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_M$ is straightforward.

Formally, the NPC estimator $\widehat{\boldsymbol{\alpha}}$ of an examinee's attribute vector is defined as the attribute vector underlying the ideal item response vector that among all ideal item response vectors minimizes the distance to an examinee's observed item response vector:

$$\widehat{\boldsymbol{\alpha}}_i = \underset{m \in \{1,2,\ldots,M\}}{\arg\min} \; d(\boldsymbol{y}_i, \boldsymbol{\eta}_m) \tag{5.3}$$

Hence, the choice of the specific distance measure $d(\cdot)$ for the loss function of Eq. 5.3 is of critical importance in determining $\widehat{\boldsymbol{\alpha}}_i$.

A distance measure often used for clustering binary data is the Hamming distance that simply counts the number of disagreements between two vectors:

$$d_H(\boldsymbol{y}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \mid y_j - \eta_j \mid$$

If the different levels of variability in the item responses are to be incorporated, then the Hamming distances can be weighted, for example, by the inverse of the item sample variance, which allows for larger impact on the distance functions of items with smaller variance:

$$d_{wH}(\boldsymbol{y}, \boldsymbol{\eta}) = \sum_{j=1}^{J} \frac{1}{\bar{p}_j(1 - \bar{p}_j)} \mid y_j - \eta_j \mid$$

($\bar{p}_j$ is the proportion of correct responses to the jth item). A purported advantage of the weighted Hamming distance is the substantial reduction in the number of ties, which can be an issue especially with short tests. (As a second variety of a weighted Hamming distance, Chiu and Douglas (2013) discuss a differential item weighting scheme that incorporates slipping and guessing.)

Simulation studies conducted by Chiu and Douglas (2013) showed that the NPC method (a) can be implemented in a computationally inexpensive and effective way; (b) can be used—different from parametric methods—essentially with any sample size, and is therefore particularly suited for small-scale testing programs; (c) is robust to Q-matrix misspecifications; and (d) can be used with observed item responses that conform to any DCM that uses the concept of an ideal item response to link the vector of required attributes $\mathbf{q}$ of an item with the attributes mastered by an examinee. The **N**oisy **I**nput **D**eterministic "**A**ND" Gate (NIDA) model (Maris,

1999) and the DINA model use conjunctive ideal item response vectors $\boldsymbol{\eta}_i^{(c)}$, with elements $\eta_{ij}^{(c)} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$ defined already in connection with the DINA model as the conjunction parameter of item j. The DINO model uses disjunctive ideal item response vectors $\boldsymbol{\eta}_i^{(d)}$, with elements $\eta_{ij}^{(d)} = 1 - \prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}}$ defined already in connection with the DINO model as the disjunction parameter ω_{ij} of item j.

Wang and Douglas (2015) proved that under certain regularity conditions $\widehat{\boldsymbol{\alpha}}$ obtained by the NPC method is a statistically consistent estimator of an examinee's attribute vector for any DCM:

> "... the only general condition required of the underlying item response function is that the probability of a correct response for masters of the attributes is bounded above 0.5 for each item, and the probability for non-masters is bounded below 0.5. If the true model satisfies these simple conditions, nonparametric classification will be consistent as the test length increases" (Wang & Douglas, 2015, p. 99).

For example, for conjunctive DCMs the conditions $P(Y_{ij} = 1 \mid \eta_{ij}^{(c)} = 0) < 0.5$ and $P(Y_{ij} = 1 \mid \eta_{ij}^{(c)} = 1) > 0.5$ guarantee consistency of $\widehat{\boldsymbol{\alpha}}$.

In summary, the NPC method allows for the computationally inexpensive and robust classification of examinees. Perhaps its most attractive feature is the ability to handle small and very small samples (say, of classroom size), without the requirement to specify a DCM supposedly underlying the data. The key idea of the NPC method—estimating examinees' proficiency class in comparing observed and ideal item responses—is integral also to the nonparametric Q-matrix refinement method and the joint maximum likelihood estimation of DCM item parameters presented below. An implementation of the NPC method is available in the `R` package `NPCD` (Zheng & Chiu, 2016).

5.4.2 The General Nonparametric Classification Method

The consistency conditions of the NPC-estimator $\widehat{\boldsymbol{\alpha}}$ identified by Wang and Douglas (2015) are often difficult to meet for more complex DCMs like the Reduced RUM and general DCMs because the probability of a correct response to an item increases as a function of the number of required attributes that are mastered by an examinee (known as the "monotonicity assumption"). Hence, $\boldsymbol{\eta}^{(c)}$ and $\boldsymbol{\eta}^{(d)}$ might not offer the necessary flexibility to model the relation between required and mastered attributes for these advanced DCMs. As an illustration, consider a domain characterized by two attributes; the realizable proficiency classes are $\boldsymbol{\alpha}_1 = (00)$, $\boldsymbol{\alpha}_2 = (10)$, $\boldsymbol{\alpha}_3 = (01)$, and $\boldsymbol{\alpha}_4 = (11)$. Given an item having attribute vector $\boldsymbol{q} = (11)$, the corresponding conjunctive ideal item responses are $\eta_1^{(c)} = 0$, $\eta_2^{(c)} = 0$, $\eta_3^{(c)} = 0$, and $\eta_4^{(c)} = 1$. Assume this item conforms to the DINA model, with $g = 0.1$ and $1 - s = 0.9$. (The equivalent parameterization using the G-DINA model is $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_{12})' = (0.1, 0, 0, 0.8)'$.) The probabilities of answering the item correctly for the four proficiency classes are 0.1, 0.1, 0.1, and 0.9, respectively.

Thus, the conjunctive ideal responses 0, 0, 0, and 1 are, indeed, the most likely responses of the four proficiency classes. However, this may no longer be true if the data conform to a more complex model, say, the saturated G-DINA model with parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_{12})' = (0.1, 0.4, 0.6, -0.2)'$. Then, for the four proficiency classes, the probabilities of a correct item response are 0.1, 0.5, 0.7, and 0.9. Thus, the conjunctive ideal item responses 0, 0, 0, and 1 are no longer the most likely responses, and using $\boldsymbol{\eta}^{(c)}$ in that instance may result in a substantial number of examinee misclassifications.

The two prototypic conjunctive and disjunctive DCMs, the DINA model and the DINO model, define the two extremes of a continuum describing the relation between $\mathbf{q}$ and $\boldsymbol{\alpha}$ (in fact, the two DCMs have a "dual" relation such that—loosely speaking—the one can be seen as the inverse of the other; Köhn & Chiu, 2016). Based on this observation, Chiu et al. (2018) proposed a weighted ideal response $\boldsymbol{\eta}^{(w)}$, defined as the convex combination of $\boldsymbol{\eta}^{(c)}$ and $\boldsymbol{\eta}^{(d)}$, that allowed to overcome the limitations of conjunctive and disjunctive ideal item responses for the development of the General Nonparametric Classification (GNPC) method.

Suppose item j requires $K_j^* \leq K$ attributes that, without loss of generality, have been permuted to the first K_j^* positions of the item attribute vector $\mathbf{q}_j$. Thus, the original K-dimensional vector $\mathbf{q}_j$ can be reduced to the K_j^*-dimensional item attribute vector $\mathbf{q}_j^*$ because the remaining $K - K_j^*$ entries beyond the first K_j^* positions of $\mathbf{q}_j$ are irrelevant for distinguishing among proficiency classes as these attributes are not required for item j. Said differently, item j requiring $K_j^* \leq K$ attributes allows to distinguish only between $2^{K_j^*}$ latent classes because the attributes beyond the first K_j^* positions of $\boldsymbol{\alpha}$ and $\boldsymbol{q}_j$ are uninformative for distinguishing among proficiency classes. Define the set $l = \{m \mid \boldsymbol{\alpha}_m = (\boldsymbol{\alpha}_l^*, \cdot)\}$, where $(\cdot)$ denotes the irrelevant entries in the original examinee attribute vector beyond position K_j^* (with values 0 or 1). Assume that a group of examinees shares the attributes in the first K_j^* positions of their attribute vectors. Consequently, they are all classified as $\mathcal{C}_l$ although the remaining $K - K_j^*$ entries in their attribute vectors might identify these examinees as belonging to different proficiency classes. But because in case of item j these additional $K - K_j^*$ attributes are disregarded they cannot contribute to distinguish further among examinees. In other words, the possibly distinct proficiency classes of these examinees have been "collapsed" into $\mathcal{C}_l = \bigcup_{m \in l} \mathcal{C}_m$.

For each item j and $\mathcal{C}_l$, the weighted ideal response $\eta_{lj}^{(w)}$ is defined as the convex combination

$$\eta_{lj}^{(w)} = w_{lj}\eta_{lj}^{(c)} + (1 - w_{lj})\eta_{lj}^{(d)} \tag{5.4}$$

where $0 \leq w_{lj} \leq 1$. (As an aside, Eq. 5.4 identifies $\eta^{(c)}$ and $\eta^{(d)}$ as special cases of $\eta^{(w)}$; for example, if the underlying model is indeed conjunctive, then $w_{lj} = 1$ and $\eta_{lj}^{(w)} = \eta_{lj}^{(c)}$.) As an important feature of the GNPC method, the weights w_{lj} are estimated from the data and do not require a priori knowledge of the DCM underlying the data that—as one might suspect—would be needed to specify the

relative contributions of $\eta_{lj}^{(c)}$ and $\eta_{lj}^{(d)}$ to $\eta_{lj}^{(w)}$; thus, w_{lj} is "automatically" adjusted to the level of complexity and variability of the data. The distance between the observed responses to item j and the weighted ideal responses $\eta_{lj}^{(w)}$ of examinees in $\mathcal{C}_l$ is defined as the sum of squared deviations:

$$d_{lj} = \sum_{i \in \mathcal{C}_l} (y_{ij} - \eta_{lj}^{(w)})^2 = \sum_{i \in \mathcal{C}_l} \left(y_{ij} - w_{lj}\eta_{lj}^{(c)} - (1 - w_{lj})\eta_{lj}^{(d)}\right)^2$$

Thus, $\widehat{w}_{lj}$ can be estimated by minimizing d_{lj}:

$$\widehat{w}_{lj} = \frac{\sum_{i \in \mathcal{C}_l} (y_{ij} - \eta_{lj}^{(d)})}{\parallel \mathcal{C}_l \parallel (\eta_{lj}^{(c)} - \eta_{lj}^{(d)})} \tag{5.5}$$

where $\parallel \mathcal{C}_l \parallel$ indicates the number of examinees in the "collapsed" proficiency class $\mathcal{C}_l$. Equation 5.5 implies that an initial classification of examinees is required as input to the estimation of w_{lj}. The NPC method is used to obtain this initial classification.

After $\widehat{w}_{lj}$ has been determined, $\widehat{w}_{mj}$ can be derived immediately because $m \in l$ and therefore, $\widehat{w}_{mj} = \widehat{w}_{lj}$ for all m. As an illustration, consider $K = 3$ attributes and item j having attribute vector $\boldsymbol{q}_j = (110)$. Because only the first two attributes are required for item j, just $2^{K_j^*} = 2^2 = 4$ proficiency classes can be identified, but not all of the $M = 2^3 = 8$ realizable proficiency classes. For example, $\mathcal{C}_2$ and $\mathcal{C}_6$ having attribute vectors $\boldsymbol{\alpha}_2 = (100)$ and $\boldsymbol{\alpha}_6 = (101)$, respectively, cannot be separated based on item j. Thus, $m = 2, 6$ and $l = \{2, 6\}$, and proficiency classes $\mathcal{C}_2$ and $\mathcal{C}_6$ are "collapsed" into $\mathcal{C}_l = \mathcal{C}_{\{2,6\}} = \mathcal{C}_2 \bigcup \mathcal{C}_6$ implying that $w_{2j} = w_{6j} = w_{\{2,6\}j}$. Then, $\widehat{w}_{\{2,6\}j}$ is estimated from the observed responses of examinees in $\mathcal{C}_l$ according to Eq. 5.5 resulting in the estimates $\widehat{w}_{2j}$ and $\widehat{w}_{6j}$: suppose $\widehat{w}_{\{2,6\}j} = 0.6$, then $\widehat{w}_{2j} = \widehat{w}_{6j} = \widehat{w}_{\{2,6\}j} = 0.6$. Subsequently, $\hat{\eta}_{mj}^{(w)}$ is computed from $\widehat{w}_{mj}$ based on Eq. 5.4. In this manner, all possible weighted ideal response patterns $\widehat{\boldsymbol{\eta}}_1^{(w)}, \widehat{\boldsymbol{\eta}}_2^{(w)}, \ldots, \widehat{\boldsymbol{\eta}}_M^{(w)}$ can be constructed for the M realizable $\boldsymbol{\alpha}_m$.

The distance between the observed item response vector and a particular weighted ideal item response vector $\boldsymbol{\eta}_m^{(w)}$ is defined as

$$d(\boldsymbol{y}_i, \widehat{\boldsymbol{\eta}}_m^{(w)}) = \sum_{j=1}^{J} d(y_{ij}, \hat{\eta}_{mj}^{(w)}) = \sum_{j=1}^{J} (y_{ij} - \hat{\eta}_{mj}^{(w)})^2$$

The GNPC estimator $\widehat{\boldsymbol{\alpha}}$ of an examinee's attribute vector is defined as the attribute vector underlying the weighted ideal item response vector that among all weighted ideal item response vectors minimizes the loss function defined in terms of the distance to an examinee's observed item response vector:

$$\widehat{\boldsymbol{\alpha}}_i = \underset{m \in \{1,2,\ldots,M\}}{\arg\min} \; d(\boldsymbol{y}_i, \widehat{\boldsymbol{\eta}}_m^{(w)})$$

A few concluding remarks seem in order. First, the GNPC method allows for estimating examinees' proficiency class when the DCMs underlying the data use a more complex approach than the DINA model and the DINO model to modeling the functional relation between mastery of attributes and the probability of a correct item response. Second, the algorithm of the GNPC method is easy to implement and computationally inexpensive. Its ability to handle even smallest sample sizes resolves the difficulties arising from unstable and unreliable estimates that parametric methods typically encounter in such situations. These features qualify the GNPC method as an analysis tool for classrooms and other small-scale educational programs, where formative assessments devised within the CD framework are needed most. Due to its fast algorithm, the GNPC method might also be an option for constructing CD-based computerized adaptive tests (CAT) to be used in classrooms. Third, one should recall that the GNPC method relies on initial estimates of $\boldsymbol{\alpha}$ (for calculating the estimated proficiency class size needed in the denominator of Eq. 5.5) that, by default, are obtained by using the NPC method with $\boldsymbol{\eta}^{(c)}$. But depending on the true model underlying the data, this choice might not provide the best estimates of $\boldsymbol{\alpha}$ for initializing the GNPC method. As a viable alternative to $\boldsymbol{\eta}^{(c)}$ for obtaining initial estimates of the proficiency classes, Chiu et al. (2018) suggested to use an ideal response with fixed weights defined as

$$\eta_{lj}^{(fw)} = \frac{\sum_{k=1}^{K} \alpha_k q_{jk}}{K} \eta_{lj}^{(c)} + \left(1 - \frac{\sum_{k=1}^{K} \alpha_k q_{jk}}{K}\right) \eta_{lj}^{(d)}. \tag{5.6}$$

In contrast to the freely estimated weight in Eq. 5.4, the weight $\frac{\sum_{k=1}^{K} \alpha_k q_{jk}}{K}$ in Eq. 5.6 is fixed for item j and proficiency class m regardless of the underlying model.

5.5 Methods in Cognitive Diagnosis That Rely on Nonparametric Classification

5.5.1 *Joint Maximum Likelihood Estimation for Cognitive Diagnosis*

The assumption of local independence of the observed item responses allows writing the joint likelihood function as

$$L(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N, \boldsymbol{\Theta}; \mathbf{Y}) = \prod_{i=1}^{N} L_i(\boldsymbol{\alpha}_i, \boldsymbol{\Theta}; \mathbf{y}_i) = \prod_{i=1}^{N} \prod_{j=1}^{J} f(y_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\alpha}_i) \tag{5.7}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_J)$ is the matrix of item parameters, and the matrix $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)'$ consists of N rows corresponding to examinees' J-dimensional observed item response vectors. Despite the mathematical convenience of simple

likelihood functions, the joint estimation of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N$ and $\boldsymbol{\Theta}$ through iteratively maximizing Eq. 5.7—called "joint maximum likelihood estimation" (JMLE)—has been mostly avoided in psychometrics because the JMLE parameter estimators typically lack statistical consistency (Baker & Kim, 2004; Haberman, 2004; Neyman & Scott, 1948).

Chiu, Köhn, Zheng, and Henson (2016) developed a JMLE procedure for CD that resolved the consistency issue by substituting the examinee attribute vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N$ in the joint likelihood function by an external, statistically consistent estimator of examinees' proficiency classes denoted by $\widehat{\boldsymbol{\alpha}}$. Thus, the joint likelihood of Eq. 5.7 is reduced to a function of only a single set of unknowns, the item parameters, $L(\boldsymbol{\Theta}; \mathbf{Y}, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2, \ldots, \widehat{\boldsymbol{\alpha}}_N)$, which then allows for the construction of item parameter estimators that are also consistent, as Chiu et al. (2016) proved.

The JMLE algorithm proposed by Chiu et al. (2016) is an adaptation of Birnbaum's paradigm (Birnbaum, 1968), a two-stage procedure for JMLE (Baker & Kim, 2004; Embretson & Reise, 2000). Examinees' attribute vectors and the item parameters are treated as two sets: the former is assumed to be known, whereas the parameters in the second set are to be estimated. The algorithm is initialized with the estimates of examinees' attribute vectors $\widehat{\boldsymbol{\alpha}}$ as input, which are obtained from one of the nonparametric classification methods, NPC or GNPC, described in the previous section. (Hence, $\widehat{\boldsymbol{\alpha}}$ does not depend on the JMLE procedure—$\widehat{\boldsymbol{\alpha}}$ is an external estimator.) The estimators of the item parameters can then be derived immediately by maximizing the item log-likelihood

$$\ln L_j(\boldsymbol{\theta}_j; \mathbf{y}_j, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2, \ldots, \widehat{\boldsymbol{\alpha}}_N) = \sum_{i=1}^{N} \ln\big(f(y_{ij} \mid \boldsymbol{\theta}_j, \widehat{\boldsymbol{\alpha}}_i)\big). \tag{5.8}$$

As an example, consider the LCDM having IRF

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp(v_{ij})}{1 + \exp(v_{ij})}.$$

(See Eq. 5.2; recall that v_{ij} was defined in Eq. 5.1 as the linear combination of all attribute main effects, two-way effects, …, K-way effects.) Suppose all proficiency class attribute vectors $\boldsymbol{\alpha}$ have been estimated using the GNPC method. Then, the item likelihood is

$$\begin{aligned} L_j(\boldsymbol{\beta}_j; \mathbf{y}_j, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2, \ldots, \widehat{\boldsymbol{\alpha}}_N) &= \prod_{i=1}^{N} f(y_{ij} \mid \boldsymbol{\beta}_j, \widehat{\boldsymbol{\alpha}}_i) \\ &= \prod_{i=1}^{N} \Big(\frac{\exp(v_{ij})}{1 + \exp(v_{ij})}\Big)^{y_{ij}} \Big(\frac{1}{1 + \exp(v_{ij})}\Big)^{1-y_{ij}}. \end{aligned}$$

Estimators of the elements of the item parameter vector, $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \beta_{j2}, \ldots, \beta_{j12\ldots K})'$, can be derived by maximizing the item log-likelihood $\ln L_j(\boldsymbol{\beta}_j; \mathbf{y}_j, \widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_2, \ldots, \widehat{\boldsymbol{\alpha}}_N)$.

The item parameter estimators have closed-form expressions that are functions of the means of the M proficiency classes (identifiable by their specific attribute vectors $\boldsymbol{\alpha}$). Define the proficiency class $\widehat{\mathcal{C}}(\mathcal{A}) = \{i \mid \hat{\alpha}_{ik} = 1, \forall k \in \mathcal{A} \text{ and } \hat{\alpha}_{ik'} = 0, \forall k' \in \mathcal{A}^c\}$. Chiu et al. (2016) derived the closed-form expressions of the estimators of the item parameters as $\hat{\beta}_{j0}$, $\hat{\beta}_{jk}$, and $\hat{\beta}_{jkk'}$,

$$\hat{\beta}_{j0} = \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\emptyset)}{1 - \bar{y}_j \hat{\mathcal{C}}(\emptyset)}\right)$$

$$\begin{aligned}\hat{\beta}_{jk} &= \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\{k\})}{1 - \bar{y}_j \hat{\mathcal{C}}(\{k\})}\right) - \hat{\beta}_{j0} \\ &= \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\{k\})}{1 - \bar{y}_j \hat{\mathcal{C}}(\{k\})}\right) - \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\emptyset)}{1 - \bar{y}_j \hat{\mathcal{C}}(\emptyset)}\right)\end{aligned}$$

$$\begin{aligned}\hat{\beta}_{jkk'} &= \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\{k,k'\})}{1-\bar{y}_j \hat{\mathcal{C}}(\{k,k'\})}\right) - \hat{\beta}_{jk} - \hat{\beta}_{jk'} - \hat{\beta}_{j0} \\ &= \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\{k,k'\})}{1-\bar{y}_j \hat{\mathcal{C}}(\{k,k'\})}\right) - \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\{k\})}{1-\bar{y}_j \hat{\mathcal{C}}(\{k\})}\right) - \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\{k'\})}{1-\bar{y}_j \hat{\mathcal{C}}(\{k'\})}\right) + \ln\left(\frac{\bar{y}_j \hat{\mathcal{C}}(\emptyset)}{1-\bar{y}_j \hat{\mathcal{C}}(\emptyset)}\right)\end{aligned}$$

The expressions of the estimators of the remaining parameters can be readily deduced from the pattern emerging from the equations of $\hat{\beta}_{j0}$, $\hat{\beta}_{jk}$, and $\hat{\beta}_{jkk'}$.

Simulation studies conducted by Chiu et al. (2016) for evaluating the performance of their JMLE algorithm showed that the accuracy of the JMLE-based examinee classification and item parameter estimates was comparable to those obtained from MMLE using the EM algorithm. (As an aside, the numerical accuracy of the estimates can be further increased by iterating the algorithm; Theorem 4.2 in Junker (1991) suggests that the consistency property of the parameter estimators is preserved while iterating.)

5.5.2 *Q-Matrix Reconstruction and Refinement*

The development of methods for the identification and validation of the Q-matrix underlying a test is one of the long-standing topics in CD that has always inspired researchers. Examples are Barnes (2010), Chen (2017), Chen, Culpepper, Chen, and Douglas (2018), de la Torre (2008), de la Torre and Chiu (2016), DeCarlo (2012), and Liu, Xu, and Ying (2012, 2013). Chiu (2013) proposed the Q-Matrix Refinement

(QMR) method for identifying and correcting misspecified entries in the Q-matrix of a given test when the underlying DCM is conjunctive (i.e., involves η_{ij}) like the DINA model and the NIDA model.[5]

The QMR method relies on the NPC method for estimating examinees' proficiency classes. Consider the attribute vector of item j, corresponding to the jth row of the Q-matrix to be evaluated; $\mathbf{q}_j$ is linked with the estimated attribute vector $\widehat{\boldsymbol{\alpha}}_i$ of examinee i to generate the ideal item response η_{ij}, the score of examinee i on item j if no perturbation had occurred. The squared difference between the observed item response Y_{ij} and the ideal item response η_{ij} is defined as the residual sum of squares (RSS) of examinee i for item j:

$$\text{RSS}_{ij} = (Y_{ij} - \eta_{ij})^2.$$

The RSS for item j across all N examinees in the M proficiency classes is

$$\text{RSS}_j = \sum_{i=1}^{N}(Y_{ij} - \eta_{ij})^2 = \sum_{m=1}^{M}\sum_{i \in C_m}(Y_{ij} - \eta_{mj})^2. \tag{5.9}$$

(Note that the index of the ideal response to item j has been changed from η_{ij} to η_{mj} because ideal item responses are class-specific: they depend on examinees' attribute vectors as they are determined by proficiency class membership such that all examinees in proficiency class $\mathcal{C}_m$ share the same attribute vector: $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{i \in \mathcal{C}_m} = \boldsymbol{\alpha}_m$.) For each item j, there are $2^K - 1$ admissible candidate item attribute vectors $\mathbf{q}_j$, and for each of them, the loss function defined in Eq. 5.9 can be computed. The value of RSS_j is expected to be smallest for the correct q-vector among the $2^K - 1$ candidates. As the item specific loss functions of RSS_j are independent of each other, the total RSS of the entire test is minimized if each of the individual RSS_j is minimized. Chiu (2013) established the validity of the rationale of the QMR method in demonstrating that if examinees have been correctly classified, then for a given item j, RSS_j of the correct q-vector is always less than RSS_j of any misspecified q-vector.

The algorithm executing the QMR method first obtains an initial estimate of examinees proficiency class using the NPC method. Based on the Q-matrix to be evaluated and the $\widehat{\boldsymbol{\alpha}}$, the ideal item responses η_{mj} and the associated RSS_j for each item are computed. The item with the largest RSS_j is identified as most likely having a misspecified q-vector. For the remaining $2^K - 2$ candidate q-vectors of this item, the RSS_j are then computed, and the jth row of the Q-matrix under inspection is replaced by the q-vector resulting in the smallest RSS_j value. In this manner, the remaining $J - 1$ items are evaluated, and, eventually, have their q-

[5]Notation: As the QMR method relies on the NPC method that can be used for conjunctive as well as disjunctive models, instead of $\eta_{ij}^{(c)}$ and $\eta_{ij}^{(d)}$, in this section only η_{ij} is used to denotes the conjunctive as well as the disjunctive case.

vectors replaced. After the first cycle through the entire set of J items is completed, examinees' $\widehat{\boldsymbol{\alpha}}$ are re-estimated based on the updated Q-matrix using again the NPC method. For the second cycle, the ideal item responses are updated, followed by the evaluation of all candidate q-vectors of the J items, and the possible replacement of q-vectors resulting in large RSS_j. After the completion of the second cycle, examinees' $\widehat{\boldsymbol{\alpha}}$ are again re-estimated based on the updated Q-matrix, and so on. The algorithm continues until the stopping criterion—the RSS_j of each item does no longer change—has been met.

In summary, the QMR method does not require large samples, which, together with its computational efficiency, recommend this approach to Q-matrix reconstruction and refinement especially for use in small and medium-sized educational testing programs. However, the QMR method is not without limitations; for example, a proficiency class may have too few examinees so that identifying a misspecified item q-vector by minimizing RSS_j might be difficult. As a solution to this problem, Chiu (2013) developed a re-scaled loss function that is independent of class size, and that was shown to improve the detection rate for misspecified q-vector entries if proficiency classes were sparse. As a downside, the re-scaled loss function consumed considerably more CPU time.

5.6 Conclusion and Final Remarks

While specialized software offering efficient implementations of parametric, parametric methods for fitting DCMs to (educational) assessment data—for example, the `R` packages `CDM` (Robitzsch et al., 2016) and `GDINA` (Ma & de la Torre, 2017) (for further software options for fitting DCMs, consult "Part IV" in this book), however, where parametric methods may fail or be difficult to implement. Recall that algorithms like MMLE-EM work best for large-scale assessments, where the data of at least several hundred examinees are available. If assessment data collected in educational micro-environments are to be analyzed, then sample sizes may be simply too small for MLE to provide reliable estimates of examinees' proficiency classes (see the small-sample simulations reported in Chiu et al., 2018). (Within an applied context, the focus is typically on the evaluation of instruction and the assessment of students' learning; hence, estimation of the item parameters is not necessarily a primary goal.) In such settings, nonparametric methods—especially, the NPC and GNPC methods—may be the only viable tools for monitoring and assisting "the teaching and learning process while it is occurring" (Stout, 2002, p. 506)—that is, at the classroom level, where CD-based methods are most useful and needed. Similar considerations apply to the use of CD-based computerized adaptive testing (CD-CAT) in small-scale educational settings, where it would be most beneficial. But due to the lack of an efficient and effective computational engine for the reliable assessment of examinees' proficiency classes, CD-CAT is currently not available for use in classrooms.

Within a wider context, one should recall that the conditions and demands of American higher education have changed dramatically over the past years, including the pressure to increase degree-completion rates and calls for greater instructional accountability in general, mandating educational institutions to change, especially in the realm of instruction (Brown & Diaz, 2011; Picciano, 2012). Initiatives in education for reform and change have required that tests for monitoring instruction not only assess overall educational progress, but provide specific diagnostic information on students knowledge and processing skills that are instrumental for problem solving in a curricular domain. The development and use of computer-/web-based facilities in teaching and learning and a shift in the focus of educational assessment to monitoring small-sample settings like individual classrooms are perhaps among the effective responses from many higher education institutions to cope with the challenges posed by calls for greater instructional accountability. Nonparametric methods, as they have been reviewed in this article, will play an important role in the development of assessment systems tailored to support a better understanding of human learning and provide guidance in improving instruction in the classroom. Within this context, a promising avenue for future research is the application of nonparametric methods to the analysis of the individual learning trajectories of students, as they have been recently studied within a CD-based framework (see, for example, Chen, Culpepper, Wang, & Douglas, 2018; Wang, Yang, Culpepper, & Douglas, 2018).

References

Arabie, P., Hubert, L. J., & De Soete, G. (Eds.). (1996). *Clustering and classification*. River Edge, NJ: World Scientific.

Ayers, E., Nugent, R., & Dean, N. (2009). A comparison of student skill knowledge estimates. In T. Barnes, M. Desmarais, C. Romero, & S. Ventura (Eds.), *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings.* Cordoba, Spain (pp. 101–110).

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Barnes, T. (2010). Novel derivation and application of skill matrices: The q-matrix method. In C. Ramero, S. Vemtora, M. Pechemizkiy, & R. S. J. de Baker (Eds.), *Handbook of educational data mining* (pp. 159–172). Boca Raton, FL: Chapman & Hall.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Load & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, H. H. (2007). Clustering methods: A history of K-means algorithms. In P. Brito, P. Bertrand, G. Cucumel, & F. De Carvalho (Eds.), *Selected contributions in data analysis and classification* (pp. 161–172). Berlin, Germany: Springer.

Brown, M. B., & Diaz, B. (2011). Seeking evidence of impact: Opportunities and needs. *EDUCAUSE Review, 46*, 41–54.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement, 41*, 277–293.
Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of DINA Q matrix. *Psychometrika, 83*, 89–108.
Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement, 42*, 5–23.
Chiu, C.-Y. (2008). *Cluster analysis for cognitive diagnosis: Theory and applications* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3337778).
Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis *Applied Psychological Measurement, 37*, 598–618.
Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response profiles. *Journal of Classification, 30*, 225–250.
Chiu, C.-Y., & Köhn, H.-F. (2015a). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement, 39*, 465–479.
Chiu, C.-Y., & Köhn, H.-F. (2015b). A general proof of consistency of heuristic classification for cognitive diagnosis models. *British Journal of Mathematical and Statistical Psychology, 68*, 387–409.
Chiu, C.-Y., & Köhn, H.-F. (2016). Consistency of cluster analysis for cognitive diagnosis: The reduced reparameterized unified model and the general diagnostic model. *Psychometrika, 81*, 585–610.
Chiu, C.-Y., & Ma, W. (2016). *ACTCD: Asymptotic classification theory for cognitive diagnosis. R package version 1.1-0*. Retrieved from the Comprehensive R Archive Network [CRAN] website http://cran.r-project.org/web/packages/ACTCD/
Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.
Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*, 355–375.
Chiu, C.-Y., Köhn, H.-F., Zheng, Y., & Henson, R. (2016). Joint maximum likelihood estimation for cognitive diagnostic models. *Psychometrika, 81*, 1069–1092.
de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.
de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362.
de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.
de la Torre, J., & Chiu, C.-Y. (2016) A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–73.
DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447–468.
DiBello, L. V., Roussos, L. A., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 979–1030). Amsterdam: Elsevier.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). New York: Arnold.
Forgy, E. W. (1965). Cluster analyses of multivariate data: Efficiency versus interpretability of classifications. *Biometrika, 61*, 621–626.
Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.

Fu, J., & Li, Y. (2007). *An integrative review of cognitively diagnostic psychometric models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Grim, J. (2006). EM cluster analysis for categorical data. In D.-Y. Yeung, J. T. Kwok, A. L. N. Fred, F. Roll, & D. de Ridder (Eds.), *Structural, syntactic, and statistical pattern recognition* (pp. 640–648). Berlin, Germany: Springer.

Haberman, S. J. (2004, May/2005, September). *Joint and conditional maximum likelihood estimation for the Rasch model for binary responses* (Research report No. RR-04-20). Princeton, NJ: Educational Testing Service.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skill diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 1031–1038). Amsterdam: Elsevier.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hartigan, J. A. (1978). Asymptotic Distributions for Clustering Criteria. *The Annals of Statistics, 6*, 117–131.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics, 28*, 100–108.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3044108).

Hartz, S. M., & Roussos, L. A. (October 2008). *The fusion model for skill diagnosis: Blending theory with practicality* (Research report No. RR-08-71). Princeton, NJ: Educational Testing Service.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Heinen, T. (1996). *Latent class and discrete latent trait models*. Newbury Park, CA: Sage.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika, 56*, 255–278.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Köhn, H.-F., & Chiu, C.-Y. (2016). A proof of the duality of the DINA model and the DINO model. *Journal of Classification, 33*, 171–184.

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*, 112–132.

Köhn, H.-F., Chiu, C.-Y., & Brusco, M. J. (2015) Heuristic cognitive diagnosis when the Q-matrix is unknown. *British Journal of Mathematical and Statistical Psychology, 68*, 268–291.

Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models*. New York: Plenum.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Leighton, J., & Gierl, M. (2007) *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli, 19*, 1790–1817.

Ma, W., & de la Torre, J. (2017). *GDINA: The generalized DINA model framework. R package version 1.4.2*. Retrieved from the Comprehensive R Archive Network [CRAN] website https://cran.r-project.org/web/packages/GDINA/

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley, CA: University of California Press.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379–416.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

McLachlan, G., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1–32.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Park, Y. S., & Lee, Y.-S. (2011). Diagnostic cluster analysis of mathematics skills. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (IERI monograph series, Vol. 4, pp. 75–107). Hamburg, Germany: IERI.

Picciano, A. G. (2012). The evolution of big data and learning analytics in American higher education. *Journal of Asynchronous Learning Networks, 16*, 9–20.

Pollard, D. (1981). Strong consistency of *K*-means clustering. *The Annals of Statistics*, *9*(1), 135–140.

Pollard, D. (1982). Quantization and the method of *K*-means. *IEEE Transactions on Information Theory*, *28*, 199–205.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2016). *CDM: Cognitive diagnosis modeling. R package version 4.7-0*. Retrieved from the Comprehensive R Archive Network [CRAN] website https://cran.r-project.org/web/packages/CDM/

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement Interdisciplinary Research and Perspectives, 6*, 219–262.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement. Theory, methods, and applications*. New York: Guilford.

Steinhaus, H. (1956). Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III, IV*(12), 801–804.

Steinley, D. (2003). Local optima in *K*-means clustering: What you don't know may hurt you. *Psychological Methods, 8*, 294–304.

Steinley, D. (2006). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*, 1–34.

Steinley, D., & Brusco, M. J. (2007). Initializing *K*-means batch clustering: A critical analysis of several techniques. *Journal of Classification, 24*, 99–121.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67*, 485–518.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational and Behavioral Statistics, 12*, 55–73.

Tatsuoka, K. K. (2009). *Cognitive assessment. An introduction to the rule space method*. New York: Routledge/Taylor & Francis.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Vermunt, J. K. (1997). *Loglinear models for event histories*. Thousand Oaks, CA: Sage.

von Davier, M. (2005, September). *A general diagnostic model applied to language testing data* (Research report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–301.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement – Interdisciplinary Research and Perspectives, 7*, 67–74.

von Davier, M. (2014a). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series, 2014*(2), 1–39.

von Davier, M. (2014b). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71.

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*, 85–100.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. (2018). Tracking skill acquisition with cognitive diagnosis models: Applications to spatial rotation skills. *Journal of Educational and Behavioral Statistics, 43*, 57–87.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Willse, J., Henson, R., & Templin, J. (2007). *Using sum scores or IRT in place of cognitive diagnosis models: Can existing or more familiar models do the job?* Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Zheng, Y., & Chiu, C.-Y. (2016). *NPCD: Nonparametric methods for cognitive diagnosis. R package version 1.0-10.* Retrieved from the Comprehensive R Archive Network [CRAN] website http://CRAN.R-project.org/package=NPCD

Chapter 6
The General Diagnostic Model

Matthias von Davier

Abstract The general diagnostic model (GDM) allows modeling dichotomous and polytomous item responses under the assumption that respondents differ with respect to multiple latent skills or attributes, and that these may be distributed differently across populations. Item responses can be of mixed format, dichotomous and/or polytomous, and skills/attributes can be binary, polytomous ordinal, or continuous. Variables that define populations can be observed, latent as in discrete mixture models, or partially missing. Unobserved grouping variables can be predicted based on hierarchical extensions of the GDM. It was shown that through reparameterization, the GDM contains the DINA as well as the logistic G-DINA, which is the same as the log-linear cognitive diagnostic model (LCDM), as special cases, and hence can fit all models that can be specified in these frameworks. Taken together, the GDM includes a wide range of diagnostic models, as well as item response theory (IRT), multidimensional IRT (MIRT), latent class models, located latent class models, multiple group and mixture versions of these models, as well as multilevel, and longitudinal extensions of these. This chapter introduces the GDM by means of a formal description of basic model assumptions and their generalizations and describes how models can be estimated in the GDM framework using the *mdltm* software. The software is free for research purposes, can handle very large databases up to millions of respondents and thousands of items, and provides efficient estimation of models through utilization of massively parallel estimation algorithms. The software was used operationally for scaling the PISA 2015, 2018, and PIAAC 2012 main study databases, which include hundreds of populations, grouping variables, and weights, and hundreds of test forms collected over five assessment cycles with a combined size of over two million respondents.

M. von Davier (✉)
National Board of Medical Examiners (NBME), Philadelphia, PA, USA
e-mail: mvondavier@nbme.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_6

6.1 The General Diagnostic Model

The general diagnostic model (GDM) is a solution for modeling dichotomous or polytomous item responses collected in mixed format assessment under the assumption that respondents differ with respect to multiple latent skills or attributes, and that these traits may be distributed differently across populations. In the GDM, latent skills/attributes can be binary as in most diagnostic models (e.g., von Davier, DiBello, & Yamamoto, 2008; Rupp, Templin, & Henson, 2010; von Davier & Lee, Chap. 1, this volume) indicating mastery or non-mastery, or ordered polytomous attributes as in discrete IRT models (e.g., Heinen, 1996; Haberman, von Davier, & Lee, 2008), or continuous latent variables, approximated by numerical integration over a discretized ability interval as customarily done in IRT and multidimensional IRT (MIRT) estimation.

Item responses can be of mixed format, i.e., dichotomous or polytomous, as can be the latent structure, combining dichotomous, polytomous, and continuous latent skill variables. Variables that define populations can be observed, missing, or partially missing. Unobserved grouping variables can be predicted based on higher-order covariates when using hierarchical extensions of the GDM.

The GDM is one of the most general models for diagnostic classification. It was shown that through reparameterization, the GDM contains the DINA as well as the log-linear cognitive diagnostic model (LCDM) (and with that, the logistic G-DINA, which is identical to the LCDM), as special cases, and hence can fit all models that can be specified within these frameworks. Taken together, these results on the GDM show that this approach includes a wide range of diagnostic models, as well as IRT, MIRT, latent class models, located latent class models, as well as multiple group and mixture distribution versions of these models, and finally multilevel extensions of these.

This chapter introduces the GDM and provides a mathematical description of model assumptions and their formalization. In addition, this chapter describes GDM extensions that provide longitudinal, as well as discrete mixture models and multiple population versions of the approach, which are particularly useful for the assessment of measurement invariance across populations. The chapter closes with a section on model equivalencies and a summary.

6.2 Notation

In this chapter capital letters are used to denote random variables, and lower-case letters for realizations of random variables. The following notation is used for response variables, covariates, and latent skill variables:

Let $\boldsymbol{X} = (X_1, \ \ldots \ , X_K)$ denote K binary or polytomous response variables and let $\boldsymbol{x}_n = (x_{n1}, \ \ldots \ , x_{nK})$ denote the observed responses for test takers $n = 1, \ \ldots \ , N$. Let $\boldsymbol{Y} = (Y_1, \ \ldots \ , Y_J)$ and $\boldsymbol{y}_n = (y_{n1}, \ \ldots \ , y_{nJ})$ denote a vector of J covariates

and their realizations observed on test taker n. Finally, consider a grouping variable with $g_n \in \{1, \ldots, G\}$ for all test takers. For mixture models, g_n is unobserved, for multiple group models, g_n is completely or partially observed (von Davier & Yamamoto, 2004a, 2004b; von Davier & Carstensen, 2007).

Let the dimensionality of the latent structure (i.e., the number of latent variables) in the model be denoted by D, and let $\boldsymbol{A} = (A_1, \ldots, A_D)$ denote the vector of latent attributes, and let the attribute pattern of person $n = 1, \ldots, N$ be denoted by $\boldsymbol{a_n} = (a_{n1}, \ldots, a_{nD})$. Alternatively for continuous traits $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_D)$ and $\boldsymbol{\theta_n} = (\theta_{n1}, \ldots, \theta_{nD})$ may be used.

Let $P(\boldsymbol{A} = \boldsymbol{a})$, $P(\boldsymbol{A} = (a_1, \ldots, a_D))$ denote probabilities of the latent trait or attribute distribution, if space requires a short form we may use $P(a_1, \ldots, a_D)$ where needed, and the same for marginal distributions of observed variables such as $P(y_1, \ldots, y_J)$. For conditional probabilities, we use $P(X_i = x_{ni} | \boldsymbol{A} = (a_1, \ldots, a_D))$ or alternatively, $P(x_{ni} | a_1, \ldots, a_D)$, as well as $P(a_1, \ldots, a_D | g)$ and $P(g | y_1, \ldots, y_J)$ for conditional attribute distributions, and conditional population distributions given covariates.

Like most latent trait, and more general, most latent structure models, the GDM assumes local independence given skill attribute vector, that is,

$$P\left(\boldsymbol{X} = (x_1, \ldots, x_K) \,|\boldsymbol{A} = (a_1, \ldots, a_D)\right) = \prod_{i=1}^{K} P\left(x_i | a_1, \ldots, a_D\right).$$

For a given attribute distribution $P(a_1, \ldots, a_D)$, the marginal probability of a response pattern can be calculated as

$$P\left(\boldsymbol{X} = (x_1, \ldots, x_K)\right) = \sum_{\boldsymbol{A}=(a_1,\ldots,a_D)} P\left(a_1, \ldots, a_D\right) \prod_{i=1}^{K} P\left(x_i | a_1, \ldots, a_D\right).$$

Note that the attribute distributions $P(a_1, \ldots, a_D)$ pose an estimation challenge if the number of skills D grows large, or the number of skill levels per skill grows, or both. For D = 5, the number of binary skill patterns is $32 = 2^5$, for 5 skills with 4 levels each, we look at 1024 different skill patterns, the same as for 10 binary skill attributes. Instead of estimating potentially thousands of nuisance parameters, von Davier and Yamamoto (2004a) and Xu and von Davier (2006, 2008) propose a log-linear skill-attribute distribution. This provides a more parsimonious approach by assuming

$$\ln P\left(a_1, \ldots, a_D\right) = \lambda_0 + \sum_{d=1}^{D} \lambda_{d1} a_d + \sum_{d=1}^{D-1} \sum_{e=d+1}^{D} \lambda_{de2} a_d a_e$$

as a model for the skill distributions. This log-linear skill attribute model involves main effects and first order interactions, and can approximate the unobserved distribution of skill attributes for a wide range of cases. Xu and von Davier (2006, 2008)

provide examples and show how this approach compares well in terms of balancing parsimony and model fit to fully parameterizing the skill attribute distribution. von Davier (2018) notes that the skill distribution defined above generalizes the second order exponential distribution (Tsao, 1967; Holland, 1990) to a model that provides a polytomous extension of the Ising (1926) model, an approach that recently gained interest in what is sometimes called network psychometrics (Marsman et al., 2018).

The GDM for binary and polytomous variables, as well as extensions involving multiple observed and unobserved populations, longitudinal data structures, as well as data structures with covariates and multilevel structure will be introduced in the subsequent sections.

6.3 GDM for Binary and Ordinal Skill Attributes

This section introduces the GDM (von Davier, 2005) for dichotomous and polytomous response variables as well as binary and ordinal skill attributes. For simplicity, all variables, responses and attributes, can be considered binary, however, the GDM does not require this, unlike other diagnostic models. In the binary case, the skill levels of any attribute can be interpreted as mastery $a_d = 1$ versus non-mastery $a_d = 0$ of skill d. Let $\boldsymbol{a} = (a_1, \ldots, a_D)$ be a D-dimensional skill profile consisting of polytomous skill attributes $a_d, d = 1, \ldots, D$. Then, the probability of a polytomous response $x \in \{0, \ldots, m_i\}$ to item i under the GDM, for a person with skill attributes $\boldsymbol{a} = (a_1, \ldots, a_D)$ is given by

$$P\left(X_i = x | \boldsymbol{A} = (a_1, \ldots, a_D)\right) = \frac{\exp\left[\beta_{ix} + \sum_{d=1}^{D} \gamma_{ixd} h_i\left(q_{id}, a_d\right)\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{iy} + \sum_{d=1}^{D} \gamma_{iyd} h_i\left(q_{id}, a_d\right)\right]}$$

with item parameters β_{ix}, γ_{ixd} for $x = 1, \ldots, m_i$ and $i = 1, \ldots, I$ and $d = 1, \ldots, D$. The rows of the Q-matrix $\boldsymbol{q_i} = (q_{i1}, \ldots, q_{id})$ are constants. As customary in other diagnostic models, the q_{id} relate item i to skill dimension d and determine whether the skill is required for that item. If the skill d is required for item i, then $q_{id} = 1$, and $q_{id} = 0$ otherwise. The helper functions $h_i(q_{id}, a_d)$ map the dichotomous or polytomous skill levels a_d and Q-matrix entries q_{id} to the real numbers. In most cases, the same mapping will be adopted for all items, so one can drop the index i. While different mappings are conceivable, for the sake of parsimony and replicability, it appears not assuming variations of item level models is sensible, unless different (mixed format) item types are used. The $h()$ mapping defines how the Q-matrix entries and the skill levels interact (von Davier, 2005; von Davier et al., 2008).

For polytomous data, the number of slope parameters γ_{ixd} in the equation above grows with the number of response categories $x = 1, \ldots, m_i$ per item, as is easily verified. For a more parsimonious model, a restriction $\gamma_{ixd} = x\gamma_{id}$ can be

implemented. In addition, the helper functions can be specified as $h_i(q_{id}, a_d) = q_{id}a_d$ for all items i. This defines the partial credit GDM:

$$P\left(X_i = x|\boldsymbol{A} = (a_1, \ldots, a_D)\right) = \frac{\exp\left[\beta_{ix} + \sum_{d=1}^{D} x\gamma_{id}q_{id}a_d\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{iy} + \sum_{d=1}^{D} y\gamma_{id}q_{id}a_d\right]}$$

with item parameters β_{ix}, γ_{id} for $x = 1, \ldots, m_i$ and $i = 1, \ldots, I$ and $d = 1, \ldots, D$. For binary responses, this model reduces to

$$P\left(X_i = 1|\boldsymbol{A} = (a_1, \ldots, a_D)\right) = \frac{\exp\left[\beta_i + \sum_{d=1}^{D} \gamma_{id}q_{id}a_d\right]}{1 + \exp\left[\beta_i + \sum_{d=1}^{D} \gamma_{id}q_{id}a_d\right]}$$

with vector-valued item parameter $(\beta_i, \gamma_{i1}, \ldots, \gamma_{iD})$ which will be also written using the alternative notation $\boldsymbol{\lambda}_i = (\lambda_{i0}, \lambda_{i1}, \ldots, \lambda_{iD})$, as used in some of the chapters in this volume. Component-wise equivalency holds, i.e., $\lambda_{id} = \gamma_{id}$ and $\lambda_{i0} = \beta_i$. Define $M_{\boldsymbol{q}_i,\boldsymbol{a}} = (1, q_{i1}a_1, \ldots, q_{iD}a_D)$. Then we can write

$$\frac{\exp\left[\beta_i + \sum_{d=1}^{D} \gamma_{id}q_{id}a_d\right]}{1 + \exp\left[\beta_i + \sum_{d=1}^{D} \gamma_{id}q_{id}a_d\right]} = \frac{\exp\left[\boldsymbol{\lambda}_i^{\boldsymbol{T}} M_{\boldsymbol{q}_i,\boldsymbol{a}}\right]}{1 + \exp\left[\boldsymbol{\lambda}_i^{\boldsymbol{T}} M_{\boldsymbol{q}_i,\boldsymbol{a}}\right]}.$$

The GDM contains a large class of well-known psychometric models as special cases, including IRT, MIRT, latent class models, located latent class models, HYBRID models, as well as MIRT models for longitudinal data (von Davier, Xu, & Carstensen, 2011).

Several cognitive diagnostic models (CDMs), including the G-DINA (de la Torre, 2011) with logistic link, the DINA (Junker & Sijtsma, 2001), and the LCDM (Henson, Templin, & Willse, 2009) turn out to be special cases of the GDM (von Davier, 2014, 2016) as well.

This following paragraph summarizes how the equivalent GDM that includes LCDM and G-DINA can be defined, while a more in depth derivation is provided in an extended section below. For the vector-valued skill variable $\boldsymbol{A} = (A_1, \ldots, A_D)$ with realizations $(a_1, \ldots, a_D)$ define the extended $E = 2^D - 1$ dimensional skill attribute space $\boldsymbol{A}^* = (A_1, \ldots, A_D, A_{12}, A_{13}, \ldots, A_{D-1D}, A_{123}, A_{124}, \ldots, A_{123\ldots D}) = (A^*_1, \ldots, A^*_E)$, where the realizations are based on a constraint of the extended skill space defined by $a_{12} = a_1a_2, a_{13} = a_1a_3, \ldots, a_{123\ldots D} = \prod_d a_d$ which implies the following restriction on the distribution of $\boldsymbol{A}^*$:

$$P\left(A_{ij} = 1|a_1, \ldots, a_D\right) = \begin{cases} 1 & a_ia_j = 1 \\ 0 & \text{otherwise} \end{cases},$$

and analog for $P(A_{ijk} = 1|\, a_1, \ \ldots\ , a_D) \ \ldots\ P(A_{1\ \ldots\ D} = 1|\, a_1, \ \ldots\ , a_D)$. This restriction ensures that the number of independent parameters remains $2^D - 1$. Then we can define

$$P\left(X_i = x|\left(a^*{}_1, \ldots, a^*{}_E\right)\right) = \frac{\exp\left[\beta_{ix} + \sum_{e=1}^{E} x\gamma_{id} q^*{}_{ie} a^*{}_e\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{iy} + \sum_{e=1}^{E} y\gamma_{ie} q^*{}_{ie} a^*{}_e\right]}$$

where Q^* is the extended Q-matrix that includes the specification which of the extended skill variables (representing the skill interactions) are included. For the DINA, for example, only $q^*_{iE} = q^*_{i1\ldots D} = 1$ while all other entries are equal to zero. The subsequent sections present how this model can be used in the extensions of the GDM that allow specification of multilevel/hierarchical as well as mixture IRT and hierarchical latent class models within the GDM framework. Within this family, multi-level, mixture and multiple group versions of the LCDM/G-DINA can be estimated for dichotomous and polytomous response variables and attributes.

6.4 Mixture Distribution Extensions of the GDM

von Davier (2008b) introduced the discrete mixture distribution version of the GDM, referred to as the MGDM subsequently. In discrete mixture models for item response data (Mislevy & Verhelst, 1990; Rost, 1990; von Davier & Rost, 2006, 2016), the probability of observed responses $\boldsymbol{x} = (x_1, \ \ldots\ , x_K)$ depends on the unobserved latent trait $\boldsymbol{x} = (a_1, \ \ldots\ , a_D)$ and a subpopulation indicator $g \in \{1, \ \ldots\ , G\}$, which may also be unobserved. The rationale for mixture distribution models is that observations from different subpopulations may either differ in their skill distribution, or in their item parameters, or both. The complete data for a test taker n is $D_n = (\boldsymbol{x_n}, \boldsymbol{a_n}, g_n)$, of which only $\boldsymbol{x_n}$ is observed in mixture distribution models. In multiple population models, $(\boldsymbol{x_n}, g_n)$ is observed, and in partially observed mixtures (von Davier & Yamamoto, 2004b) $\boldsymbol{a_n}$ is unobserved and some (or most) of the g_n are missing.

To accommodate multiple observed or unobserved populations, the conditional independence assumption as well as the expression for the marginal probability are augmented by a population indicator g. The conditional independence assumption becomes

$$P\left(\boldsymbol{X} = (x_1, \ldots, x_K)\,|\,(a_1, \ldots, a_D)\,, g\right) = \prod_{i=1}^{K} P\left(x_i|\,(a_1, \ldots, a_D)\,, g\right)$$

and the marginal probability of the response vector is extended to include population indicator g as well as the conditional attribute distribution given population g. That is

$$P\left(x_1,\ldots,x_K\right)=\sum_{g=1}^{G}\pi_g\sum_{a_1,\ldots,a_D}P\left(a_1,\ldots,a_D|g\right)\prod_{i=1}^{K}P\left(x_i|a_1,\ldots,a_D,g\right)$$

with population specific weights $\pi_g = P(G = g)$, also referred to as mixing proportions or class sizes. The probability of a response vector $(x_1, \ldots, x_K)$ for a respondent in population g with skill attribute pattern $(a_1, \ldots, a_D)$ is

$$P\left(x_1,\ldots,x_K|\left(a_1,\ldots,a_D\right),g\right)=\prod_{i=1}^{K}\frac{\exp\left[\beta_{ixg}+\sum_{d=1}^{D}\gamma_{ixdg}h\left(q_{id},a_{\mathrm{d}}\right)\right]}{1+\sum_{y=1}^{m_i}\exp\left[\beta_{iyg}+\sum_{d=1}^{D}\gamma_{iydg}h\left(q_{id},a_d\right)\right]}$$

with class-specific item difficulties β_{ixg} and class specific slope parameters γ_{ixdg}. The item parameters of the MGDM can be constrained across populations, or estimated freely across populations, with only those constraints imposed that are required for identification (von Davier, 2008b). Three important special cases of the MGDM can be distinguished:

1. The measurement invariance model, which assumes the same item parameters across populations.
2. The equivalent groups model, which assumes equivalent skill distributions, while item parameters may differ across populations.
3. The naïve scale alignment or equating model, which allows different item parameters, and different ability distributions, while identification constraints assure that item parameters maintain the same average difficulties and slope parameters across populations.

The measurement invariance model assures that item functions are the same across populations, while ability (or skill attribute) distributions $P(a_1, \ldots, a_D|\, g)$ may vary. Xu and von Davier (2008) show how to apply this approach in a cross-sectional design, and von Davier et al. (2011) demonstrate the use of the multi-group GDM in a longitudinal design. The model that assumes the same ability distributions across populations, and allows different item parameters across populations is useful when looking at differences in item functioning across populations. Models of this type can be used to look at adaptations of assessments to new delivery platforms, for example when moving tests from paper to computer, or from PCs to tablets (von Davier et al., in press). A multiple group model that only includes basic constraints to ensure identifiability may or may not use the same types of constraints across populations. The sum of item difficulties could be equal to -1.0 in one group, and $+$ 1.5 in another group, as any constant would work. Using the same constants for item parameter constraints across populations does not provide any stronger equality constraint than using different constants (von Davier & von Davier, 2007, 2011). However, the use of same constants is used for example in IRT equating and mimics classical observed score equating in this regard.

The MGDM can be used to implement complex population structures to reflect linking designs and to test invariance assumptions across multiple populations, including mode effect studies and differential item functioning (von Davier et al., 2011; von Davier & Rost, 2016). In order to utilize covariates that are available in addition to response data and population indicators, the MGDM can be further extended to a multilevel/hierarchical diagnostic model. The next section introduces the hierarchical GDM, which is based on the mixture distribution extensions of the GDM.

6.5 Hierarchical/Multilevel Extensions of the GDM

This section introduces the hierarchical GDM, an extension of the mixture distribution GDM based on a multilevel version of latent class analysis (Vermunt, 2003). There are many examples of models that accommodate grouping variables as the driver of varying associations between observed and latent variables. Hierarchical linear models allow random intercepts and random slopes (Bryk & Raudenbush, 1992). Hybrid models (Yamamoto, 1989; von Davier, 1996) assume that in some subpopulations, there is systematic co-variation between latent trait and observed response variables, whereas in other subpopulations, there is no such relationship. Multiple group models (e.g., Bock & Zimowski, 1997) assume that the same item response model with different sets of parameters holds in different groups. Mixture distribution item response models (von Davier & Rost, 2016) assume that an IRT or MIRT holds, but with different parameters in different subpopulations.

The hierarchical extension of the GDM presented by von Davier (2010) allow checking the impact of clustering or nesting of the sample, such as data collected for students nested within schools in large scale educational surveys, on the structural parameter estimates of the model. Moreover, the hierarchical version of the GDM allows studying differences of skill attribute distributions across clusters.

For the developments presented here, the extension of the LCA to a hierarchical model (e.g., Vermunt, 2003, 2004) is of importance. In addition to the latent class or grouping variable g, the hierarchical extension of the LCA assumes that each observation n is characterized by additional variables $(y_{1n}, \ldots, y_{xn})$. Respondents are then sorted into equivalency classes or clusters $s = S(y_{1n}, \ldots, y_{xn})$ with the same vector of covariates, that is, we have

$$s_n = s_m \text{ if } (y_{1n}, \ldots, y_{xn}) = (y_{1m}, \ldots, y_{xm}) .$$

The clusters s identified by these covariates may represent schools or school types, or groups with homogeneous background, so that the cluster s may either represent the hierarchical structure of the data collection, or equivalency classes based on respondents who share a common background. Class membership g_n is thought of as an individual outcome. While two respondents may share the same

cluster, i.e., $s_n = s_m$, they may belong to the same, or to different unobserved populations, or latent classes: Both $g_n \neq g_m$ and $g_n = g_m$ are permissible. In addition, it is assumed that the skill attribute distribution depends only on the group indicator g and no other variable, that is,

$$P\left(a_1, \ldots, a_D | g, z\right) = P\left(a_1, \ldots, a_D | g\right)$$

for any random variable z, including the clustering of respondents. More specifically, the hierarchical GDM (HGDM) assumes that the distribution of classes g may differ across clusters s, so that one may have $\pi_{g|s_1} = P\left(g|s_1\right) \neq P\left(g|s_2\right) = \pi_{g|s_2}$, while the differences in distribution of skill attributes across clusters are fully explained by differences across classes, i.e., it is assumed that

$$P\left(a_1, \ldots, a_D | s\right) = \sum_g P\left(g|s\right) P\left(a_1, \ldots, a_D | g\right).$$

The marginal distribution of observed responses $x_1, \ldots, x_K$ under the HGDM is given by

$$P\left(x_1, \ldots, x_K\right) = \sum_s P(s) \sum_{g=1}^{G} \pi_{g|s} \sum_{a_1, \ldots, a_D} P\left(a_1, \ldots, a_D | g\right) \prod_{i=1}^{K} P\left(x_i | a_1, \ldots, a_D, g\right)$$

with the $P(a_1, \ldots, a_D|\, g)$ representing the distribution of the skill patterns in group g, and the $P(x_i|\, a_1, \ldots, a_D, g)$ denote the distribution of the responses x_i conditional on skill pattern $a_1, \ldots, a_D$ and group g. A HGDM that assumes measurement invariance across clusters and across groups can be written as

$$P\left(x_1, \ldots, x_K\right) = \sum_s P(s) \sum_{g=1}^{G} \pi_{g|s} \sum_{a_1, \ldots, a_D} P\left(a_1, \ldots, a_D | g\right) \prod_{i=1}^{K} P\left(x_i | a_1, \ldots, a_D\right)$$

which assumes conditional response probabilities $P(x_i|\, a_1, \ldots, a_D)$ that do not depend on g. The increased complexity of HGDMs over nonhierarchical versions lies in the fact that the mixing proportions $P(g|\, s)$ depend on the cluster variable s. If effects of the group membership are considered fixed effects, this increases the number of group or class size parameters linearly with the number of clusters. If the clusters are considered to be random draws from a population, the effects $\pi_{g\,|\,s}$ can be modeled with a Dirichlet distribution. von Davier (2010) describes estimation of class-specific item difficulties β_{ixg} and the class specific slope parameters γ_{idg} as well as the estimation of the other quantities, including the fixed and random effect versions of the $\pi_{g\,|\,s}$.

6.6 Estimation of the GDMs

The GDM, being a constrained latent class model (von Davier, 2009a, 2009b) can be estimated with the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). While other approaches are also viable, the EM-algorithm has a proven track record of providing a solution to a wide range of missing data problems. Estimation of complex models for large scale data collections with sample sizes in the millions is possible with this approach, as well as estimation involving moderate sample sizes.

6.7 Standard EM-Algorithm

The EM-algorithm is an iterative solution for maximizing a likelihood function that involves an incomplete data problem. Recall that in terms of respondent level variables, the complete data for respondent n is $(x_{n1}, \ldots, x_{nK}, s_n, g_n, a_{n1}, \ldots, a_{nD})$,of which only $x_{n1}, \ldots, x_{nK}, s_n$ is observed. If there are no clusters s, or no subpopulations g, we may leave out s_n, g_n. This means that, at a minimum, the skill attribute vector $(a_{n1}, \ldots, a_{nD})$ is missing for all respondents.

The necessary expressions for expected counts needed in the E-step and estimation equations for the M-step are provided by von Davier (2005). The EM algorithm for estimating the GDM was implemented in the *mdltm* software (von Davier, 2005, 2008a). von Davier (2008b, 2010) describes the estimation of the mixture GDM, and the hierarchical GDM, respectively. Here, we provide the steps and calculations involved in the EM-algorithm as pseudo code for the GDM without mixture or hierarchical extensions:

- E-step: Calculate expected counts of the skill attribute distribution, and the conditional distribution of observed variables given latent skills. The E-step involves Bayesian methods to calculate the posterior distribution of skill attributes given observed variables and aggregates these posteriors across respondents in order to generate the required expected counts.
- M-Step: Maximize the likelihood of the observed data, given expected counts, by conducting a single optimization step and updating item parameters by means of a Newton Raphson, a Quasi-Newton, or Gradient based method.

E-step and M-step are alternated and repeated until a convergence criterion is reached. von Davier (2005, 2008a) provides detailed maximization and expectation equations for each of the parameter types. Typically, convergence is assessed by comparing the log-likelihood of the data under two sets of parameters, the current ones, and the ones from the previous iteration. If the difference is smaller than a pre-specified amount, iterations are terminated and the current parameters are considered final.

6.8 Parallel EM Algorithm

von Davier (2016) develops a parallel EM algorithm for the latent variable models presented above. In an empirical comparison of results from a parallel EM algorithm developed for the GDM and the customary (serial) execution of the EM iterations it is shown that parallel execution is able to speed up the calculations by orders of magnitude. In particular, both versions of the algorithm yield identical results, while the runtime reduction, due to the parallel implementation of the EM, is substantial. von Davier (2017) reports on the second generation of this parallel EM algorithm for generalized latent variable models on the basis of the GDM (von Davier, 2005, 2008a, 2014). This new development further improves the performance of the parallel-E parallel-M algorithm presented by von Davier (2016). This is achieved by means of parallel language constructions that produce additional gains in performance. The additional gain over the factors of 6 to 20 seen when testing the first parallel version is ranging from 20–30% depending on the nature of the model being estimated. The estimation of a MIRT model for large scale data may show a larger reduction in runtime compared to a multiple-group model which has a structure that is more conducive to parallel processing of the E-step. Multiple population models can be arranged such that the parallelism directly exploits the ability to estimate multiple latent variable distributions separately in independent threads of the algorithm. Table 6.1 shows results for a 12-core Intel XEON workstation. von Davier (2016, 2017) also report results obtained from a 32-core AMD server with four 8-core CPUs, in which case speedups are obtained that reach a factor of 20 or more.

The examples cover a wide range of latent variable models from IRT to MIRT, confirmatory models, multi-trait multi-method models, and latent class models. The items are mixed format, and their number varies from 54 to 293, the number of respondents varies from 2026 to 1.8 million, the number of populations ranges from 1 to 312, and the number of dimensions d ranges from 1 to 5.

Table 6.1 Results of the comparison of parallel-E parallel-M versions 1 and 2 on a 12-Core Xeon workstation as well as the sequentially executed algorithm

Scales	Model	Groups	Items	Sample	Serial EM	Parallel EM	Speedup
1	2PL/GPCM	312	293	1,614,281	1356	153	807%
1	2PL/GPCM	283	133	1,803,599	1127	96	963%
7	MTMM	1	214	7377	2465	343	785%
2	MIRT	1	175	5763	44	11	400%
2	MIRT	1	175	5763	1155	145	708%
NA	LCA	54	54	246,112	7444	964	716%
5	MIRT	1	150	2026	2499	726	263%

Times for Serial EM and Parallel EM are given in seconds. Likelihoods of the converged solutions and numbers of iterations are identical for parallel and serial EM

The development of this high-performance computational approach facilitates estimating advanced psychometric models for very large datasets in a matter of minutes rather than hours. Unlike methods relying on simplifications of the likelihood equations that are only available for some constrained problems such as the bifactor model (Gibbons & Hedeker, 1992; Rijmen & Jeon, 2013), the approach presented by von Davier (2016, 2017) is applicable to any multidimensional latent variable model, including MIRT models, multi-group and mixture models, as well as longitudinal approaches such as growth curve and growth mixture models. Massive gains in processing speed can be realized by using the Parallel-E Parallel-M algorithm with Tile Reduction (PEPM-TR) for estimating generalized latent variable models.

The parallel version of the EM algorithm was utilized operationally in the analyses of the PISA 2015 main assessment data. This dataset was linked to previous cycles of the PISA assessment through data stemming from three prior assessment rounds. Together, the calibration sample that covers four cycles from 2006 to 2015 contains response patterns from roughly two million students, sampled from 300 or more populations, and up to 300 items. While the serial EM algorithm would take up to 20 min or more per calibration, depending on the actual size of the database, the parallel EM based calibrations took between 2–3 min. per run on a workstation with two CPUs and six CPU-cores each. Given that multiple calibrations (often 8 or more) are needed to iteratively evaluate fit and differential item functioning and apply item treatments, these gains reduce the operational burden considerably.

6.9 Testing Model Fit

Since the GDM and many other diagnostic models can be understood as constrained latent class models, all methods that are available in the context of model checking and goodness of fit for categorical data analyses can be applied to these models as well. Examples of these global model fit approaches that aim at the level of selecting a model will be discussed in the next section.

In addition, more specific fit diagnostics commonly used in IRT can be readily applied to the GDM, as it contains IRT and MIRT models as special cases. The second subsection discusses examples of fit diagnostics such as item and person fit measures that are available in the *mdltm* software evaluating the fit of the GDM.

6.10 Global Goodness of Fit

Goodness of fit, in the context of models for categorical data analysis can be either assessed using statistical testing procedures, if certain regularity conditions are met, or by means of resimulation (von Davier, 1997), or posterior predictive checks (e.g., Sinharay, 2003). Alternatively, information criteria providing a heuristic decision

rule can be applied (e.g., Akaike, 1974). In this section, the focus is on model selection using heuristics. The reason for this is that for very large sample sizes, and sparse data structures that come naturally with applications of these models to data from educational assessments, the power to detect deviations at the global level is such that practically all models would be rejected if the selection was based on a significance testing procedure. It is a long-standing observation that when the sample is large enough, all models can be proven wrong (Berkson, 1938; Box, 1976). Model selection heuristics based on information criteria, however, are helping in the comparison of a range of models in order to select the relatively best fitting model.

A list of model selection tools that can be used for selection among competing model specifications within the family of GDM is provided below. These model selection tools are available through the *mdltm* software (von Davier, 2005, 2008a) and are provided by default in the output that is produced upon completion of a GDM estimation.

- **Information criteria:** Among the most commonly used model selection heuristics, information criteria (Akaike, 1974) are customary for use with models for continuous and categorical data. Information criteria have the general form

$$AnyIC_{Model} = -2 * LogLik_{Model} + W_{AnyIC} * NPar_{Model}$$

  where $NPar_{Model}$ is the number of free model parameters, $LogLik_{Model}$ as the log-likelihood of the data under the model, and W_{AnyIC} is the weight of the parameter penalty term. The Akaike (1974) information criterion (AIC) specifies this as a constant $W_{AnyIC} = 2$, for Schwarz (1978) BIC it is $W_{AnyIC} = \ln(N)$ where N is the sample size. For Bozdogan's consistent AIC, it is $W_{AnyIC} = (1 + \ln(N))$.
- **Log-penalty:** The log-penalty criterion is a transform of an estimate of the expected log-likelihood per response. It is based on work by Gilula and Haberman (1994) referencing Savage's (1972) work. The log-penalty uses the expected log likelihood per item response, but is also subjected to a penalty function. The log-penalty can be calculated as a simple transformation that modifies the AIC index, namely,

$$\text{LogPen} = \frac{AIC}{N^*K}$$

  where N is the sample size and K is the number of items, in the case of complete data without missing responses. For designs with planned missingness or for data with a lot of omitted responses the ratio needs to involve only the actual number of observed responses.
- **Classification consistency:** Vermunt (2010) emphasizes that classification-based approaches are often evaluated with respect to classification accuracy. For models such as LCA, mixture IRT, and mixture GDM the following two related fit measures can be used: The average classification error

$E = \frac{1}{N}\sum_{n=1}^{N}\left[1 - \max_{\{c\}}\left(P\left(C = c|x_{1n}, .., x_{kn}\right)\right)\right]$ and the reduction of classification error $\lambda = 1 - \frac{E}{1-\max_{\{c\}}(P(C=c))}$ compared to the classification based on marginal probabilities only. The closer this measure is to 1.0, the better the classification accuracy is compared to classification based on the marginals only.

The GDM software *mdltm* provides the above global measures of fit for all models in the GDM framework. This allows model comparisons across models with discrete and continuous latent traits, and with varying numbers of latent populations.

6.11 Item and Person Fit

Item and Person Fit are customary fit indices (Molenaar, 1983) used to assess the extent to which a test-takers or an item response profile agree with the assumed model regularities that are at the basis of the inferences we wish to make. While there are likelihood-based item fit indices (e.g., Rost & von Davier, 1994; von Davier & Molenaar, 2003), we focus here on response residuals. To test whether a response vector is 'expected' or 'aberrant', response residuals have been used for extended periods of time in categorical data analysis (e.g., Haberman, 2009). For a binary response x_{ui} of person u on item i, define

$$Z\left(x_{ui}\right) = \frac{x_{ui} - P\left(X = 1|u, i\right)}{\sqrt{P\left(X = 1|u, i\right)\left[1 - P\left(X = 1|u, i\right)\right]}}.$$

Note that the probabilities in the above equation are not known constants, but need to be approximated. In *mdltm*, these are estimated by means of plugging in the item parameter estimates, and integrating over the posterior distribution of skill-attributes for each respondent.

For polytomous responses, the expected and the observed response are compared, and the variance is calculated accordingly. To identify aberrant response vectors, sums of squared residuals across respondents are used to evaluate items, or sum of squared residuals across all responses of a person to evaluate person fit. A standardized person fit index is given by

$$\xi_n = \frac{\sum_{i=1}^{I}\left[Z\left(x_{ni}\right)\right]^2 - I}{\sqrt{2I}}$$

which can be considered approximately normally distributed for larger numbers of items. For small I, the statistic $\sum_{i=1}^{I}\left[Z\left(x_{ui}\right)\right]^2$ is used with the chi-square distribution for I degrees of freedom (see, e.g., Haberman, 2009; Adams, 2010). A standardized item fit index can be calculated as

$$\xi_i = \frac{\sum_{n=1}^{N}\left[Z\left(x_{ni}\right)\right]^2 - N}{\sqrt{2N}}.$$

The *mdltm* software allows generating response residuals that are appended to the output file containing person estimates. In addition, the GDM software *mdltm* also generates measures of root mean square deviation and mean deviation of item response functions.

6.12 The GDM Includes LCDM and DINA

The sections above presented the GDM for dichotomous, polytomous, and mixed format tests, as well as with latent skill structures that can contain combinations of binary skills, ordinal skills, and continuous latent variables, for data that is potentially sampled from multiple populations. This yields a modeling framework that contains IRT, MIRT, discrete latent trait, and latent class, localized latent class models, as well as mixture and multilevel extensions of these models.

In addition to these properties of the GDM, it can be shown that the approach is also a more general model than the LCDM (Henson & Templin, Chap. 8, this volume) and the G-DINA with logistic link function (de la Torre & Minchen, Chap. 7, this volume). von Davier (2011, 2013) showed that the GDM contains the DINA model as a special case, and this result was generalized using the LCDM as an example of a GDM special case by von Davier (2014). The following section demonstrates how the DINA and the LCDM can be specified as special cases in the GDM framework.

This presentation follows von Davier (2014) and shows that the LCDM (and with it, the DINA, and logistic G-DINA, which is the same as the LCDM) are equivalent to a special case of the GDM. This constrained GDM yields identical parameter estimates compared to the LCDM (and G-DINA ...) based on a transformed set of compensatory skills. Recall that the GDM for binary data can be rewritten with $M_{\boldsymbol{q}_i,\boldsymbol{a}} = (1, q_{i1}a_1, \ldots, q_{iD}a_D)$. Then we can write

$$P\left(X_i = 1|a_1, \ldots, a_D\right) = \frac{\exp\left[\lambda_{i0} + \sum_{d=1}^{D}\lambda_{id}q_{id}a_d\right]}{1 + \exp\left[\lambda_{i0} + \sum_{d=1}^{D}\lambda_{id}q_{id}a_d\right]} = \frac{\exp\left[\boldsymbol{\lambda}_i^T M_{\boldsymbol{q}_i,\boldsymbol{a}}\right]}{1 + \exp\left[\boldsymbol{\lambda}_i^T M_{\boldsymbol{q}_i,\boldsymbol{a}}\right]}.$$

With item parameter $(\beta_i, \gamma_{i1}, \ \ldots \ , \gamma_{iD})$ written as $\boldsymbol{\lambda}_i = (\lambda_{i0}, \lambda_{i1}, \ \ldots \ , \lambda_{iD})$. Using the same notation, the LCDM was defined by Henson, Templin, and Willse (2009) as a model that additionally contains skill interactions, that is, the exponent in the above equation, $\lambda_{i0} + \sum_{d=1}^{D}\lambda_{id}q_{id}a_d$, becomes

$$\lambda_{i0} + \sum\nolimits_{d=1}^{D}\lambda_{id}q_{id}a_d + \sum_{d<e}\lambda_{ied}q_{ide}a_d a_e + \sum_{d<e<f}\lambda_{idef}r_{idef}a_d a_e a_f + \cdots + \lambda_{i1..D}r_{i1..D}a_1..a_D$$

where the interaction terms of up to order D-1 are included if the interaction inclusions indicators are $r_{i\ldots} = 1$ for that term.

Reorganizing and collecting terms by defining the conjunctions $a_{de} = a_d a_e$, $a_{def} = a_d a_e a_f, \ldots, a_{1..D} = \prod_{d=1}^{D} a_d$ and optionally renaming $r_{i\ldots} = q_{i\ldots}$ yields

$$\lambda_{i0} + \sum\nolimits_{d=1}^{D} \lambda_{id} q_{id} a_d + \sum_{d<e} \lambda_{ied} q_{ide} a_{de} + \sum_{d<e<f} \lambda_{idef} q_{idef} a_{def} + \cdots + \lambda_{i1..D} q_{i1..D} a_{1..D}.$$

Note that the $r_{i\ldots} = q_{i\ldots}$ in the LCDM and G-DINA are essentially the rows of a second design matrix that includes information about which interactions of order 2 to D are included in the model. This extra design matrix can be made part of the Q-matrix by defining additional skills, and assigning these skills a constrained distribution that reflects whether the underlying source skills are present or not.

With the notations introduced earlier, let $\boldsymbol{a}^* = \left(a_1^*, \ldots, a_I^*\right)$ denote an $E = (2^D - 1)$-dimensional skill vector that we refer to as the transformed skill space. Similarly, let $\boldsymbol{q}_i^*$ denote the E-dimensional transformed Q-matrix entry for item i. We define the entries of the transformed $\boldsymbol{q}_i^*$ for a given source $\boldsymbol{q}_i$ and the distributional constraints required for $P\left(\boldsymbol{A}^* = \left(a_1^*, \ldots, a_E^*\right)\right)$ in the following. Von Davier (2014) defines the extended Q-matrix, $\boldsymbol{Q}^*$, as follows. The total number of skills in the transformed skill space is given by

$$E = 2^D - 1 = \sum_{d=1}^{D} \binom{D}{d} = \frac{D!}{d!\,(D-d)!}$$

where each of the $d = 1 \ldots D$ summands in this equation represents the number of terms $\binom{D}{d}$ that involve d skills. This is the number of transformed skill columns required for each interaction of order $v - 1$. If we look at a Q-matrix with four skills, there are four main effects, corresponding to the λ_{id} parameters. There are $6 = \frac{4*3}{2}$ interactions involving two skills (for the λ_{ide} parameters) and $\binom{4}{3} = 4$ interactions involving three skills (for the λ_{idef} parameters), and one, $\binom{4}{4} = 1$, that involves all four skills, $\lambda_{i1..4}$. Implicitly, there is also $\binom{4}{0} = 1$ zero-skill "interaction" that corresponds to the baseline item parameter λ_{i0} present in all items. Note that $1 + 4 + 6 + 4 + 1 = 16 = 2^4$. Instead of having different types of skills and skill interaction parameters, one can rename the skills again, that is,

$$\left(a_1, a_2, a_3, a_{4,} a_{12}, a_{13}, a_{14}, a_{23,} a_{24,} a_{34,} a_{123,} a_{124,} a_{134,} a_{234}, a_{1234}\right)$$

becomes

$$(a_1, \ldots, a_{15})$$

and

$$\left(q_1, q_2, q_3, q_4, r_{12}, r_{13}, r_{14}, r_{23}, r_{24}, r_{34}, r_{123}, r_{124}, r_{134}, r_{234}, r_{1234}\right)$$

becomes

$$(q_1, \ldots, q_{15})$$

and one obtains a GDM with 2^4 - 1 = 15 skills. von Davier (2014) provides general equations that allow an algorithmic renaming to retain the association between the interaction based LCDM and the transformed skills GDM.

The only additional requirement to make this model equivalent to the LCDM is a constraint on the skill attribute distribution. Each transformed GDM skill a_v^* corresponds to a conjunction of some number of source skills, i.e., for each a_v^* there are skill indices d, e, ... with

$$P\left(a_v^* = 1 | a_d = 1, a_e = 1, \ldots\right) = 1.$$

The general case showing that this produces an LCDM-equivalent compensatory skill space that is used in the GDM as presented in von Davier (2014), together with an empirical example that provides evidence that the LCDM, and the transformed skill space based GDM indeed produce identical results. The key to seeing this is that for each of the potential skill interactions in the source LCDM, one defines the required skill entries in the transformed skill space as additional skills, but with a constrained skill distribution that fulfills the condition given above. This defines an item model that is equivalent to the LCDM definition but operates on a transformed additive/compensatory skill space.

One could argue that this 'unnecessarily' explodes the skill space beyond need: However, this should be evaluated on the basis that the transformed skill space distribution is constrained. For $2^D - 1 = E$ skills, an unconstrained distribution would involve $2^E - 1 = 2^{(2^D-1)} - 1$ free probabilities, in our example with 4 skills LCDM, this would mean $2^{15} - 1 = 32767$ free parameters. However, the constraints on the skill space enforce that

$$1 - P\left(a_v^* = 1 | a_d = 1, a_e = 1, \ldots\right) =$$

$$P\left(a_v^* = 1 | a_d = 0 \; or \; a_e = 0 \; or \ldots\right) = 0.$$

Hence, among the potentially 32,767 free probabilities, many are fixed to zero, and the remainder includes large numbers of equality constraints. Indeed, it turns there are only $2^4 - 1 = 15$ freely estimated skill probabilities in the example with 4 skills, just as in the equivalent LCDM formulation. This also holds for the general case of D skills. For a discussion of the general case, refer to von Davier (2013, 2014).

6.13 Summary

The GDM provides a general modeling framework for skills diagnosis suitable for dichotomous and polytomous response variables, and for skill variables with binary or ordinal levels. The GDM framework allows comparative analyses of data from educational assessments with a wide range of models. IRT and MIRT approaches as well as diagnostic models such as the DINA and the LCDM, and logistic G-DINA turn out to be special cases of the GDM (von Davier, 2013, 2014). This chapter provided and overview of these results as well as relevant references for further study.

The GDM can be estimated with the standard EM-algorithm, and multiple implementations are available, either as stand-alone programs, or as packages implemented on top of statistical computation tools such as R. In addition, the GDM can be estimated with high-performance tools that utilize state-of-the-art parallel programming paradigms (von Davier, 2016, 2017) allowing analysis of very large complex data bases with multidimensional models in considerably less time than standard algorithms.

Examples of operational use include the analysis of the international databases of the PIAAC 2012 and the PISA 2015 and 2018 data collections, which include estimation of models with more than 300 populations, about 2,000,000 student response patterns, and up to 300 items. These applications included IRT, MIRT and analyses with diagnostic models. These examples show that this general modeling family that includes diagnostic modeling approaches has become a part of large scale operational data processing, while providing innovative approaches to modeling and linking assessments across samples.

References

Adams, R. (2010). *Case (Person) Fit and Residuals*. Notes on ConQuest 3.0 software features. https://www.acer.org/files/Conquest-Notes-3-CaseFitAndResiduals.pdf

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chisquare test. *Journal of the American Statistical Association, 33*, 526–542.

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791–799. https://doi.org/10.1080/01621459.1976.10480949

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods* (1st ed.). Newbury Park, CA: Sage Publications.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Minchen, N. D. (this volume). The G-DINA model framework. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1–38.

Gibbons, R. D., & Hedeker, D. R. (1992, September). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436. https://doi.org/10.1007/BF02295430

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*, 645–656.

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models*. ETS RR-09-15.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. RR-08-45. ETS Research Report.

Heinen, T. (1996). *Latent class and discrete latent trait models, similarities and differences*. Thousand Oaks, CA: Sage Publications.

Henson, R., & Templin, J. L. (this volume). Loglinear cognitive diagnostic model (LCDM). In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika, 74*, 191–210.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55*(1), 5–18.

Junker, B. W., & Sijtsma, K. (2001). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement, 25*(3), 211–220. https://doi.org/10.1177/01466210122032028

Marsman, M., Borsboom, D., Kruis, J., Epskamp, S., van Bork, R., Waldorp, L. J., … Maris, G. (2018). An introduction to network psychometrics: Relating Ising network models to item response theory models. *Multivariate Behavioral Research, 53*(1), 15–35. https://doi.org/10.1080/00273171.2017.1379379

Mislevy, R. J., & Verhelst, N. D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195–215.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48*, 49–72.

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research, 206*, 647–662.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271–282.

Rost, J., & von Davier, M. (1994). A conditional item fit index for Rasch models. *Applied Psychological Measurement, 18*, 171–182.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: The Guilford Press.

Savage, L. J. (1972). *The foundation of statistics*. Dover publications.

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464. https://doi.org/10.1214/aos/1176344136

Sinharay, S. (2003). Practical applications of posterior predictive model checking for assessing fit of the common item response theory models (Research Report RR03–33). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-03-33-Sinharay.pdf

Tsao, R. (1967). A second order exponential model for multidimensional dichotomous contingency tables with applications in medical diagnosis. Unpublished doctoral thesis, Harvard University, Department of Statistics.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213–239.

Vermunt, J. K. (2004). An EM-algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica, 58*(2), 220–233.

Vermunt, J. K. (2010). Latent class models. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (Vol. 7, pp. 238–244). Oxford, UK: Elsevier.

von Davier, M. (1996). Mixtures of polytomous Rasch models and latent class models for ordinal variables. In F. Faulbaum & W. Bandilla (Eds.), *Softstat 95 – advances in statistical software 5*. Stuttgart, Germany: Lucius & Lucius.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research, 2*, 29–48. Retrieved January 14, 2010, from: http://www.dgps.de/fachgruppen/methoden/mpr-online/issue3/art5/davier.pdf.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series, 2005*, 1–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2008a). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2008b). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models* (pp. 255–276). Charlotte, NC: Information Age Publishing.

von Davier, M. (2009a). Mixture distribution item response theory, latent class analysis, and diagnostic mixture models. In S. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 11–34). Washington, DC: APA Press.

von Davier, M. (2009b). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement – Interdisciplinary Research and Perspectives, 7*(1, March), 67–74.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28. Retrieved April 26, 2012, from: http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf

von Davier, M. (2011). *Equivalency of the DINA model and a constrained general diagnostic model. ETS-RR-11-37*. Princeton: ETS Research Report Series.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model—two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (Research Report No. RR-14-40). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12043

von Davier, M. (2016). *High-performance psychometrics: The parallel-E parallel-M algorithm for generalized latent variable models*. (ETS Research Report ETS-RR-16-34).

von Davier M. (2017) New results on an improved parallel EM algorithm for estimating generalized latent variable models. In van der Ark L., Wiberg M., Culpepper S., Douglas J., Wang WC. (eds) *Quantitative psychology*. IMPS 2016. Springer Proceedings in Mathematics & Statistics (Vol 196). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-56294-0_1

von Davier, M. (2018). Diagnosing diagnostic models: From von Neumann's elephant to model equivalencies and network psychometrics. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 59–70. https://doi.org/10.1080/15366367.2018.1436827

von Davier, M., & Lee, Y.-S. (this volume). Introduction: From latent class analysis to DINA and beyond. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linkage and scale transformations. *Methodology, 3*, 115–124.

von Davier, M., & von Davier, A. A. (2011). A general model for IRT scale linking and scale transformations. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking*. New York: Springer.

von Davier, M., & Carstensen, C. H. (Eds.). (2007). *Multivariate and mixture distribution Rasch models*. New York, NY: Springer.

von Davier, M., & Molenaar, I. W. (2003). A person-fit index for Polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68*, 213–228.

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 643–768). Amsterdam, The Netherlands: Elsevier.

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, 2nd ed., pp. 393–406). Boca Raton, FL: CRC Press.

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151–176). Toronto, Canada: Hogrefe & Huber Publishers.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336.

von Davier, M., & Yamamoto, K. (2004a, October). *A class of models for cognitive diagnosis*. Paper presented at the 4th Spearman Conference, Philadelphia, PA.

von Davier, M., & Yamamoto, K. (2004b). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measurement, 28*(6), 389–406.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (Research Report, RR-06-08). Princeton, NJ: ETS.

Xu, X., & von Davier, M. (2008). *Linking with the general diagnostic model*. (Research Report RR-08-08). Princeton, NJ: Educational Testing Service.

Yamamoto, K. (1989). *A hybrid model of IRT and latent class models*.. Research Report RR-89-41. Princeton, NJ: Educational Testing Service.

Chapter 7
The G-DINA Model Framework

Jimmy de la Torre and Nathan D. Minchen

Abstract The development of cognitive diagnosis models (CDMs) has been prolific since the turn of the century; however, they have often been developed in such a way that they lack an overall connective framework. The purpose of this chapter is to review the G-DINA framework. As a general model, it subsumes several simpler and widely-known CDMs; as a general framework, it has also served as the foundation for a variety of model extensions and new methodological developments. We will also discuss associated topics, which include model estimation, Q-matrix validation, computerized adaptive testing, and model selection as they relate to the reviewed models.

7.1 Introduction

Cognitive diagnosis models (CDMs) can be viewed as restricted versions of the more general latent class models. In particular, the number of latent classes, as well as their interpretation, are known a priori when CDMs are involved. Further restrictions can be posited regarding how the underlying attributes interact to produce the observed responses. These interactions (or condensation rules; Maris, 1999) include conjunctive, disjunctive, and additive processes (de la Torre, 2011). Assuming a specific underlying process involves the use of a reduced or constrained CDM such the DINA model (Haertel, 1989; Junker & Sijtsma, 2001), DINO model (Templin & Henson, 2006), LLM (Maris, 1999), R-RUM (Hartz, 2002), and *A*-CDM (de la Torre, 2011). Although more interpretable and requiring smaller sample sizes, reduced models can also lead to poorer model-data fit when they

J. de la Torre (✉)
Division of Learning, Development and Diversity, University of Hong Kong, Hong Kong, China
e-mail: j.delatorre@hku.hk

N. D. Minchen
Pearson, Bronx, NY, USA
e-mail: nathan.minchen@pearson.com

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_7

are incorrectly specified (e.g., Chen & de la Torre, 2013). Notwithstanding their own shortcomings, general or saturated CDMs, such as the G-DINA model (de la Torre, 2011), LCDM (Henson, Templin, & Willse, 2009), and GDM (von Davier, 2008), can be used as an alternative to reduced CDMs to minimize the impact of potential model misspecifications. With the exception of the GDM, which can be specified more generally, the CDMs above are designed for dichotomous attributes and dichotomous responses. It should be noted that when dichotomous attributes and dichotomous responses are involved, the G-DINA model, which is typically written using the identity link function, the LCDM and GDM, which are based on the logit link function, and any saturated CDMs in other link functions (e.g., log) are equivalent to each other. To accommodate a wider range of attribute and response types, extensions of CDMs need to be considered.

An integral component of most, if not all, CDM specifications, general or otherwise, is the Q-matrix (Tatsuoka, 1983). In its typical formulation, a Q-matrix is a $K \times D$ matrix that identifies the subset of attributes measured by each item, where K is the number of items and D the number of attributes measured by the test. The attribute specification for item j is given in the binary $D-$length vector, $\boldsymbol{q}_j$. Correspondingly, the latent variable in CDM is typically a binary $D-$length vector, $\boldsymbol{\alpha}_l$, where $l = 1, \ldots, L = 2^D$, the number of latent classes. As will be shown later, both $\boldsymbol{q}_k$ and $\boldsymbol{\alpha}_l$ may require some modifications before they can be used in conjunction with CDM extensions.

The valid use of scores derived from CDMs presupposes that the model is adequate for the data. To this end, steps need to be taken to ensure that a discrete latent variable can fit the data, the correct CDMs are employed, and Q-matrix entries are correctly specified. In addition, for greater efficiency, simpler models should be preferred over more complex models whenever appropriate.

Given the large number of CDMs that currently exists, a unifying framework from which these models can be viewed is needed to better understand their unique natures and the extent to which these models relate to each other. Moreover, a coherent framework that permits implementation of various CDM-related procedures can allow for the appropriate use of CDMs to be evaluated more systematically. As will be discussed below, the G-DINA model framework aims to accomplish this two-pronged objective. In addition, the G-DINA as a model can serve as the foundation on which CDM extensions can be built.

7.2 The G-DINA Model Framework

7.2.1 The G-DINA Model

Without loss of generality, assume that the first D_k^* attributes are required for item k, and let $\boldsymbol{\alpha}_{lk}^*$ be the D_k^*-length *reduced* attribute vector, $l = 1, \ldots, 2^{D_k^*}$, which retains only the attributes required for item k. The item response function (IRF) of

the G-DINA model is given by

$$g[P(X_k = 1 \mid \boldsymbol{a}_{lk}^*)] = \phi_{k0} + \sum_{d=1}^{D_k^*} \phi_{kd} a_{ld} + \sum_{d'=d+1}^{D_k^*} \sum_{d=1}^{D_k^*-1} \phi_{kdd'} a_{ld} a_{ld'} + \ldots + \phi_{12\ldots D_k^*} \prod_{d=1}^{D_k^*} a_{ld}, \tag{7.1}$$

where $g[\cdot]$ is either the identity, log, or logit link function, ϕ_{k0} is the intercept, ϕ_{kd} is the main effect due to mastering a_d, and each of the remaining $\phi_{k\cdot}$ represent all possible higher-order interaction effects, ranging from two-way to D_k^*-way. When $g[\cdot]$ is the logit link, it is equivalent to the LCDM, which has also been shown to be equivalent to a GDM with an extended skill space (von Davier, 2014).

The G-DINA model is considered a saturated CDM because it contains $2^{D_k^*}$ parameters corresponding to the $2^{D_k^*}$ latent groups in item k. As shown by de la Torre (2011), several reduced models can be derived from the G-DINA model by constraining its parameters. The DINA model is equivalent to the G-DINA model with all but the intercept and the highest-order interaction effect set to zero. Its IRF in the G-DINA notation is

$$g[P(X_k = 1 \mid \boldsymbol{a}_{lk}^*)] = \phi_{k0} + \phi_{12\ldots D_k^*} \prod_{d=1}^{D_k^*} a_{ld}. \tag{7.2}$$

Similarly, the DINO model can be obtained from the G-DINA model using the following constraints: $\phi_{kd} a_{ld} = -\phi_{kdd'} = \cdots = (-1)^{D_k^*+1} \phi_{12\ldots D_k^*}$. Thus, its IRF can be written as

$$g[P(X_k = 1 \mid \boldsymbol{a}_{lk}^*)] = \phi_{k0} + \phi_{kd} a_{ld}. \tag{7.3}$$

Finally, when all the interaction effects are set to zero, as in,

$$g[P(X_k = 1 \mid \boldsymbol{a}_{lk}^*)] = \phi_{k0} + \sum_{D=1}^{D_k^*} \phi_{kd} a_{ld}, \tag{7.4}$$

the G-DINA model in the identity, log, or logit link is equivalent to the *A*-CDM, R-RUM, or LLM, respectively. Although the additive property is inherent to a particular link function (e.g., R-RUM is multiplicative when converted to the identity link), Ma, Iaconangelo, and de la Torre (2016) noted the interchangeability of the three additive models for some item parameter combinations. As a whole, recognizing that the G-DINA model subsumes a number of reduced CDMs has important implications in model comparison and model-data fit evaluation.

7.3 Model Extensions

7.3.1 G-DINA Model for Polytomous Attributes

Although the G-DINA model is a general CDM, it is only so with respect to dichotomous attributes. However, some educational applications may benefit from a finer-grained, and therefore, more instructionally-relevant classification of students. For example, classifying students as having no mastery, basic mastery, and advanced mastery of the skills might be of interest. The middle-school proportional reasoning (PR) assessment described by Tjoe and de la Torre (2013a,b) measures two polytomous attributes, namely, (a) comparing and ordering of fractions, where level 0 represents nonmastery of the attributes, level 1 the ability to compare two fractions, and level 2 the ability to order three or more fractions; and (b) constructing ratios and proportions, where level 0 again represents nonmastery, level 1 the ability to construct a single ratio, and level 2 the ability to construct a proportion, which is made up of two ratios. Such classifications require polytomous attributes.

Define $\boldsymbol{\alpha}_l = \{\alpha_{ld} \mid \alpha_{ld} \in (0, 1, \ldots, M_d)\}$ as the polytomous attribute vector, and again, assume that the first D_k^* attributes are required for item k. The reduced attribute vector in this context can be written as $\boldsymbol{\alpha}_{lk}^* = \{\alpha_{ld}, \ldots, \alpha_{lD_k^*}\}$. When there are no constraints on the model, item k involves $M_1 \cdot M_2 \cdots M_{D_k^*}$ latent groups. A saturated CDM for this item would require the same number of parameters, making it too complex to be viable in most practical testing situations. Chen and de la Torre (2013) proposed the polytomous G-DINA (pG-DINA) model as a lower-complexity CDM that can accommodate polytomous attributes. To reduce the number of latent groups, and hence complexity of the corresponding CDM, the pG-DINA model assumes that, for each attribute within an item, an examinee can be classified as either at or below the required attribute level. Examinees on or above the cutoff are assumed to have the necessary attribute mastery level to answer the item correctly, whereas those below it do not. Chen and de la Torre (2013) referred to this as the *specific attribute level mastery* (SALM) assumption. The reduced polytomous attribute vector $\boldsymbol{\alpha}_{lk}^*$ can be converted to a reduced dichotomous attribute vector $\boldsymbol{\alpha}_{lk}^*$ as follows: $\boldsymbol{\alpha}_{lk}^* = \{I(\alpha_{ld} \geq q_{kd})\}$, for $d = 1, \ldots, D_k^*$. After the conversion, $\boldsymbol{\alpha}_{lk}^*$ can be used in the IRF given in (7.1) to model a wide variety of attribute interactions.

In general, the conversion process in the pG-DINA model reduces the number of latent groups to $2^{D_k^*}$ for item k regardless of the number of levels of the attributes involved. It should also be noted that the pG-DINA model differs from other polytomous CDMs (e.g., GDM) in that the attribute level required for an item is defined by domain or subject-matter experts a priori, whereas in other CDMs, only the attribute, but not the level, need to be specified. This distinct feature of the pG-DINA model implies a modification of the Q-matrix – instead of only 0 and 1, $q_{kd} \in (0, 1, \ldots, M_d - 1)$. Using the PR assessment data, Chen and de la Torre (2013) and de la Torre (2015) have shown that the pG-DINA model provides a better fit when compared to the G-DINA model. These results indicate that the pG-DINA model is not only theoretically appealing, but also empirically more appropriate.

7.3.2 G-DINA Model for Polytomous Response

Although items that can be scored as either right or wrong (i.e., 1/0) remains the most common item type in large-scale assessments, items that can be scored with ordered polytomous categories are also available. In the CDM literature, it is not uncommon for these scores to be dichotomized and analyzed using existing CDMs for dichotomous response. In recent years, a number of CDMs for ordered polytomous response have been proposed, including the GDM for graded responses (von Davier, 2008), the polytomous LCDM (Hansen, 2013) and the sequential G-DINA (sG-DINA; Ma & de la Torre, 2016) model. Of these, only the sG-DINA model considers the possibility that, within the same item, the subset of attributes being measured can vary from one response category to another.

The sG-DINA model assumes that the problem-solving process is sequential in nature, and allows for different subsets of attributes to be associated with different steps or categories. In the sG-DINA model, the Q-matrix is modified to accommodate $\boldsymbol{q}_{kh}$, the q-vector for category h of item k, where $h = 1, 2, \ldots, H_k$. Note that for ordered polytomous response, 0 is one of the response categories (i.e., $X_k = \{h \mid h \in (0, 1, \ldots, H_k)\}$), but this category does not require a q-vector. Hence, instead of K rows, the modified Q-matrix contains $\sum_{k=1}^{K} H_k$ rows.

We can again assume that the first D_k^* are the required attributes for category h of item k. Conditional on the reduced attribute pattern $\boldsymbol{\alpha}_{lh}^*$, the probability of a correct response to category h of item k given the previous step is correctly answered is denoted by

$$S_k(h|\boldsymbol{\alpha}_{lh}^*) = P(X_{k,h} = 1 \mid X_{k,h-1}, \boldsymbol{\alpha}_{lh}^*). \tag{7.5}$$

$S_k(h|\boldsymbol{\alpha}_{lh}^*)$ is referred to as the *processing function* in the item response theory literature (Samejima, 1973). The processing function can be more generally formulated by using various link functions. In doing so, the IRF of the G-DINA model given in (7.1) can be used as the processing function to model a range of attribute interactions associated with the category response. Based on the sG-DINA model, the probability of obtaining a score of h on item k is given by

$$P(X_k = 1|\boldsymbol{\alpha}_{lh}^*) = [1 - S_k(h+1|\boldsymbol{\alpha}_{lh}^*)] \prod_{h'=1}^{h} S_k(h'|\boldsymbol{\alpha}_{lh'}^*), \tag{7.6}$$

where

$$S_k(h|\boldsymbol{\alpha}_{lh}^*) = \begin{cases} 1, & \text{if } h = 0 \\ 0, & \text{if } h = H_k + 1 \end{cases}.$$

The sG-DINA model is said to be restricted when the attribute-category associations are known. However, for some items, only the attribute-item associations

can be ascertained. For these items, the unrestricted version of the sG-DINA is used, where the same subset of attributes are specified for all categories. Although more general, and therefore more flexible, fitting the unrestricted sG-DINA model when the restrictions are appropriate can lead to suboptimal results. Originally the unrestricted sG-DINA model was designed for ordered responses; however, Ma and de la Torre (2016) have shown that the model can also be used in conjunction with nominal response, and is equivalent to the partial credit DINA model (de la Torre, 2010) and the nominal response diagnostic model (Templin, Rupp, Henson, Jang, & Ahmed, 2008). Finally, as expected, the sG-DINA model performs better than the G-DINA model fitted to dichotomized polytomous data.

7.3.3 G-DINA Model for Continuous Response

Although a number of CDMs for dichotomous and polytomous responses are available, modeling continuous response in the CDM context is in its infancy. With the proliferation of computer-based testing, perhaps the most obvious and readily-available source of continuous response is latency, or response time. However, other item formats such as placing a mark on a line segment (e.g., Noel, 2014; Noel & Dauvier, 2007) and probability testing (e.g., Ben-Simon, Budescu, & Nevo, 1997) can also yield continuous responses. For illustration purposes, we will use response time to represent continuous response throughout the chapter. As de la Torre and Minchen (2016), Minchen and de la Torre (2018) and Minchen, de la Torre, and Liu (2017) have shown, response time in the CDM context may itself be the work product of interest, or it could be viewed as a type of process data and used in conjunction with response accuracy.

The first CDM to handle responses of a strictly continuous type is the continuous DINA (cDINA) model proposed by Minchen et al. (2017). Like the DINA model, the cDINA model involves the same latent variable $\boldsymbol{a}_l$, classifies the examinees into one of two latent groups – those who have the required attributes for the items ($\eta_{lk} = 1$), and those who do not ($\eta_{lk} = 0$). However, instead of a single parameter (i.e., slip or guessing) governing the response of one particular group, the item response of a latent group in the cDINA model is governed by two parameters, representing the mean and standard deviation of the group's, say, response time on item k. It should also be noted that unlike dichotomous response where examinees in group $\eta_{lk} = 1$ are expected to score higher, the expected response time of the same examinees can be longer or shorter depending on the context of application. The real data example in Minchen et al. shows that examinees in $\eta_{lk} = 1$ are more engaged with problems that they are equipped to handle, resulting in longer response times.

Using the cDINA model, the cumulative distribution function for the response t_{lk} on item k given $\boldsymbol{a}_l$ can be written as

$$P(T_k \leq t|\boldsymbol{a}_l) = \int_0^t f_{k\eta}(t_k)dt_k, \tag{7.7}$$

where

$$f_{k\eta}(t_k) = \frac{1}{t_k\sqrt{2\pi\sigma_{k\eta}^2}} \exp\Big[-\frac{(\ln t_k - \mu_{k\eta})^2}{2\sigma_{k\eta}^2}\Big], \tag{7.8}$$

which is the lognormal distribution with group-specific parameters $\mu_{k\eta}$ and $\sigma_{k\eta}$ for $\eta = 0, 1$.

The continuous G-DINA (cG-DINA; Minchen & de la Torre, 2018) is a straightforward generalization of the cDINA model. Instead of two latent groups, the cG-DINA model allows for a unique response distribution to be associated with each of the $2^{D_k^*}$ latent groups; thus it is characterized by $2^{D_k^*+1}$ parameters. The cumulative distribution of the cG-DINA model for the response t_{lk} is similar to that in (7.7) with the exception that the lognormal distribution in (7.8) involves $\mu_{k\eta}$ and $\sigma_{k\eta}$ for $\eta = 1, 2, \ldots, 2^{D_k^*}$, and a one-to-one correspondence between η and $\boldsymbol{a}_{lk}^*$ can be made.

The cG-DINA model is a saturated model because each of the $2^{D_k^*}$ latent groups is characterized by a unique parameter set $(\mu_{k\eta}, \sigma_{k\eta})$. By imposing the constraints $\mu_{k1} = \cdots = \mu_{k,2^{D_k^*}-1}$ and $\sigma_{k1} = \cdots = \sigma_{k,2^{D_k^*}-1}$, the cDINA model can be easily derived from the cG-DINA model. Similar constraints can be imposed to derive a disjunctive CDM from the cG-DINA model. However, as noted earlier, CDMs for continuous response are in their nascent stages. At present, it is not clear how additive CDMs in this context should be formulated or what constraints on $\mu_{k\eta}$ and $\sigma_{k\eta}$ are needed to derive them from the saturated model. Furthermore, the existence of two parameters per latent group raises the possibility that the constrained model for $\mu_{k\eta}$ may not be the same as that for $\sigma_{k\eta}$.

7.4 Estimation

An expectation-maximization (EM) implementation of marginalized maximum likelihood estimation (MMLE) can be used to obtain parameter estimates of the CDMs discussed above (e.g., de la Torre, 2009, 2011). Specifically, under the assumption of local independence, the log-marginalized likelihood of the dichotomous response data can be written as

$$\ell(\boldsymbol{X}) = \log \prod_{n=1}^{N} \sum_{l=1}^{2^D} P(\boldsymbol{X}_n \mid \boldsymbol{a}_l) p(\boldsymbol{a}_l), \tag{7.9}$$

where

$$P(\boldsymbol{X}_n \mid \boldsymbol{a}_l) = \prod_{k=1}^{K} P(X_{nk} = 1 \mid \boldsymbol{a}_l)^{X_{nk}} [1 - P(X_{nk} = 1 \mid \boldsymbol{a}_l)]^{1-X_{nk}}. \tag{7.10}$$

The MMLE/EM algorithm implements E-step and M-step iteratively item by item until convergence. In particular, the E-step calculates $N_{\boldsymbol{a}^*_{lk}} = \sum_{n=1}^{N} P(\boldsymbol{a}^*_{lk}|\boldsymbol{X}_n)$, the expected number of individuals having the attribute pattern $\boldsymbol{a}^*_{lk}$, and $R_{\boldsymbol{a}^*_{lk}} = \sum_{n=1}^{N} x_{nk} P(\boldsymbol{a}^*_{lk}|\boldsymbol{X}_n)$, the number of individuals with attribute pattern $\boldsymbol{a}^*_{lk}$ expected to answer item k correctly. Note that $P(\boldsymbol{a}^*_{lk}|\boldsymbol{X}_n)$ is the posterior probability of individual n having attribute pattern $\boldsymbol{a}^*_{lk}$. In the M-step, as shown in de la Torre (2011), the maximum likelihood estimate of $P(X_k = 1 \mid \boldsymbol{a}^*_{lk})$ is given by

$$\hat{P}(X_k = 1 \mid \boldsymbol{a}^*_{lk}) = \frac{R_{\boldsymbol{a}^*_{lk}}}{N_{\boldsymbol{a}^*_{lk}}}. \tag{7.11}$$

The item parameters $\boldsymbol{\phi}$ in (7.1) can be derived from (7.11) via the ordinal least-squares approach.

For the DINA and DINO models, the $2^{D^*_k}$ latent groups are further partitioned into two non-overlapping groups η_{k0} and η_{k1}, where individuals in the former and latter groups are expected to answer item k incorrectly and correctly, respectively. The maximum likelihood estimate of the probability of success for individuals in group η_{ku} where $u \in (0, 1)$ is

$$\hat{P}(X_k = 1 \mid \eta_{ku}) = \frac{\sum_{\boldsymbol{a}^*_{lk} \in \eta_{ku}} R_{\boldsymbol{a}^*_{lk}}}{\sum_{\boldsymbol{a}^*_{lk} \in \eta_{ku}} N_{\boldsymbol{a}^*_{lk}}}. \tag{7.12}$$

For A-CDM, LLM and R-RUM, the maximum likelihood estimate can be found using various optimization functions based on $R_{\boldsymbol{a}^*_{lk}}$ and $N_{\boldsymbol{a}^*_{lk}}$. The parameters of the pG-DINA model can be estimated as in the G-DINA model after converting $\boldsymbol{\alpha}^*_{lk}$ to reduced dichotomous attribute vector $\boldsymbol{a}^*_{lk}$. For the sG-DINA model, the following objective function is maximized in the M-step,

$$\ell_k = \sum_{l=1}^{2^{D^*_k}} \sum_{h=0}^{H_k} R_{\boldsymbol{a}^*_{lkh}} \log\left[P(X_k = h|\boldsymbol{a}^*_{lk})\right],$$

where $R_{\boldsymbol{a}^*_{lkh}} = \sum_{n=1}^{N} I(x_{nk} = h) P(\boldsymbol{a}^*_{lk}|\boldsymbol{X}_n)$ is the number of individuals with attribute pattern $\boldsymbol{a}^*_{lk}$ expected to obtain a score of h on item k. Note that the EM algorithm for estimating the sG-DINA model can also be implemented at the category level after transforming the polytomous data to dichotomous data with missing values using the mapping matrix (Ma, 2018).

For the cG-DINA model, the conditional likelihood given in (7.10) can be written as

$$P(\boldsymbol{t}_n \mid \boldsymbol{a}_l) = \prod_{k=1}^{K} f_j(t_{nk}|\boldsymbol{a}_l). \tag{7.13}$$

Following several steps of derivation, the maximum likelihood estimates of $\mu_{k\eta}$ and $\sigma^2_{k\eta}$ can be shown to be equal to

$$\hat{\mu}_{k\eta} = \sum_{n=1}^{N} p^*(\boldsymbol{a}_{lk}|\boldsymbol{t}_n) \log t_{nk}, \tag{7.14}$$

and

$$\hat{\sigma}^2_{k\eta} = \sum_{n=1}^{N} p^*(\boldsymbol{a}_{lk}|\boldsymbol{t}_n)(\log t_{nk} - \hat{\mu}_{k\eta})^2, \tag{7.15}$$

respectively, where $p^*(\boldsymbol{a}_{lk}|\boldsymbol{t}_n)$ is the posterior probability (normalized across the N examinees) of examinee n being in the reduced attribute pattern $\boldsymbol{a}_{lk}$.

Unlike traditional IRT, where the prior ability distribution can be reasonably specified, for example, using $N(0, 1)$, the multinomial attribute distribution $p(\boldsymbol{a}_l)$ in CDM cannot be readily determined a priori. A convenient way of specifying $p(\boldsymbol{a}_l)$ is to employ the empirical Bayes estimate. In particular, we let $p^{(c+1)}(\boldsymbol{a}_l)$, the prior distribution at iteration $c + 1$, be equal to the $p^{(c)}(\boldsymbol{a}_l \mid \boldsymbol{X})$, the posterior distribution at iteration c. It should be noted that in the CDM context, estimation of the item response model can impact the joint attribute distribution estimate, and vice versa. Therefore, in situations where the impact of model misspecification on item parameter estimates needs to be isolated, one can use the G-DINA model to arrive at the correct attribute distribution estimate in the first step, and, fixing the attribute distribution, use the EM algorithm to obtain the item parameter estimates of the reduced model in the second step.

7.5 G-DINA Model-Based Methodologies

7.5.1 Q-Matrix Validation

In typical CDM applications, Q-matrices are built by subject-matter experts. In addition to subjective judgments, experts may not reach complete agreement on each of the Q-matrix entries. For these reasons, the correctness of the entire Q-matrix cannot be guaranteed. To address this issue, statistical procedures, referred to in the literature as *empirical Q-matrix validation* methods, have been proposed.

De la Torre and Chiu (2016) proposed the G-DINA model discrimination index (GDI) for an item with any q-vector. For simplicity of notation, let us assume again that the first D^*_k attributes are required for item k. The GDI is defined as

$$\varsigma^2_{1:D^*_k} = \sum_{l=0}^{2^{D^*_k}} p(\boldsymbol{a}^*_l)\Big[P(X_k = 1|\boldsymbol{a}^*_l) - \bar{p}_k\Big]^2 \tag{7.16}$$

where $p(\boldsymbol{a}_l^*)$ is relative size of the reduced attribute pattern $\boldsymbol{a}_l^*$, and $\bar{p}_k$ is the mean success probability on item k. As can be seen from (7.16), the GDI is simply the variance of the success probabilities given a particular q-vector. For each item, $2^D - 1$ q-vectors can be specified, each corresponding to one GDI. De la Torre and Chiu (2016) defined a q-vector that results in the maximum ς_k^2 as an *appropriate* q-vector to item k. Of the appropriate q-vectors, the q-vector with the minimum number of attributes specified is deemed *correct*.

The GDI serves as the basis of the EM-based data-driven algorithm (de la Torre & Chiu, 2016) developed to validate the expert-based provisional Q-matrix. Compared to other data-driven Q-matrix validation methods that are designed for specific CDMs (e.g., the δ-method for the DINA model; de la Torre, 2008), the GDI is based on a general model so it can be used with any reduced CDMs the G-DINA model subsumes. In practice, the inequality established by de la Torre and Chiu (2016) may not hold due to potential misspecifications in the provisional Q-matrix as well as noise in the data. As a matter of fact, the maximum ς_k^2 is always achieved when $\boldsymbol{q}_k = \mathbf{1}$, which, more often than not, is an overspecification. To address this issue, they recommended examining the proportion of variance accounted for a particular q-vector relative to the maximum ς_k^2, and suggested selecting the simplest q-vector from a set of q-vectors with GDIs above a particular cutoff (e.g., $\varsigma^2 > 0.95$). Although it has been shown that the GDI-based procedure can be a reliable method of empirically validating a provisional Q-matrix, particularly when high quality items are involved, determining a single cutoff that is optimal across a variety of conditions remains a challenge. To minimize dependence on a single cutoff and to allow for quantitative and graphical information to be combined in determining the correct q-vector for an item, de la Torre and Ma (2016) proposed the use of the *mesa plot*. The mesa plot displays the GDIs of different q-vectors in ascending order. Instead of a single recommendation, a number of q-vectors in the vicinity where the plot plateaus or forms a tabletop are suggested from which the correct q-vector can be selected.

7.5.2 Cognitive Diagnosis Computerized Adaptive Testing

As in traditional IRT, computerized adaptive testing can also be used to improve test efficiency (i.e., shorter test or greater accuracy) in the CDM context by administering items that are tailored to an examinee's most current attribute estimate. However, due to the discrete and multidimensional nature of the attributes, the method for determining the optimal item in cognitive diagnosis computerized adaptive testing (CD-CAT) differs.

Kaplan, de la Torre, and Barrada (2015) used the GDI as an item selection index for CD-CAT. Specifically, for examinee n, the GDI for item k at time t (i.e., after t items have been administered) is computed as

$$\varsigma_k^{2(t)} = \sum_{l=0}^{2^{D_k^*}} p(\boldsymbol{a}_l \mid \boldsymbol{X}_n^{(t)}) \left[P(X_k = 1 | \boldsymbol{a}_l^*) - \bar{p}_{nk}^{(t)} \right]^2, \tag{7.17}$$

where $p(\boldsymbol{\alpha}_l \mid \boldsymbol{X}_n^{(t)})$ is posterior probability of $\boldsymbol{\alpha}_l^*$ at time t, $P(X_k = 1|\boldsymbol{\alpha}_l^*)$ is the time-invariant success probability on item k given $\boldsymbol{\alpha}_l^*$, and $\bar{p}_{nk}^{(t)}$ current overall item difficulty. Note that, as a CD-CAT item selection index, (7.17) is a function of $p(\boldsymbol{\alpha}_l \mid \boldsymbol{X}_n^{(t)})$, which changes over time. The item with the largest $\varsigma_k^{2(t)}$ is deemed most informative, and hence administered at time $t + 1$.

To examine the viability of the GDI as a CD-CAT item selection index, Kaplan et al. (2015) compared it with the posterior-weighted Kullback-Leibler (PWKL; Cheng, 2009) index, as well as the doubly-posterior-weighted modified PWKL (MPWKL) index, which they also introduced. They found that the GDI and MPWKL outperformed the PWKL when the reduced model is either the DINA or DINO model, but not when it is the A-CDM. In addition, although GDI and MPWKL performed similarly in terms of correct classification rate or average test length, the former was deemed more efficient in that it only required a fraction of the time to be implemented.

7.5.3 Item-Level Model Comparison

Given the variety of CDMs currently available, it is not obvious how the choice between these models can be made in practice. Previously, researchers assume a particular underlying process (e.g., conjunctive, additive) to fit a particular CDM (i.e., DINA model, R-RUM) to the data. With the availability of general models, fitting CDMs with less restrictive assumptions has been advocated. However, recent analyses of real data show that different items may require different types of CDMs, both reduced and saturated. These findings imply that a single reduced CDM would likely not provide a sufficient model-data fit. Moreover, even if a general model may provide an adequate fit assuming CDMs are appropriate, the parsimony principle (Beck, 1943) dictates that the simplest set of models that can provide equally good fit to the data be chosen. These findings also imply that using a test-level comparison using, say, Akaike (1973) or Bayesian information criterion, or the likelihood ratio test to choose *en masse* from among the CDMs that have been specified a priori may not lead to the selection of the optimal CDMs for the data.

To determine empirically (i.e., *post hoc*) the most appropriate CDM for each item, de la Torre (2011) developed an item-level model selection method using the Wald test. Assuming the Q-matrix has been validated, the Wald test can be used to determine whether one or more reduced CDMs can be used in place of the saturated CDM. For item k, the Wald statistic for comparing the reduced CDM ϱ against the saturated model is defined as

$$W_{k\varrho} = \left[\boldsymbol{R}_{k\varrho} \times g(\boldsymbol{P}_k)\right]' \left[\boldsymbol{R}_{k\varrho} \times Var[g(\boldsymbol{P}_k)] \times \boldsymbol{R}'_{k\varrho}\right] \left[\boldsymbol{R}_{k\varrho} \times g(\boldsymbol{P}_k)\right], \quad (7.18)$$

where $g(\boldsymbol{P}_k)$ is $g[P(X_k = 1 \mid \boldsymbol{\alpha}_{lk}^*)]$, $Var[g(\boldsymbol{P}_k)]$ is the corresponding variance matrix, and $\boldsymbol{R}_{k\varrho}$ is the restriction matrix associated with the reduced CDM ϱ. The

restriction matrix $\boldsymbol{R}_{k\varrho}$ is of size $(2^{D_k^*} - p) \times 2^{D_k^*}$, where p is the number of parameters in model ϱ. Below are examples of $\boldsymbol{R}$ for the (1) DINA model, (2) DINO model, and (3) additive models when $D_k^* = 2$:

$$\boldsymbol{R}^{(1)} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{bmatrix}, \quad \boldsymbol{R}^{(2)} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{R}^{(3)} = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}.$$

The Wald statistic $W_{k\varrho}$ is assumed to be asymptotically χ^2-distributed with $(2^{D_k^*} - p)$ degrees of freedom. It should be noted that using the Wald test for the purpose of evaluating the appropriateness of reduced CDMs is only meaningful when $D_k^* \geq 2$.

With a sufficiently large sample size and reasonable item quality, the Wald test has acceptable Type I error and power across various reduced models (de la Torre & Lee, 2013; Ma & de la Torre, 2016). Furthermore, in comparing the fit of CDMs selected via the Wald test against that of the G-DINA model, Ma, Iaconangelo, and de la Torre (2016) found using simulated and real data that the former provided higher correct classification rate than the latter, particularly when lower item qualities and smaller sample sizes are involved. More recently, de la Torre and Ma (2017) have shown that performing the Wald test is a necessary step to accurately evaluate whether or not a test can potentially identify all the possible attribute patterns. An evaluation of the expected item response profiles derived from fitting a saturated model without considering the appropriateness of reduced CDMs can lead to incorrect conclusions about the identifiability, or lack thereof, of the attribute patterns. Lastly, the use of the Wald test in the CDM context extends beyond item-level model comparison – it has also been used to evaluate differential item functioning (e.g., Hou & Terzi, 2017).

7.6 Discussion

This chapter presents the G-DINA model as framework for conducting analysis in the CDM context. As a general model and with appropriate link functions, the G-DINA model can be shown to subsume a number of familiar reduced CDMs in the literature. With it as the base model, the G-DINA model can be extended in various directions to address a wider range of practical testing situation needs. As a framework, the G-DINA model provides a coherent environment where CDM-related procedures can be developed and implemented. Thus far, the CDM-based methodologies that have been developed are largely applicable to CDM for dichotomous responses and attributes. To further improve the practicability of CDMs, these methodologies should be expanded to also apply to other CDM types.

The surge in the development of CDMs and related methodologies in recent years is without a doubt a positive development in this field. However, using these models and methodologies systematically and integratively can be daunting, particularly to many applied researchers. If any suggestions could be proffered regarding this

matter, they would be as follows. First, validate the Q-matrix specification. To do so without conflating Q-matrix misspecifications with potential CDM misspecifications, fit the G-DINA model. Second, check whether reduced CDMs can be used in place of the G-DINA model for items where $D_k^* \geq 2$. More likely than not, this would result in different items retaining different CDMs. Third, recalibrate the data using the CDMs selected in the previous step to update the estimates of the item parameters and attribute distributions. These are the estimates that one can use in estimating the examinees' attribute patterns. Optionally, in some applications, one can also consider whether the attribute distribution, which is typically estimated in saturated form (i.e., without constraints), can be simplified. An alternative is to specify the attribute distribution using a higher-order formulation (de la Torre & Douglas, 2004). As a final step, evaluate the absolute fit (i.e., goodness-of-fit) of the model to the data. One way this can be accomplished is by comparing the expected and observed moments, particularly the correlation and log-odds ratio, of each item pair (Chen, de la Torre, & Zhang, 2013; de la Torre & Douglas, 2008).

As a last word, we should be cognizant that, despite the numerous developments pertaining to CDM and related methodologies in the last two decades, these advances represent but one side of the coin. To fully take advantage of the potential of CDMs, we should also focus our attention on the other side of the same coin, which is developing cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014). In particular, we need to develop diagnostic assessments from the ground up using a CDM framework. On one hand, without the appropriate data, psychometric tools no matter how sophisticated cannot produce the rich information needed for precise diagnosis of student needs. On the other hand, without the appropriate psychometric tools, information no matter how rich cannot be properly extracted and utilized. Thus, for cognitive diagnosis modeling to break new ground in the near future, CDMs and CDAs must be used hand in hand.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Beck, L. W. (1943). The principle of parsimony in empirical science. *The Journal of Philosophy, 40*, 617–633.

Ben-Simon, A., Budescu, D. V., & Nevo, B. A. (1997). Comparative study of measures of partial knowledge in multiple choice tests. *Applied Psychological Measurement, 21*, 65–88.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*, 419–437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.
de la Torre, J. (2010, July). The partial-credit DINA model. Paper Presented at the International Meeting of the Psychometric Society, Athens, GA.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.
de la Torre, J. (2015, December). *Cognitively diagnostic assessment and cognitive diagnosis modeling: An example from start to finish*. Invited Presentation at the Global Chinese Conference on Educational Information and Assessment and Chinese Association of Psychological Testing Annual Conference, Taichung, Taiwan.
de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.
de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.
de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595–624.
de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373.
de la Torre, J., & Ma, W. (2016, August). *Cognitive diagnosis modeling: A general framework approach and its implementation in R*. A Short Course at the Fourth Conference on Statistical Methods in Psychometrics, Columbia University, New York.
de la Torre, J., & Ma, W. (2017, November). *Do* $\boldsymbol{I}$ *complete* $\boldsymbol{Q}$*?* Invited Presentation at the Fifth Conference on the Statistical Methods in Psychometrics, Department of Statistics, Columbia University, New York.
de la Torre, J., & Minchen, N. D. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*, 89–97.
de la Torre, J., & Minchen, N. D. (2016, May). *Modeling response time in cognitive diagnosis*. Invited Presentation at the Graduate Institute of Educational Measurement and Statistics Colloquium, National Taichung University of Education, Taiwan.
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.
Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis*. (Unpublished doctoral dissertation). Los Angeles: University of California.
Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation).
Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.
Hou, L., & Terzi, R. (2017, April). *Examining DIF in the proportional reasoning test using various Wald test formulations.* Paper Presented at the Meeting of National Council on Measurement in Education, San Antonio, TX.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.
Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188.
Ma, W. (2018). A diagnostic tree model for polytomous responses with multiple strategies. *British Journal of Mathematical and Statistical Psychology.* Advanced online publication. https://doi.org/10.1111/bmsp.12137
Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*, 253–275.
Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*, 200–217.
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Minchen, N. D., & de la Torre, J. (2018). A general cognitive diagnosis model for continuous-response data. *Measurement: Interdisciplinary research and perspective, 16*, 30–44.

Minchen, N. D., de la Torre, J., & Liu, Y. (2017). A cognitive diagnosis model for continuous response. *Journal of Educational and Behavioral Statistics, 42*, 651–677.

Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika, 79*, 647–674.

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement, 31*, 47–73.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203–219.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Templin, J. L., Rupp, A., Henson, R., Jang, E., & Ahmed, M. (2008, March). *Cognitive diagnosis models for nominal response data*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Tjoe, H., & de la Torre, J. (2013a). Designing cognitively-based proportional reasoning problems as an application of modern psychological measurement models. *Journal of Mathematics Education, 6*, 17–22.

Tjoe, H., & de la Torre, J. (2013b). The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Mathematics Education Research Journal, 26*, 237–255.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (Research Report No. RR-14-40). Princeton, NJ: Educational Testing Service. Retrieved from https://doi.org/10.1002/ets2.12043

Chapter 8
Loglinear Cognitive Diagnostic Model (LCDM)

Robert Henson and Jonathan L. Templin

Abstract The Log-Linear Cognitive Diagnosis Model (LCDM; Henson RA, Templin J, Willse J, Psychometrika 74:191–210, 2009) provides a general approach to diagnostic modeling that is deeply tied to log-linear models. As a result, the parameterization and concepts of the LCDM can be directly tied to the concepts of a general linear model for use of a multiple way ANOVA. By using these concepts, the LCDM provides a general framework that does not require the user to specifically determine the relationship between the attributes and the probability of a correct response. Furthermore, because of its flexibility, this chapter will show that the LCDM can be used to discuss similarities and differences between many common diagnostic models. This chapter will first provide a theoretical introduction to the motivation and the definition of the LCDM. Next, typical tools that have been used to estimate the LCDM and measures of fit are discussed. Finally, this chapter will discuss the relationship of other diagnostic models to the LCDM and, as a result, provide a succinct definition of what is meant by disjunctive, compensatory, and conjunctive models.

8.1 The Log-Linear Cognitive Diagnostic Model

Diagnostic classification models (DCMs; also known as cognitive diagnosis models (CDMs)) were becoming more popular in the early 2000's. Prior to this time some approaches such as the work on Rule Space (Tatsuoka, 1983), the Unified Model (DiBello, Stout, & Roussos, 1995), and Knowledge Space Theory (Doignon & Falmagne, 1999) had been published that focused on determining the state of a

R. Henson (✉)
Educational Research Methodology (ERM) Department, The University of North Carolina at Greensboro, Greensboro, NC, USA
e-mail: rahenson@uncg.edu

J. L. Templin
Educational Measurement and Statistics Program, University of Iowa, Iowa City, IA, USA

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_8

test taker's knowledge. As this field continues to expand, many of the diagnostic models define the probability of a correct response for the n^{th} person to the k^{th} item, $P(x_{nk} = 1)$, as a function of a set of item parameters and a test takers knowledge, an attribute profile. An attribute profile is an indicator vector of ability that is sometimes referred to as a mastery profile. More specifically, many diagnostic models define ability of a test taker as a D-dimensional vector $\boldsymbol{A} = (A_1, A_2, \ldots, A_D)$, where A_1, A_2, $\ldots$ A_D are thought to be indicators of mastery or non-mastery of a set of D attributes. Note that attributes have also be described or defined as specific skills or facets that influence performance on a test.

Most models do not assume that all attributes influence the performance of all items. For example, it is possible that there are two attributes that describe math ability such as basic addition (A_1) and basic subtraction (A_2). Thus, the item "$2 + 3 =$?" is expected to only be influenced by whether the test taker has mastered basic addition, A_1, whereas the item "$4-1 + 3 =$?" would require mastery of both basic addition and basic subtraction. The Q-matrix is used as an indicator matrix such that any element $q_{kd} = 1$ of the d^{th} attribute is measured by the k^{th} item and otherwise $q_{kd} = 0$. In this example, the vector for the Q-matrix related to the item "$2 + 3 =$?" would be $\boldsymbol{q}_k = (1, 0)$ indicating that only mastery/non-mastery of A_1 influences the probability of correctly answering that item. Whereas, $\boldsymbol{q}_k = (1, 1)$ would be specified for the second example item, "$4-1 + 3 =$?"

Given the attribute profile and the Q-matrix, the relationship between the attribute profile and the probability of a correct response varies based on the model used. Each model provides a specific functional form with model specific parameters. However, very few models considered the possibility of modeling the response patterns based on a log-linear model with latent variables. Specifically, dichotomous items and mastery/non-mastery on each of a set of D attributes can be thought of as creating a large contingency table. For example, if there are only two items and two attributes, then the contingency table would be a $2 \times 2 \times 2 \times 2$ table representing the two dichotomous items and the two dichotomous attributes. Furthermore, because all items are assumed to be independent given the examinees attributes (the assumption of conditional independence) one could consider a separate contingency table for each item crossed with the attributes measured by that item. For example, focusing only on the item, "$4-1 + 3 =$?", a three-way table could be created to represent the joint distribution of the k^{th} item with attributes A_1 and A_2 (basic math and basic subtraction respectively). Table 8.1 provides an example table that contains this joint distribution $P(X_K, A_1, A_2)$. Note that the sum across all of the cells is 1. In addition, it can be said that, across the total population, 20% of the examinees did not master either attribute and missed the item, $P(X_k = 0, A_1 = 0, A_1 = 0) = 0.20$.

This table shows that for test takers that missed this item, the attribute pattern $A_n = (0, 0)$ is most likely. Whereas of the attribute patterns that answered the item correctly, $A_n = (1, 1)$ is most likely. Note that these probabilities could also be expressed in terms of the probability of a correct response given a specific attribute pattern. For example, given that a test taker has the attribute pattern

Table 8.1 Example of the joint distribution of response and attributes

	$X_k = 0$			$X_k = 1$	
	$A_2 = 0$	$A_2 = 1$		$A_2 = 0$	$A_2 = 1$
$A_1 = 0$	0.200	0.125	$A_1 = 0$	0.050	0.125
$A_1 = 1$	0.150	0.025	$A_1 = 1$	0.100	0.225

$A_n = (0, 0)$, the probability of a correct response is $\frac{0.05}{0.05+0.20} = 0.20$, which is equal to $(X_k = 1, A_1 = 0, A_1 = 0)/[(X_k = 0, A_1 = 0, A_1 = 0) + (X_k = 1, A_1 = 0, A_1 = 0)]$. Furthermore, if the attributes were observed quantities, this would represent a contingency table, where one could use the log-linear model as a diagnostic model that predicts the probability of any combination of X_k, A_1, and A_2 (Henson et al., 2009).

In this setting, the log-linear model would be used to model the joint distribution of X_k, A_1, and A_2 (i.e., Table 8.1) for each k. Using the example item in Table 8.1, this log-linear model would include main effects for each variable (λ_{X_K}, λ_{A_1}, and λ_{A_2}, respectively), in addition to all 2-way interactions ($\lambda_{X_K A1}$, $\lambda_{X_k A_2}$, and $\lambda_{A_1 A_2}$) and a three way interaction ($\lambda_{X_k A_1 A_2}$). For a more thorough description of the log-linear model see Agresti (2003). Of particular importance are effects that only describe the attributes. For example, the main effect λ_{A_1} is related to a shift or change in the joint distribution when considering $A_1 = 1$ versus the reference category $A_1 = 0$, whereas the value $\lambda_{A_1 A_2}$describes the shift of joint probability above and beyond what would be expected based on the marginal when both $A_1 = 1$ and $A_2 = 1$. Finally, in this example, the three-way interaction expresses the fact that when computing the joint probability there is an effect for the specific combination of $X_k = 1$, $A_1 = 1$, and $A_2 = 1$ that is above and beyond what could be exampled by the marginal and two-way interactions. This specific parameterization assumes that non-mastery represents a reference category for each attribute.

Although the log-linear model could be used to predict the joint probability, $P(X_k, A_1, A_2)$, in diagnostic models, the focus is not typically on this joint distribution, but instead it is on the conditional distribution of the item given the attribute profile. Furthermore, because it is assumed that the X_k are dichotomous 0/1 items (i.e., right/wrong scoring), the conditional distribution can be specified in terms of the log-odds of the probability that $X_k = 1$. Thus, the Log-Linear Model for Cognitive Diagnosis Models (LCDM; Henson et al., 2009) re-expresses the log-linear model in terms of the log-odds of correctly responding (Agresti, 2003). Note that in doing this, all effects with respect to X_k are not included because the conditional distribution of the item is modeled as the log-odds of X_k. Thus, in this example, the LCDM is defined as

$$\ln\left(\frac{P\left(X_k = 1|A_n\right)}{P\left(X_k = 0|A_n\right)}\right) = \lambda_0 + \lambda_{A_1} + \lambda_{A_2} + \lambda_{A_1 A_2}. \tag{8.1}$$

Here, the value λ_0 represents the log-odds of a correct response for the reference category, which are test takers who have not mastered either of the attributes (i.e.,

$A_1 = 0$ and $A_2 = 0$). The main effects, λ_{A_1} and λ_{A_2}, represent the marginal change in the log-odds when that specific attribute is mastered. Finally, the interaction term, $\lambda_{A_1A_2}$ represents the additional change in the log-odds when both of the attributes have been mastered. As described in Henson et al. (2009) and in more detail in Rupp, Templin, and Henson (2010), the model parameters are subject to monotonicity constraints that specify that the probability of correct response increases monotonically with the number of attributes mastered. Without such constraints, the LCDM is incomplete as the specification of constraints aids in estimation and attribute interpretation. In addition to the typical assumptions of a log-linear model, conditional independence of the item responses is assumed given an examinees attribute mastery profile.

By defining the log-odds of a correct response in this way, the LCDM is directly related to traditional multidimensional models that assume continuous traits. In addition, the LCDM is related to two other models that were defined previously in the literature, the Compensatory Reparameterized Unified Model (Compensatory RUM; Hartz, 2001) and the General Diagnostic Model (GDM; von Davier, 2008). Although both the Compensatory RUM and the GDM where defined from different perspectives, both where expressed as linear models with a logit link for the probability of a correct response (the GDM also models polytomous responses and polytomous attributes). Both the Compensatory RUM and the GDM had a primary focus on the main effects. However, given the general definition of the GDM, it can be shown that the LCDM is a special case of the GDM that emphasizes the interactions. Specifically, the GDM defines the probability for a dichotomous response as

$$P\left(X_{nk} = 1 | A_n\right) = \frac{e^{\boldsymbol{\lambda}^T \boldsymbol{h}(A_n, q_k)}}{1 + e^{\boldsymbol{\lambda}^T \boldsymbol{h}(A_n, q_k)}} \tag{8.2}$$

where the LCDM specifically defines the value

$$\boldsymbol{\lambda}^T \boldsymbol{h}\left(A_n, q_k\right) = \lambda_0 + \sum_{d=1}^{D} \lambda_d A_d q_{kd} + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \lambda_{A_i A_j} A_{ni} A_{nj} q_{ki} q_{kj} \ldots \tag{8.3}$$

Thus, the LCDM defines the log-odds as a linear function that includes all attribute main effects and all possible interactions. Often, the model is expressed as in Eq. 8.2, as the probability of a correct response as opposed to modeling the log-odds. In addition, for those attributes that are not measured by the item, $q_{kd} = 0$ and therefore those effects drop out and are not estimated.

The presentation of the LCDM in this chapter demonstrates the relationship between diagnostic models and the log-linear model. In addition, because the relationship of the log-linear model to the logistic model, we ultimately express the LCDM as a generalized linear model where a linear relationship between the attributes and their potential interactions and the log-odds of a correct response is defined. Alternative parameterizations were later explored and named the GDINA

by de la Tore (2011) by modeling different links of the probability of a correct response as opposed to the log-odds (e.g., the identity link or the log link). The general fit of the GDINA should be identical to the LCDM although the interpretation of the parameters is different depending on the link used. That said, the interpretation will be the same when the log-odds link is used with the GDINA.

In defining the LCDM in this way, the LCDM is a general model that subsumes many of the diagnostic models that have been discussed in the literature. However, there may be concerns related to the number of parameters that would need to be estimated. For example, if an item measured four attributes, the LCDM would define the probability of a correct response as a function of 16 parameters, which correspond to an intercept, 4 main effects (one for each attribute), 6 two-way interactions, 4 three-way interactions, and a four-way interaction. As a result, an approach similar to what is used in an ANOVA or other linear models may be applied where it is desirable to reduce the model by eliminating the higher-order interactions. Note that reducing the higher order interactions naturally moves the LCDM toward a compensatory model, which may be reasonable. However it is also possible to reduce the model by eliminating lower order effect or to place constraints on the model such that it is equivalent to many of the models defined in the literature. Submodels of the LCDM will be discussed later in this chapter.

8.2 LCDM Estimation

In the original LCDM paper, Henson et al. (2009) estimated the LCDM through use of a Metropolis-Hastings within Gibbs MCMC algorithm. As outlined in the paper, uniform priors for all item parameters were assumed and constraints of monotonicity were defined. In addition, a hierarchical Bayes approach was described such that the attribute space was defined to be the result of a dichotomized multivariate normal distribution and thus a "cut point" was estimated for each attribute representing the portion of masters for each attribute. In addition, the correlation matrix, which represented the tetrachoric correlation between attributes was estimated using a common factor model. However, a number of possible methods could be used to model the attribute space as described by Rupp et al. (2010).

Shortly thereafter, marginal maximum likelihood algorithms were also used for the LCDM, starting in Mplus, which is described in Chap. 28 (see also Rupp et al., 2010; Templin & Hoffman, 2013), and subsequently in several R packages. It should also be noted that because the LCDM is a special case of the GDM and because the GDINA builds on the concepts of the LCDM, software that fits both the GDM and GDINA could also be used to estimate the LCDM. The specific estimation of the GDM and GDINA are discussed in the Chaps. 6 and 7 of this handbook. Regardless of the type of estimation approach, a primary consideration of any algorithm is the inclusion of the LCDM monotonicity constraints. Without such constraints, the LCDM becomes a very general latent class model, and is subject to

optimality issues prevalent in latent class models, primarily a multi-mode likelihood surface and nominally-defined classes that may switch meaning during the analysis. Both of which are demonstrated to occur when using estimation procedures without constraints (for example see Lao, 2016). At the time of this writing, two common packages that allow the estimation of the LDCM, the CDM package in R (George, Robitzch, Kiefer, Gross, & Uenlue, 2016) and flexMIRT (Cai, 2017) do not impose constraints and therefore may suffer from inaccurate results. The user should be aware of this potential limitation in that results in higher dimensional may be difficult to interpret, which would be true even if alternative software packages are used. In addition to the monotonicity constraints it is possible that very complex Q-matrices may result in a non-identified solution. The identifiability conditions of diagnostic models are discussed to some extent in Chap. 16 of the volume.

The other big consideration that has largely driven the choice of an estimation framework is estimation time. Marginal maximum likelihood algorithms often see algorithm completion times that increase exponentially as the number of attributes in an analysis increases linearly. Most MCMC algorithms have completion times that increase linearly in the number of attributes. That said, most Q-matrices with five or fewer attributes tend to be faster using MML whereas MCMC is faster for larger Q-matrices.

8.3 Evaluation of Model Fit

The evaluation of model fit for the LCDM is not generally different than the evaluation of model fit for any other latent variable model, particularly for models for categorical data. Two broad classes of model fit exist: (1) absolute fit – where a model is compared against properties of the data used for estimation and (2) relative fit – where competing models are compared against each other. In general, for estimated models to be considered as candidates for use, each must achieve a reasonable level of absolute fit prior to being compared for relative fit. Many methods for model fit have been developed, stemming from multiple disciplines including factor analysis/structural equation modeling, item response theory, and computer science, including machine learning. As such, the treatment given in this chapter is appropriately incomplete. Here, we focus on several methods for absolute and relative fit that have been used in empirical research with the LCDM. We further restrict our conversation to cases where the LCDM is used with binary items (i.e., scored dichotomously with two values: 0 and 1). For more information on broader methods of model fit, we refer the reader to the model fit chapter of Rupp et al. (2010) for a broader set of references.

8.3.1 Absolute Fit

Absolute fit methods involve a comparison of estimated model predictions with that of the data set used in the estimation process. The most general method for absolute model fit is the comparison of the distribution of observed item response patterns in the data with the distribution expected by the estimated LCDM. With $k = 1, \ldots, K$ binary items, a total of 2^K response patterns are possible. The model-based expected frequency is given by the LCDM likelihood function:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{c=1}^{C} \eta_e \prod_{k=1}^{K} \pi_{kc}^{x_k} (1 - \pi_{kc})^{1-x_k}. \tag{8.4}$$

Here, the summation is across the $c = 1, \ldots, C$ attribute mastery profiles (latent classes), marginalizing the conditional item response likelihood function (provided by the product term, which is in place due to the assumption of conditional independence of items given attribute profile). The η_c parameter provides the proportion of people with attribute profile c from the structural model. The π_{kc} parameter is the LCDM item response function for item k for a person with attribute profile c. The model-based expected probability can be found for each of the 2^I response patterns, leading to the expected number of examinees with that response pattern when multiplied by the number of people in the sample. From here, a Pearson or likelihood χ^2 value can be obtained through customary methods of comparing expected versus observed responses. The degrees of freedom for the test statistic is given by the number of response profiles minus the number of model parameters minus one. If the test statistic is significant (at a tolerable level of Type-I error), then the expected distribution of the data does not match the observed distribution of the data, which indicates the model does not fit absolutely. One concern of such a test is that, in a reasonably large sample size, the power may be too high and one would determine the model does not fit even when the model does an reasonable job of describing the data. Because these models are sometimes estimated using Bayesian approaches, it might also be reasonable to use posterior predictive checks, which are not the focus of this chapter.

In principle, the global absolute fit test using response patterns is ideal; however, in practice, few situations exist where the test can be used. The reason the test is limited comes from the use of the χ^2 test, which is invalid when there are response patterns with zero observed examinees present. As the number of items increases, the number of possible response patterns increases exponentially, and with that exponential increase comes an increasingly high chance of observing many response patterns with zero examinees, particularly for tests with more than eight or nine items (with a total of 256 and 512 response patterns, respectively). In such situations, limited information goodness-of-fit measures are available for use (e.g., Maydeu-Olivares & Joe, 2005). Limited information fit measures examine absolute model fit to contingency tables with smaller sets of items (such as all item pairs), which are much more likely to have non-zero counts of examinees observed. The measures

also create comparisons where if a test statistic is significant, the model is said to not fit the data. The downside of these measures is that if a model fit measure is not significant, leading one to believe a model fits the data, then it is still possible for the model to not fit the data absolutely if a fit to a higher-level table (e.g., three-way) is poor. Such measures are increasingly available for psychometric models with categorical data and exist in software such as the *mirt* package in R (Chalmers, 2012) and flexMIRT (Cai, 2017), the latter of which allows for such measures to be calculated with the LCDM.

Although absolute fit measures indicate model fit with an overall test across all items, more localized versions of fit measures are available and are useful in considering how to modify an ill-fitting model. One easily attainable measure of localized model fit is the comparison of expected and observed counts of examinees for pairs of items. Each pair of items (across the $\binom{K}{2}$ possible combinations) are the atomic portions that together constitute the limited-information goodness-of-fit to two-way tables, but can be used individually to investigate why a model is not fitting. Such tables are obtainable in Mplus (Muthén & Muthén, 2017) using the TECH10 option and in the *CDM* package in R (George et al., 2016; although at the time of writing this chapter, the latter has suspect estimation results due to a lack of monotonicity constraints). If an item consistently shows up in mis-fitting pairs, that item may not be modeled correctly (missing Q-matrix entries or columns) or it may be a poorly worded or understood item.

8.3.2 Relative Fit

Relative fit measures for the LCDM also follow relative fit measures used in conventional psychometric models. Relative model fit measures compare the model fit of two or more candidate models, often contrasting the quality of fit of a model with the number of parameters in the model. It is worth reiterating, however, some conditions necessary for relative model fit. First, all candidate models must approximately fit the data based on some standard (a standard that seems to be rarely applied in practice). Second, the candidate models must also be estimated using the exact same data (the same examinees taking the same items). Third, if one model is nested within another (meaning the model with fewer parameters is achieved through a set of parameter constraints placed on the model with more parameters), then the model with fewer parameters can never achieve absolute fit better than the model with more parameters and at best can only fit as well as the model with more parameters (in which case the model with fewer parameters is more parsimonious). Should these conditions be satisfied, then model comparisons using relative model fit measures can be conducted.

For models estimated using marginal maximum likelihood, relative fit measures often involve likelihood ratio tests (for pairs of models where one is nested within another) and, for non-nested models, the calculation and comparison of information

criteria such as (but not limited to) the Akaike and Bayesian information criteria (AIC and BIC, respectively). Examples of reduced nested models are specifically discussed in the next section. Thus, it will be shown that popular models such as the DINA are a nested model of the LCDM and so a likelihood ration test could be used to compare these two models. The likelihood ratio tests provide a *p*-value that, when below a pre-specified level of Type-I error, indicates that the model with fewer parameters *does not fit as well* as the model with more parameters, or that the model with more parameters should be chosen. For information criteria, the selection is typically done by choosing the model with the lowest value of a given criterion, although what happens when different models are selected when different criteria are used is subject to debate. In models estimated through Bayesian methods, Bayes Factors may be able to be constructed and Bayesian versions of information criteria, such as the Deviance Information Criterion (DIC) exist as well.

8.4 Submodels of the LCDM

The LCDM is among a few diagnostic models that are considered general models. Diagnostic models define the probability of a correct response based on mastery or non-mastery of a set attributes. Assume an exam is designed to measure for attributes. If an item in that exam measures the first two attributes, then a diagnostic model predicts the probability of a correct response for four different types of test takers. The four different types of test takers are (i) a test taker who has mastered both attributes, $\boldsymbol{A} = (1, 1, {}_*, {}_*)$, (ii) a test taker who has mastered only the first, but not the second, $\boldsymbol{A} = (1, 0, {}_*, {}_*)$, (iii) a test taker who has mastered the second, but not the first, $\boldsymbol{A} = (0, 1, {}_*, {}_*)$, and (iv) a test taker who has not mastered either attribute, $\boldsymbol{A} = (0, 0, {}_*, {}_*)$. The "*" indicates that the mastery of that attribute does not matter and thus could be either a master or a non-master. A saturated model is a model that has as many parameters as the number of classes in the model, which would be $2^4 = 16$ in this example. Although not completely saturated, the LCDM defines as many possible parameters estimated as there are unique attribute patterns for each item, which is based on the attributes measured by each item. In this example, the LCDM would have an intercept, two main effects, and a single interaction. Although, the complexity of such a model can be a possible detriment (e.g., it may require large sample sizes), the advantage of complex models is that they provide a framework that can be used to discuss other models in the literature in addition to defining specific concepts of diagnostic models. Specifically, the LCDM can be used to define previous models in the literature and, in doing so, a better understanding of how they are similar or different can be provided. For example, the LCDM provides a more specific definition of what should be meant by disjunctive, compensatory, and conjunctive. Note that these terms are not always used consistently in the literature. First, an example of a compensatory model that is similar to the traditional compensatory IRT model is discussed. Then examples of disjunctive and conjunctive models are provided. It should also be noted that while

the sub-models presented are fairly common in the literature, they are usually an over simplification and, thus, the LCDM with possibility lower order interactions would be more appropriate.

One of the most natural reductions of the LCDM is the compensatory RUM (Hartz, 2001), which is also commonly discussed in the context of the GDM (von Davier, 2008). Using the compensatory RUM the log-odds of the probability of a correct response is defined as a function of an item specific intercept and a main effect for each attribute measured by that item. Because no interactions are defined for this model, the Compensatory RUM is the model most similar to a traditional multidimensional IRT model. The Compensatory RUM is defined, such that

$$\boldsymbol{\lambda}^T \boldsymbol{h}\left(A_n, q_k\right) = \lambda_0 + \sum_{d=1}^{D} \lambda_d A_d q_{kd} \tag{8.5}$$

where λ_0 is related to the probability of a correct response for non-masters of all measured attributes and each λ_d represents the change in the log-odds when a specific attribute is mastered. Often the sum of a set of latent abilities is referred to as a compensatory model because of the ability to compensate or "make up" for what is lacked in one ability by having even more of another ability. Although the compensatory RUM does have a summation of effects, the idea of making-up even more for an attribute is complicated by the fact that often attributes are only dichotomous. As will be discussed later in this chapter, using the LCDM a more consistent definition of what is meant to be compensatory model will be provided.

Although the Compensatory RUM has been provided as an example of a compensatory model in which an examinee can "make up" for what is lacked in one attribute by mastering another, the best example of a fully compensatory model is the Deterministic Input Noisy "Or" gate model (DINO; Templin & Henson, 2006), which is also referred to as a disjunctive model. For any given item, the DINO model divides test takers in to two different groups. The first group represents those test takers who have not mastered any of the measured attributes by that item. The second group has mastered at least one of the attributes measured by that item. The DINO model then defines the probability of getting the item correct for each of the two groups. Specifically, the probability of a correct response for the group that has not mastered any of the measured attributes is equal to the probability that they "guess" the correct answer, which is indicated by the parameter g_k. Whereas the probability of a correct response for the group that has mastered at least one attribute is equal to one minus the probability that an examinee "slips" up, s_k, even though he or she should have correctly responded to the item.

Because of how the two groups are defined, an examinee can make up for lacking a given attribute by having mastered another attribute and thus the DINO does provide the best example of compensation for diagnostic models. That is, an examinee must only master one of the measured attributes to be considered a master for that item and, as a result, have a high probability of correctly responding to the item. It is also true that the DINO does not distinguish as to which attributes have been mastered (i.e., all attributes have equal weights).

Table 8.2 The probabilities defined using the DINO and LCDM parameterizations

Attribute Profile	DINO	LCDM
$\boldsymbol{A} = \{0, 0\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = g_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0}}{1+e^{\lambda_0}}$
$\boldsymbol{A} = \{1, 0\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = 1 - s_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0+\lambda_A}}{1+e^{\lambda_0+\lambda_A}}$
$\boldsymbol{A} = \{0, 1\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = 1 - s_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0+\lambda_A}}{1+e^{\lambda_0+\lambda_A}}$
$\boldsymbol{A} = \{1, 1\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = 1 - s_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0+\lambda_A}}{1+e^{\lambda_0+\lambda_A}}$

The LCDM can model the DINO by imposing a set of constraints of the estimated weights. For example, if an item measured only two attributes A_1 and A_2 then the LCDM would be defined using eq. (8.2) where

$$\boldsymbol{\lambda}^T \boldsymbol{h}\left(A_n, q_k\right) = \lambda_0 + \lambda_A A_1 + \lambda_A A_2 + (-1)\,\lambda_A A_1 A_2 \tag{8.6}$$

Notice that Eq. 8.6 defines all main effects as being equal to a single value λ_A. Furthermore, the interaction is defined as the negative of any of the main effects (all main effects are equal). Table 8.2 provides the probabilities for all possible combinations of mastery for these two attributes.

Note in Table 8.2 that ultimately both models, the DINO and the LCDM, use only two parameters. As will be emphasized later, disjunctive models will have negative two-way interactions, which is in contrast to the compensatory RUM where all interactions were equal to zero.

The last submodel that will be discussed is the Deterministic Input; Noisy "And" gate model (DINA; Junker & Sijtsma, 2001). This model is commonly referred to as a non-compensatory or conjunctive model because a test taker cannot make up for lacking an attribute by having mastered another attribute. Like the DINO, the DINA model divides people in two groups for each item. However, the DINA assumes that test takers who have not mastered *at least* one of the measured attributes will be in a group. Whereas test takers who have mastered *all* measured attributes will be in another group. Given the two groups, like the DINO, a slip and guess parameter are used to define the probability of a correct response. Where the profile mastering all measured attributes by that item has a probability of a correct response equal to one minus the probability of slipping and all other profiles (those who lack at least one of the measured attributes) have a probability of a correct response equal to the probability of guessing. To model the DINA using the LCDM all main effects and lower order interaction terms must be fixed to zero. Thus, only the intercept and the highest order interaction are estimated. Furthermore, that interaction must be positive. Table 8.3 provides the probability of a correct response for the same theoretical item, which measures only two attributes) using both the DINA parameterization and the LCDM.

Table 8.3 The probabilities defined using the DINA and LCDM parameterizations

Attribute profile	DINA	LCDM
$\boldsymbol{A} = \{0, 0\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = g_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0}}{1+e^{\lambda_0}}$
$\boldsymbol{A} = \{1, 0\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = g_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0}}{1+e^{\lambda_0}}$
$\boldsymbol{A} = \{0, 1\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = g_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0}}{1+e^{\lambda_0}}$
$\boldsymbol{A} = \{1, 1\}$	$P(X_{nk} = 1 \mid \boldsymbol{A}_n) = 1 - s_k$	$P\left(X_{nk} = 1 \mid \boldsymbol{A}_n\right) = \frac{e^{\lambda_0+\lambda_{A_1A_2}}}{1+e^{\lambda_0+\lambda_{A_1A_2}}}$

Notice that again, when constrained, the LCDM has as many estimated parameters as the DINA. In addition, as will be discussed next, because the interaction is positive this will be referred to as a conjunctive model.

8.4.1 Disjunctive, Compensatory, and Conjunctive Models

The LCDM provides a flexible framework such that, when fully estimated, the specific nature of the relationship between the attributes and the probability of a correct response does not have to be defined a priori. Prior to the LCDM, researchers would usually select the model and thus assume the relationship between the attributes and the response. For example, a researcher may select the DINA (assuming a conjunctive model) or the DINO (assuming a disjunctive model). Using the LCDM one can determine the nature of the relationship between the attributes and a correct response based on the interaction. Furthermore, it is possible that the estimated item parameters can be reduced or constrained to less complex models. There may even be instances where the relatively extreme models such as the DINO or DINA or any other sub-model are appropriate. For a more complete description see Henson et al. (2009).

Given the specific parameterization of the LCDM it is possible to provide a more complete definition of what is meant when terms such as compensatory, disjunctive or conjunctive are used. The following discussion will be given in the context of an item that measures only two attributes, however these concepts can be extended to items measuring more than two attributes.

For a more complete definition what is meant by disjunctive, compensatory, and conjunctive using the LCDM, one must focus on how the LCDM is reduced to fit the sub-models discussed in the chapter, which represent extreme examples of disjunctive and conjunctive models. Table 8.4 provides a summary of each of the three models discussed previously, the type of model, and the needed constraints for the LCDM to fit these models.

Although only three specific examples are provided, in many ways the distinction of disjunctive, compensatory, and conjunctive models can be thought of more along a continuum. Specifically, the DINO can be thought of as the "most" disjunctive

Table 8.4 Description of the Models, Type of relationship between attributes and response and the necessary constraints for the LCDM

Model	Type	LCDM constraints
DINO	Disjunctive	Main effects are estimated
		Negative interaction
Compensatory RUM	Compensatory	Main effects are estimated and must be positive
		No interactions
DINA	Conjunctive	Main effects are fixed to 0
		Only the highest order interaction is estimated and must be positive

model in which the main effects are large and positive and the interaction is as large as the main effects and negative. A model becomes less disjunctive as the interaction becomes smaller in magnitude, but is still negative, and the main effects become slightly smaller. Models should be considered disjunctive until the interaction term becomes zero. At that point, a model with no interaction (the weight associated with the interaction term is equal to zero) and smaller positive main effects should be considered a compensatory model. Finally, as the interaction becomes positive and the main effects continue to approach zero the model should be considered conjunctive. Notice that this continuum ends when the interaction term is large and positive and all main effects are zero, which results in the DINA model.

To further describe the continuum of models ranging from disjunctive to conjunctive the two-attribute example will be expanded. Using two attributes, the log-odds of the probability of a correct response can be written out as is shown in Eq. 8.7,

$$\ln\left(\frac{P\left(X_k=1|A_n\right)}{P\left(X_k=0|A_n\right)}\right)=\lambda_0+\lambda_{A_1}A_1+\lambda_{A_2}A_2+\lambda_{A_1A_2}A_1A_2. \tag{8.7}$$

In addition, Eq. 8.7 can be rewritten to focus on the effect of the first attribute, A_1,

$$\ln\left(\frac{P\left(X_k=1|A_n\right)}{P\left(X_k=0|A_n\right)}\right)=\lambda_0+\left(\lambda_{A_1}+\lambda_{A_1A_2}A_2\right)A_1+\lambda_{A_2}A_2. \tag{8.8}$$

Equation 8.8 shows that the effect of a test taker moving from non-mastery of Attribute 1 ($A_1=0$) to mastery of Attribute 1 (i.e., $A_1=1$) is equal to $\lambda_{A_1}+\lambda_{A_1A_2}A_2$. Note that this effect could also be described as the difference in the log-odds of the probability of a correct response when comparing masters to non-masters of Attribute 1. This effect depends on the main effect of Attribute 1, which is always nonnegative and on the interaction. The dependency on the interaction term means that the actual effect of becoming a master for Attribute 1 also depends on whether Attribute 2 has or has not been mastered. Although this concept can become more complex when an item measures more than two attributes, the point can be made without the loss of generality using only two.

Recall that in Table 8.3 this interaction is equal to a large negative number for the DINO model (a disjunctive model), zero for the compensatory RUM (a

compensatory model), and a larger positive value for the DINA (a conjunctive model). Using these three models as examples along the continuum of possible models, a general more accurate definition of disjunctive, compensatory, and conjunctive models can be provided.

The DINO model as a disjunctive model is defined in such a way that the two-way interaction is negative and equal to the negative of all main effects. Thus, the effect of becoming a master of Attribute 1 for a test taker is equal to $\lambda_{A_1} - \lambda_{A_1} A_2$. Because the value A_2 can only be 0 or 1, this change or difference is either 0, which is when $A_2 = 1$ or equal to λ_{A_1} when $A_2 = 0$. Thus, in the DINO model, the effect of mastering Attribute 1 is largest when A_2 *has not* been mastered.

The effect of becoming a master of Attribute 1 when assuming the compensatory RUM is slightly different. Specifically, all interactions of the LCDM are equal to zero when modeling the compensatory RUM. Thus, the effect of mastering Attribute 1 is equal to $\lambda_{A_1} + (0)A_2$. Because A_2 is always multiplied by 0 (the interaction), this effect is always the same value. When using a compensatory RUM the effect of becoming a master is always equal to the main effect and does not depend on whether any other attributes have been mastered.

Finally, the DINA model, when specified by the LCDM has main effects equal to zero and the highest order interaction is large and positive. In this example, using only 2 attributes the effect of a test taker mastering Attribute 1 is equal to $(0) + \lambda_{A_1 A_2} A_2$. Again, because A_2 is either equal to 0 or 1, the effect of mastering Attribute 1 is either equal to 0 (when $A_2 = 0$) or equal to a large positive value when $A_2 = 1$. Thus, the effect of mastering Attribute 1 is only nonzero when all other measured attributes have been mastered.

The LCDM allows for a direct comparison between disjunctive, compensatory, and conjunctive models based on the estimation of the interaction terms. As opposed to using "loose" terminology that describes whether an individual can make up for what is lacked in one attribute based on mastery of another attribute, which is particularly difficult for dichotomous attributes. Because of the interaction, one can see that in disjunctive models the effect of mastering a given attribute is maximized only when all other attributes measured by that item have not be mastered. This effect is smaller as additional attributes measured by that item have been mastered. In contrast, when considering conjunctive models, the effect of mastering an attribute is maximized when all other attributes measured by that item have been mastered. As these attributes are not mastered, the effect of mastering a given attribute are reduced. Finally, in a compensatory model the effect of mastering a given attribute is constant, regardless of mastery or non-mastery of all other attributes.

8.5 Summary

In this chapter, the general motivation of the LCDM was discussed to provide direct ties to basic concepts of linear models and more specifically log-linear models and the use of interactions. Then the LCDM was specifically defined and available

methods of estimation were discussed. Given the LCDM and its estimation, a summary of methods used to measure data fit, which could also be used for model reduction where provided. Finally, using the LCDM many models that have been previously defined in the literature were discussed as a tool to present a more succinct definition of what is meant by disjunctive, compensatory, and conjunctive models.

References

Agresti, A. (2003). *Categorical data analysis*. Hoboken, NJ: John Wiley and Son.

Cai, L. (2017). *flexMIRT version 3.51: Flexible multilevel multidimensional item analysis and test scoring [computer software]*. Chapel Hill, NC: Vector Psychometric Group.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179. https://doi.org/10.1007/s11336-011-9207-7

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Mahwah, NJ: Erlbaum.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin, Germany: Springer, Verlag.

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

Hartz, S. M. (2001) *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality* (Ph. D. thesis, University of Illinois at Urbana-Champaign).

Henson, R. A., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Lao, H. (2016). *Estimation of diagnostic classification models without constraints: Issues with class label switching* (Doctoral dissertation, University of Kansas).

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009–1020.

Muthén, L., & Muthén, B. O. (2017). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guildford Press.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *The British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

Chapter 9
Diagnostic Modeling of Skill Hierarchies and Cognitive Processes with MLTM-D

Susan E. Embretson

Abstract This chapter formally describes the multicomponent latent trait model for diagnosis (MLTM-D; Embretson S.E., Yang X, Psychometrika 78:14–36, 2013) and then provides examples of applications to diagnose broad and narrow skills, as well as measure processing complexity and attainment. MLTM-D can be applied to diagnose either skill mastery or cognitive processing capabilities of examinees. MLTM-D is readily applicable to diagnose hierarchically-structured skills or to assess cognitive processes with postulated sources of complexity. That is, MLTM-D is a multidimensional conjunctive model for item responses that are impacted by varying underlying components with specifiable sources of complexity. MLTM-D can be applied to assess both processing competencies of examinees and the impact of the postulated features on process difficulty.

9.1 Introduction

Successful problem solving on complex tests, such as mathematical or reading achievement tests, depends not only on both broad and narrow skills, but also on the cognitive processes that are involved in item solving. The multicomponent latent trait model for diagnosis (MLTM-D; Embretson & Yang, 2013) can be applied to diagnose either skill mastery or cognitive processing capabilities of examinees. Several diagnostic models are applicable to complex combinations of skills or attributes in items (e.g., Henson, Templin, & Willse, 2009; von Davier, 2005, 2008). MLTM-D, however, is readily applicable to diagnose hierarchically-structured skills or to assess cognitive processes with postulated sources of complexity. That is, MLTM-D is a multidimensional conjunctive model for item responses that are impacted by varying underlying components with specifiable sources of complexity.

S. E. Embretson (✉)
School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA
e-mail: susan.embretson@psych.gatech.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_9

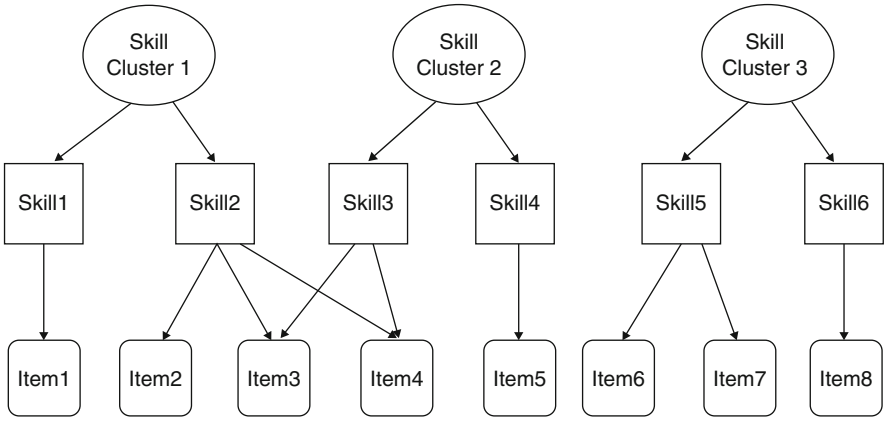


Fig. 9.1 Involvement of hierarchically organized skills in items

Blueprints for achievement tests often distinguish between two or more areas or skill clusters. For example, the blueprints for mathematical achievement tests typically distinguish between four or five general areas. The National Assessment for Educational Progress (NAGB, 2017) has test content organized into five areas; Number, Algebra, Geometry, Measurement and Data skills. Diagnosis of mastery at this level can have implications for remedial classes available to examinees. Consider the skill hierarchy shown on Fig. 9.1. The skill clusters represent three domains of skills that differ in content. That is, students with non-mastery only for Skill Cluster 2 would not need remedial instruction for Skill Cluster 1 or Skill Cluster 3. Under each skill cluster are more narrowly defined skills. Diagnosis of mastery at this level would have implications for specific remedial instructional units. MLTM-D can provide diagnosis at both levels.

Cognitive processing requirements also can vary substantially between items on global tests of aptitude or achievement. Consider the four cognitive processes that are postulated to be involved in mathematics items (Mayer, 2003) as follows: (1) Translation, encoding the meaning of the words and terms in the item, (2) Integration, bringing together the encoded aspects of the problem into equations to be solved, (3) Solution Planning, developing a strategy to solve for the unknowns and (4) Solution Execution, finding numerical solutions for the unknowns. Figure 9.2 shows mathematical problems that vary in process involvement, which are assumed to be executed sequentially. Notice that although Item 1 and Item 2 involve all four processes, the other items involve only one or two processes. For some multiple-choice items, a fifth stage (Decision) could be added to the model (see Daniel & Embretson, 2010; Embretson & Daniel, 2008).

Underlying the difficulty of the processes in specific items are postulated content features that impact cognitive complexity. For example, Translation becomes more difficult as vocabulary level, number of words and density of propositions increases (Morrison & Embretson, 2014). MLTM-D can be applied to assess both processing

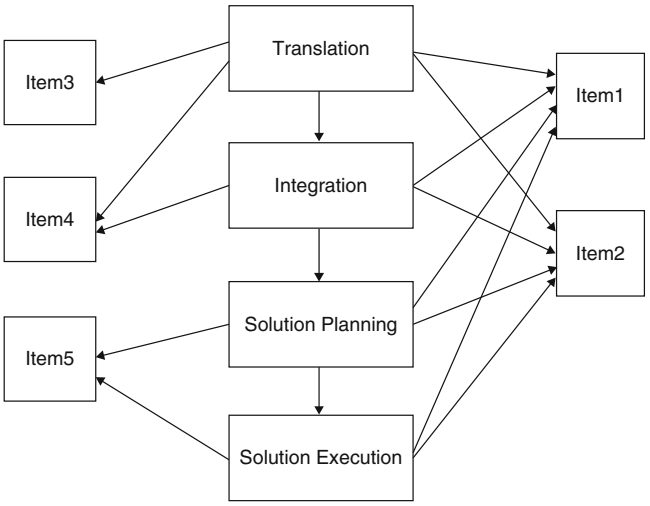


Fig. 9.2 Involvement of processes in mathematical items

competencies of examinees and the impact of the postulated features on process difficulty.

This chapter formally describes MLTM-D and then provides examples of applications to diagnose broad and narrow skills, as well as measure processing complexity and attainment.

9.2 Multicomponent Latent Trait Model for Diagnosis

In this section, MLTM-D will be presented as an explanatory conjunctive model for diagnosis for specifiable attributes at two levels. The first section includes a formulation of the basic model, followed by a consideration of the alternative types of scores that are appropriate for explanatory modeling. Also, an extended consideration of diagnosing component and attribute mastery will be presented. The second section concerns model estimation and includes model identification, estimation of item parameters and person trait levels. The third section concerns assessment of fit. This section includes methods for assessing model and item fit, as well as the assessment of score reliability and decision confidence for mastery categorizations.

9.2.1 Formulation of the Model

Basic Form of MLTM-D It is assumed that the test is complex, with items, $X = (X_1, \ldots, X_K)$, varying in which latent components underlie the probability of a correct response. The probability that examinee n solves item k, $P(X_{kn} = 1)$, is given by MLTM-D as the product of the component probabilities $P(X_{dkn} = 1)$ for item k as follows:

$$P\left(X_{kn} = 1\right) = \prod_{d=1}^{D} P(X_{dkn} = 1)^{c_{dk}} \tag{9.1}$$

where $c_{dk}$is a binary variable for the involvement of component d in item k. For components in item k, where $c_{dk} = 0$, then $P(X_{dkn} = 1)^{c_{dk}} = 1$. If $c_{dk} = 1$ for only one component d in item k, then $P(X_{kn} = 1) = P(X_{dkn} = 1)$.

For all components in which $c_{dk} = 1$, the component probabilities depend on component ability, θ_{dn}, for person n and the difficulty of component d in item k as follows:

$$P\left(X_{dkn} = 1\right) = \frac{\exp\left(\theta_{dn} - \sum_{m=1}^{M} \eta_{dm} q_{dkm} + \eta_{d0}\right)}{1 + \exp\left(\theta_{dn} - \sum_{m=1}^{M} \eta_{dm} q_{dkm} + \eta_{d0}\right)}, \tag{9.2}$$

where q_{dkm} is the score for stimulus feature m in component d for item k, η_{dm} is the weight of feature m on the difficulty of component d and η_{d0} is the intercept for component d.

Combining Eqs. 9.1 and 9.2 yields MLTM-D as follows:

$$P\left(X_{kn} = 1\right) = \prod_{d=1}^{D} \left[\frac{\exp\left(\theta_{dn} - \sum_{m=1}^{M} \eta_{dm} q_{dkm} + \eta_{d0}\right)}{1 + \exp\left(\theta_{dn} - \sum_{m=1}^{M} \eta_{dm} q_{dkm} + \eta_{d0}\right)} \right]^{c_{dk}} \tag{9.3}$$

If q_{dkm} is the involvement of a specific skill on item k in component d, as in a skill hierarchy as shown on Fig. 9.1, then η_{dm} represents the relative difficulty of skill m. If components in items represent cognitive processes, q_{dkm} could be a score to represent the cognitive complexity of the stimulus features for component d.

A special case of MLTM-D occurs if items are scored within components for difficulty rather than for stimulus features or skills. That is, q_{dkm} would consist of K binary scores (0,1) to represent difficulty for item k on component d. In this case, $\eta_{d0} = 0$ and η_{dm} in Eqs. 9.2 and 9.3 would be item difficulties in a Rasch model for component d. That is, if β_{dk} is item difficulty, then $\sum_{m=1}^{K} \eta_{dm} q_{dkm} = \beta_{dk}$. MLTM-D would be written as follows:

$$P\left(X_{kn}=1\right)=\prod_{d=1}^{D}\left[\frac{\exp\left(\theta_{dn}-\beta_{dk}\right)}{1+\exp\left(\theta_{dn}-\beta_{dk}\right)}\right]^{c_{dk}}. \tag{9.4}$$

If all items involve a single component, which is the same across items, then Eq. 9.4 becomes the unidimensional Rasch model.

Alternative Forms of MLTM-D MLTM-D is expressed in Eq. 9.4 as a logistic multidimensional Rasch model with LLTM at the component level. For convenience in specific applications, MLTM-D can be expressed in alternative forms. That is, MLTM-D can be specified in the normal metric as follows:

$$P\left(X_{kn}=1\right)=\prod_{d=1}^{D}\left[\frac{\exp\left(1.7\left(\theta_{dn}-\sum_{m=1}^{M}\eta_{dm}q_{dkm}+\eta_{d0}\right)\right)}{1+\exp\left(1.7\left(\theta_{dn}-\sum_{m=1}^{M}\eta_{dm}q_{dkm}+\eta_{d0}\right)\right)}\right]^{c_{dk}}. \tag{9.5}$$

Also, MLTM-D can be expressed as one parameter (1PL) logistic model, with the constant item discrimination α,

$$P\left(X_{kn}=1\right)=\prod_{d=1}^{D}\left[\frac{\exp\left(\alpha_{d}\left(\theta_{dn}-\sum_{m=1}^{M}\eta_{dm}q_{dkm}+\eta_{d0}\right)\right)}{1+\exp\left(\alpha_{d}\left(\theta_{dn}-\sum_{m=1}^{M}\eta_{dm}q_{dkm}+\eta_{d0}\right)\right)}\right]^{c_{dk}}. \tag{9.6}$$

Scoring Items for Components and Attributes For K items, a component structure matrix, C_{kxd}, must be specified to determine the component d in item k, c_{dk}. Scoring for components should be conducted by two or more raters with relevant expertise. The structure of C_{kxd} can vary substantially in different applications. Table 9.1 shows four different sets of hypothetical component scores for a small set of items. In Set 1, all items involve a single component which varies across items. This specification results in four unidimensional models. In Set 2, component involvement follows a simplex pattern in which progressively more components are involved in the items. In Set 3, all items involve two of the three components, but each component is not involved in some items. In Set 4, component involvement varies from one to all three. Not all structures, however, are feasible. See the section below on model identification.

Within components, item difficulty is modeled as a weighted combination of attributes that are relevant for the component. The attribute structure matrix, Q^{d}_{kxm}, contains scores on the M_d attributes that impact component item difficulty for the K items in which $c_{dk}=1$. The q_{dkm} may consist of binary or continuous variables to predict item difficulty. The M_d variables scored for the attribute structure matrices Q^{d}_{kxm} typically will differ between components. For MLTM-D hierarchical skill structures, attributes defined within each component represent the narrow skills within a broad skill cluster. For MLTM-D components that represent cognitive

Table 9.1 Four different sets of scores for component involvement in items

	Set 1			Set 2			Set 3			Set 4		
Item	c_{1k}	c_{2k}	c_{3k}	c_{1k}	c_{2k}	c_{3k}	c_{1k}	c_{2k}	c_{3k}	c_{1k}	c_{2k}	c_{3k}
1	1	0	0	1	0	0	1	1	0	1	0	0
2	1	0	0	1	0	0	1	1	0	0	1	0
3	0	1	0	1	1	0	0	1	1	1	1	0
4	0	1	0	1	1	0	0	1	1	1	0	1
5	0	0	1	1	1	1	1	0	1	0	1	1
6	0	0	1	1	1	1	1	0	1	1	1	1

processes, the relevant stimulus features are derived from theory and/or empirical results.

As noted above, if Q^d_{kxm} consists of M binary scores (i.e., dummy variables) to represent each item involving the component, then the resulting weights are item difficulties. As for the component score matrix, C_{kxd}, model identification at the attribute level requires some limits on Q^d_{kxm}, which will be described in the estimation section.

9.2.2 Diagnosing Component and Attribute Mastery

Diagnosis of persons' mastery can be obtained at both the component and attribute level. Since MLTM-D is formulated for binary items, classification for mastery can be obtained by specifying a mastery probability, y. Applied at the component level, a mastery cutline, γ_d can be obtained and applied to each person's component θ_{dn} to define mastery versus non-mastery. That is, define $\overline{P_d}$ as the mean predicted probability of solving component d on the items for θ_d. Then the cutline, γ_d, for mastery on component d is the value of θ_d for which $\overline{P_d} \geq y$. Of course, the value of γ_d, will vary substantially with the specified value of y, with the lower bound often defined at $y = .50$. However, higher values are often specified by expert panels in the substantive area for which diagnosis is obtained.

For diagnosis of specific attributes or skill mastery, a probability for mastery, y, also must be specified. Since MLTM-D is a Rasch family model, common scale measurement of specific attributes and component traits can be used to obtain mastery diagnosis. If q_{dkm} are binary variables to represent narrow skills within component d, the predicted location for the skills, $\eta_{dm}q_{dkm} + \eta_{d0}$, indicates the position on the theta scale where $P(X_{dkn} = 1) = .50$. As for component mastery, if the specified mastery probability for skills within a component is greater than .50, then the skill location can be adjusted accordingly. That is, the location of skill m in component d, τ_{dm}, is determined by the θ_d for which the probability of solving skill m equal y.

Figure 9.3 presents an alignment of Geometry skills from a Grade 6 mathematical achievement test with τ_{dm} located at $P(X_{dkn} = 1) = .70$. Fig. 9.3 also shows the

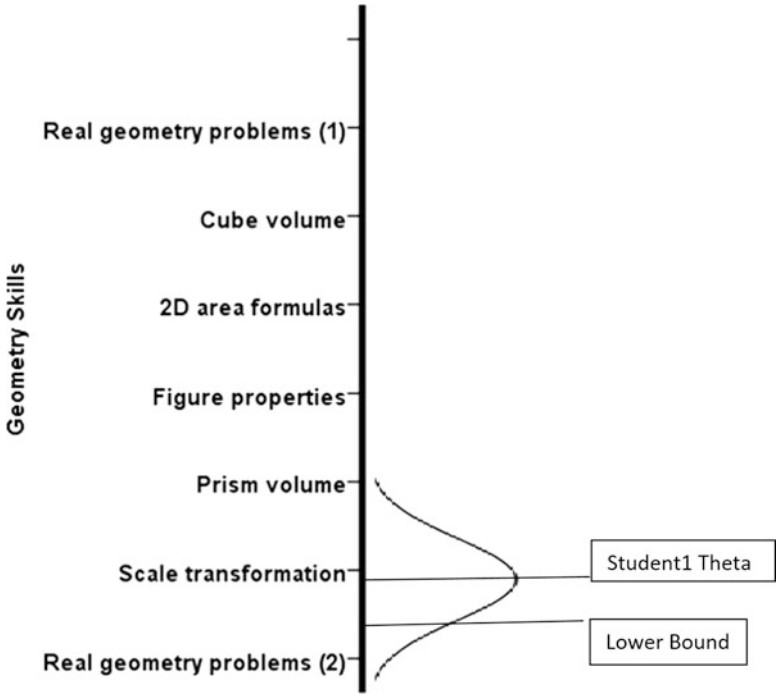


Fig. 9.3 Geometry skill difficulty continuum and an examinee's estimated trait distribution

plausible trait distribution from an examinee with the estimated component theta as the mean and the standard error of measurement as the standard deviation (i.e., $\theta_d \sim N(\theta_{dn}, \sigma_{\theta_{dn}})$).

Mastery for examinee n on skill m in component d is scored as 1 if $\theta_{dn} \geq \tau_{dm}$, otherwise skill mastery is scored as 0. The number of skills mastered for component d is the sum of the mastered skills. For the examinee on Fig. 9.3, $\theta_{dn} \geq \tau_{dm}$ for only one skill. It should be noted that interpretability of skill mastery depends on the strength of prediction of item difficulty from variables that represent the narrow skills involved.

9.3 Estimation

9.3.1 Model Identification

MLTM-D can be applied to tests in which component involvement varies across items. For K items, a component structure matrix, C_{kxd}, must be specified to determine the involvement of component d in item k. The number of item blocks

involving the same combination of components is $2^D - 1$, as items that involve no components is not a viable pattern. To identify the model, the matrix resulting by pre-multiplying C_{kxd} by its transpose must result in a matrix with full rank, such that $C_{kxd}{}' C_{kxd} = S_{dxd}$. Similarly, requirements for model identification involve the structure of, Q^d_{kxm} within each component. That is, pre-multiply by the transpose, $Q^{d'}_{kxm}{}'$, should result in a matrix of full rank for each component d.

Further, as for other multidimensional latent trait models, model identification requires fixing the scale of measurement. Assume for convenience that trait level is distributed as multivariate normal ($\boldsymbol{\theta} \sim MVN(\mathbf{0}, \sum)$). If MLTM-D is estimated as traditional Rasch models at the component level, as in Eq. 9.3, then the mean component trait levels must be fixed (e.g., $\bar{\boldsymbol{\theta}} = \mathbf{0}$). If a 1PL variant of MLTM-D is estimated, then the diagonal of Σ must be set to 1.

9.3.2 Estimating Item Parameters

The item parameters for MLTM-D may be estimated by a variety of methods, including marginal maximum likelihood (MML), which will be reviewed in this section. Let $\mathbf{x}_n = \{X_{1n}, X_{2n}, \ldots, X_{Kn}\}$ be the response pattern for examinee n, on K items, $k = 1,2, \ldots, K$. Then, the probability of the response pattern for the component trait levels, $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_D\}$, may be given as follows:

$$P(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{k=1}^{K} {P_{kn}}^{X_{kn}} (1 - P_{kn})^{1-X_{kn}}. \tag{9.7}$$

Thus, the log likelihood for the observed data, $\mathbf{X}$, on K items for N examinees is

$$\ln L(\mathbf{X}) = \sum_{n=1}^{N} \ln P(x_n). \tag{9.8}$$

For MML, the probability of a specific response pattern, $\mathbf{x}_n$, is expressed in terms of the population distribution as follows:

$$P(\mathbf{x}_n) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} P(\mathbf{x}_n|\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} P(\mathbf{x}_n|\boldsymbol{\theta})\, g(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{9.9}$$

where $g(\boldsymbol{\theta})$ is the probability density function of $\boldsymbol{\theta}$. Typically $\boldsymbol{\theta}$ is assumed to be distributed as $MVN(\mathbf{0}, \Sigma)$ and, for convenience, assume that $\Sigma = \mathbf{I}$.

The log likelihood for $\mathbf{X}$ may be expressed as

$$\ln L(X) = \sum_{n=1}^{N} \ln \left[\int_{\boldsymbol{\theta}} P\left(\mathbf{x}_n | \boldsymbol{\theta}\right) g\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} \right]. \tag{9.10}$$

The expectation-maximization (EM; Bock & Aitkin, 1981) algorithm involves expectations at various trait levels, $\boldsymbol{\theta}$, for the number of persons, $\tilde{N}$, and for the number of persons passing item k, $\tilde{\tau}_k$. For a particular response pattern, $\mathbf{x}_n$, these expectations are given as follows:

$$\tilde{N} = \left[\frac{\sum_{n=1}^{N} L\left(\mathbf{x}_n | \boldsymbol{\theta}\right)}{P\left(\mathbf{x}_n\right)} \right] \quad \tilde{\tau}_k = \left[\frac{\sum_{n=1}^{N} x_{kn} L\left(\mathbf{x}_n | \boldsymbol{\theta}\right)}{P\left(\mathbf{x}_n\right)} \right]. \tag{9.11}$$

To estimate item parameters, η_{dm}, for MLTM-D with MML, the derivative of the log likelihood of the data may be expressed with the expectations as follows:

$$\frac{\partial \ln L}{\partial \eta_{dm}} = \int_{\boldsymbol{\theta}} \left[\sum_{k=1}^{K} \left(\frac{\tilde{r}_k - \tilde{N} P_{kn}}{P_{kn}(1 - P_{kn})} \frac{\partial P_{kn}}{\partial \eta_{dm}} \right) \right] g\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}. \tag{9.12}$$

And, finally,

$$\frac{\partial \ln L}{\partial \eta_{dm}} = \int_{\boldsymbol{\theta}} \left[\sum_{k=1}^{n} \left(\frac{\tilde{r}_k - \tilde{N} P_{kn}}{P_{kn}\left(1 - P_{kn}\right)} c_{dk} \prod_{h=1, h \neq d}^{D} P_{knh}^{c_{kh}} \left(-q_{dkm}\right) P_{dkn}\left(1 - P_{dkn}\right) \right) \right] g\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}, \tag{9.13}$$

where q_{dkm} consists of K binary variables to define component item difficulty.

Integration for the D components can be implemented with Hermite-Gaussian quadrature. For S quadrature points within each of the D components, $\mathbf{X} = \{X_{q1}, X_{q2}, \Lambda X_{qS}\}$, the total number of quadrature points is S^D. Each quadrature point, X_q, has an associated weight $W(X_q)$. The parameter $\tilde{r}_{k \bullet q_1 q_2 \ldots q_D}$ represents the number of examinees expected to respond positively to item k with trait vector defined by a D-tuple of quadrature points. Similarly, $\tilde{N}_{q_1 q_2 \ldots q_D}$ is the expected number of examinees in the sample with the D-tuple of quadrature points. The integrals in Eq. 9.11 may be approximated by applying Hermite-Gauss quadrature as follows:

$$\frac{\partial \ln L}{\partial \eta_{dk}} \approx \sum_{q_1=1}^{Q} \sum_{q_2=1}^{Q} \cdots \sum_{q_D=1}^{Q} \left[\sum_{k=1}^{K} \left(\frac{\tilde{r}_{k \bullet q_1 q_2 \ldots q_D} - \tilde{N}_{q_1 q_2 \ldots q_M} P_{kn}}{P_{kn}\left(1 - P_{kn}\right)} \frac{\partial P_{kn}}{\partial \eta_{dk}} \right) \right] W\left(X_{q_1}\right) W\left(X_{q_2}\right) \ldots W\left(X_{q_D}\right). \tag{9.14}$$

9.3.3 Person Estimation

Person estimates for component trait levels and associated standard errors can be estimated by customary methods such as *expected a posteriori* (EAP) and *maximum a posteriori* (MAP) as well as by Bayesian sampling methods (e.g., Markov-chain Monte Carlo). For either EAP or MAP, normal priors are typically specified, $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \Sigma)$. Multidimensional EAP, as implemented for MLTM-D, involves Q^D quadrature points.

9.4 Assessing Fit of MLTM-D

9.4.1 Model and Item Fit

Comparisons based on the log likelihood of the data can be used to examine the fit of MLTM-D. Nested models can be compared using likelihood ratio chi-square tests. For example, the multicomponent nature of the data is examined by comparing the log likelihood of MLTM-D with an appropriate unidimensional model as follows:

$$\chi^2 = (-2lnL_{1PL}) - (-2lnL_{MLTM-D}), \tag{9.15}$$

since the unidimensional 1PL model is a special case of MLTM-D (i.e., the 1PL model is a single component model). Non-nested models can be compared with the AIC index, where a lower AIC index indicates better model fit.

Another global index, the delta statistic (Embretson, 1997), involves comparing the log likelihoods of alternative MLTM-D's; (1) lnL_{null}, for a null model in which items are modeled as equally difficult within components, (2) $lnL_{saturated}$, for the saturated model in which items within a component have unique difficulty estimates and (3) $lnL_{restricted}$, for a restricted model in which predictors of item difficulty within a component are postulated (e.g., narrow skill categories). The index may be written as follows:

$$\Delta = \sqrt{lnL_{null} - lnL_{restricted} / lnL_{null} - lnL_{saturated}}. \tag{9.16}$$

The magnitude of Δ is similar to a multiple correlation.

At the item level, fit can be analyzed by comparing observed to expected frequencies within groups of examinees with similar expected total scores on the test. That is, examinees are categorized by their mean expected probabilities of item solving on the test, $\overline{P}\ (X_{kn} = 1)$. The observed and expected frequencies for each item can be compared by either likelihood ratio fit statistics or standardized residuals (see Embretson, 2015 for further details).

9.4.2 *Assessment of Person Reliability and Decision Confidence*

Component trait levels and their associated standard errors can be estimated by customary methods, such as *expected a posteriori* (EAP) and *maximum a posteriori* (MAP). As noted in du Toit (2003), a composite reliability estimate can be obtained for a population. For EAP estimates, the empirical reliability on component d is given as follows:

$$\rho_d = \sigma^2_{\theta d} / \left(\sigma^2_{\theta d} + \overline{\sigma^2_{nd}}\right), \tag{9.17}$$

where $\sigma^2_{\theta d}$ and $\overline{\sigma^2_{nd}}$ are the variance of θ_d and the mean error variance, respectively.

For MLTM-D estimates of person *mastery*, determined using the cutlines as described above, IRT-based indices of decision confidence (e.g., Lewis & Sheehan, 1990; Rudner, 2005) are appropriate. Assume that the component d trait estimate for person n, θ_{nd} , with associated error variance, $\sigma^2_{\theta_{dn}}$ defines a plausible normal distribution of component trait levels, $\theta^*_{nd} \sim N\left(\theta_{nd}, \sigma^2_{\theta_{dn}}\right)$. The mastery cutline for the component, γ_d, as defined above, can be located on the estimated plausible distribution of theta for each person. Decision accuracy depends on both distance from the cutline and the standard error of measurement. That is, the proportion of $\theta^*_{nd} \geq \gamma_d$ indicates decision accuracy for person n if $\theta_{nd} > \gamma_d$. However, if $\theta_{nd} < \gamma_d$, the proportion of $\theta^*_{nd} < \gamma_d$ indicates decision accuracy.

Diagnosing skills within components reliably also depends on the distribution of plausible trait levels for each person. As explicated above, assessing skill mastery depends on the location of skill m at defined mastery level, τ_{dm}, on the component latent trait continuum. If $\theta_{nd} > \tau_{dm}$, then the proportion of the plausible distribution of $\theta^*_{nd} \geq \tau_{dm}$ is decision confidence. If $\theta_{nd} < \tau_{dm}$, then the proportion of the plausible distribution of $\theta^*_{nd} < \tau_{dm}$ is decision confidence.

A more complete formulation of decision accuracy for components and attributes is presented in Embretson, Morrison and Jun (2015).

9.5 Application

In this study, MLTM-D is applied to a year-end mathematical achievement test for Grade 7 that is used for state accountability purposes. Thus, the test is high stakes for assessing overall student competency estimates. MLTM-D is applied to item responses to assess student mastery of both broad and narrow skills. The data were also analyzed to assess students' cognitive processing capabilities in solving the mathematical items and compared between students with varying proficiency levels and background variables.

9.5.1 Method

Tests The Grade 7 mathematics achievement test consisted of 71 items with heterogeneous content to represent broad achievement. Item content was specified by hierarchically organized skills, as shown on Fig. 9.1, with the specific skills nested within four areas, Number, Algebra, Geometry and Data. The blueprint for each area included specifications for several skills. For example, the Number area contained 6 skills, ranging from "Understanding the Impact of Multiples of 10" to "Multiplication and Division with Decimals". Similar ranges of skills are specified for the other three areas. The items were developed to represent the narrow skills defined within the areas.

The standard for basic proficiency on the overall test was determined by expert panels at two levels. This evaluation resulted in a test cutline of 54% ($y = .54$) of items passed.

9.5.1.1 Examinees

A random sample of 5087 Grade 7 students was selected from the complete operational data in the participating state. All tests were computer administered at the end of the school year.

9.5.1.2 Item Scores

Items were scored to represent the components (i.e., skill clusters) and specific skills for MLTM-D. These scores were included in the estimation of MLTM-D item parameters as the C and Q matrices, respectively, as described above. Although the items were written for a specific skill, some items involved more than one skill. A panel, consisting of educators, a mathematician and an educational psychologist, scored the items for involvement of multiple skills. The skill clusters for the component C matrix were defined by the four areas in the blueprint (i.e., Number,

Algebra, Geometry and Data) and the specific skills for the Q matrices by the indicators listed within each area. While rater scores always included the intended indicator for an item, several items were reliably scored as involving additional indicators. If an item involved a specific skill within an area, it was scored as involving the area. Rater reliabilities for skill involvement were sufficiently high (the mean for Cohen's alpha was .709). Items with different categorizations across raters were resolved with panel discussion.

The 71 mathematical items were also scored for the involvement of Mayer's (2003) postulated cognitive processes by a panel with expertise in cognitive psychology. It was found that three stages could be distinguished in the 71 items, defined as follows: (1) Translation, interpreting words into mathematically relevant terminology, (2) Integration, organizing the terms in the problem into equations and (3) Solution Planning/Execution, finding solutions for the unknown quantities. The item scores provided a C_{kxd} for the test to estimate individual differences in processing capabilities. Both rater reliabilities (Cohen's alpha = .690) and procedures for discrepant categorization were similar to the skill panel. The items were also scored for stimulus complexity variables that were postulated to impact the difficulty of the cognitive components, based on prior research (Morrison & Embretson, 2014). These variables were as follows: (1) Translation included *Number of Context Words, Number of Symbols or Numbers, Undefined Mathematical Terms*, (2) Integration included *Generate Equations, Translate Equations, Translate Diagrams Visualization, Infer Patterns* and (3) Solution Planning/Execution included *Number of Subgoals, Relative Definitions of Variables, Number of Procedures, Number of Computations* and *Procedural Knowledge Level*.

9.5.1.3 Mastery Status

Mastery status for the four skill clusters, Number, Algebra, Geometry and Data, was determined by cutlines on θ as described above. In this analysis, the probability, y, to establish the mastery cutlines, γ_d, was the minimum proportion of items solved for basic proficiency on the overall test (i.e., $y = .54$). Also, in some comparisons, *instructional proficiency* is included (where y is specified as a traditional standard of .70). Mastery of specific skills within areas for each examinee was obtained by estimating skill location on the relevant components, in accordance with y, and then aligning the skill locations with individual component competencies.

9.5.2 Results

9.5.2.1 Item Parameter Estimates

Item parameters for the 1PL variant of MLTM-D were estimated by MML. For all models, it was assumed that $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \Sigma)$, with the diagonal of Σ set to 1.0.

For skill cluster diagnosis, MLTM-D parameters for four components were estimated to represent the skill clusters of Number, Algebra, Geometry and Data, with component involvement based on the scored matrix, C_{kxd}. Three variants of MLTM-D were estimated; a null model, a restricted model and a saturated model. The saturated MLTM-D (−2lnL = 365,390, AIC = 365,574) had 94 parameters, with unique item difficulty estimates for each relevant component. For comparison, a unidimensional 1PL model (−2lnL = 369,081, AIC = 369,225) with 72 parameters was estimated. The unidimensional model fit significantly worse than the saturated MLTM-D ($\chi^2 = 3691$, df = 22, p < .001) and had a higher AIC index. Thus, the multidimensionality of the data was supported.

The restricted model contained parameters to represent the narrow skills (i.e., indicators) within each component, Q^d_{kxm}. That is, q_{dkm} are binary variables to represent the narrow skill category m for item k on component d. The fit of this restricted model (−2lnL = 386,978, AIC = 387,062) with 42 parameters differed significantly from the saturated model ($\chi^2 = 21{,}588$, df = 52, p < .001), indicating that the skill clusters do not fully account for item differences within components. However, the strength of the relationship of skill clusters to item difficulty was examined by further comparisons with the null model. That is, the null model (−2lnL = 418,283, AIC = 418,311), with a single item difficulty and item discrimination within each component, was estimated. The likelihood ratio fit statistic of .769 obtained from in Eq. 9.16 above, which quantifies the relative differences of the saturated and restricted model from the null model, indicated strong alignment of the skills on the latent dimensions underlying the four components.

Item fit was examined using the standardized residuals (*SR*) of expected versus observed frequencies of item solving. A total of 14 score categories of examinees with similar overall item solving probabilities were used to compute *SR* for each item. Only two of the 71 items had standardized residual (SR) values that exceeded expectation from *SR* $N(0, 1)$.

MLTM-D item parameter estimates for the three cognitive processes were also obtained by MML as described above. As for skills, three variants of MLTM-D were estimated; a null model, a restricted model and a saturated model. The saturated MLTM-D for cognitive processes, with unique item difficulty estimates within each relevant component, had 111 parameters. The model overall fit (−2lnL = 364,915, AIC = 365,319) differed significantly from the unidimensional 1PL model ($\chi^2 = 4166$, df = 39, p < .001) and had a lower AIC index. Thus, as for the saturated MLTM-D for skills, the multicomponent nature of the data also was supported for the cognitive process model. However, the AIC index for the cognitive model was somewhat lower than for the saturated skill cluster model ($AIC_{diff} = 255$), indicating better overall fit.

Model comparisons of the null, restricted and saturated models indicated moderately strong prediction of component item difficulty by the stimulus complexity factors. The restricted model with 14 cognitive complexity predictors had 23 parameters and the model overall fit (−2lnL = 395,808, AIC = 395,854) differed significantly from the saturated cognitive model ($\chi^2 = 30{,}893$, df = 88, p < .001)

and had a higher AIC index. The null model ($-2\text{lnL} = 417{,}830$, AIC = 418,311) had 9 parameters, a single item difficulty and item discrimination within each component, plus three covariances of the theta estimates. The likelihood ratio fit statistic, based on comparisons to the restricted and saturated model to the null model, indicated moderately strong fit ($\Delta = .645$).

9.5.3 Competency and Mastery of Broad and Narrow Skills

Component competency levels for persons were estimated by EAP, using normal priors. Table 9.2 presents descriptive statistics on competency levels for the four skill cluster components. It can be seen that empirical reliability (r_{tt}) is moderately strong for Number, Algebra, and Geometry, but somewhat weaker for Data. Basic mastery was obtained for the four areas using the proficiency cutline of $y = .540$, based on the state standards for proficiency on the test as a whole. Proportions of examinees assessed for basic mastery ranged from .783 to .862. The decision confidence indices (DCI) were high, ranging from .853 to .939, which indicates a high degree of reliability for basic mastery assessments. A repeated measures analysis of variance, with the Huyhn-Feldt correction for sphericity, indicated that mastery proportions differed significantly between the four areas ($F = 102.462$, $p < .001$). Relatively fewer students had proficiency in Number and Data than in Algebra and Geometry. Further, the correlations between mastery categorizations were moderate, ranging from .418 to .493.

Table 9.3 presents mastery pattern frequencies from two different cutlines; basic proficiency ($y = .54$) and instructional proficiency ($y = .70$). These are the actual cutlines as established by the department of education in the participating state for this particular test. Although most students are classified as having basic proficiency (i.e., 65.1%), the higher cutline for instructional proficiency classifies the majority of students as having one or more non-mastered area. For both basic and instructional proficiency, the two patterns with the highest frequencies are (1) non-mastery in all four areas and (2) non-mastery in only Data. However, all 15 patterns of non-mastery were observed.

Narrow skills were assessed based on skill alignment and the plausible distributions of component competencies for each student, as described above. That is, skills were aligned on a continuum within each component based on the estimated parameters for the skill in the restricted MLTM-D. The aligned skills within each

Table 9.2 Descriptive statistics on components and mastery for skill clusters from 1PL variant of MLTM-D

	Component trait level			Component mastery		
	Mean	SD	r_{tt}	Mean	SD	DCI
Number	−.000	.907	.767	.783	.412	.889
Algebra	.189	.935	.802	.862	.345	.939
Geometry	−.013	.941	.801	.848	.359	.912
Data	−.001	.811	.654	.790	.407	.853

Table 9.3 Mastery patterns based on basic proficiency and instructional proficiency cutlines

Pattern	Basic Proficiency $y = .54$		Instructional $y = .70$	
NAGD	Frequency	Percent	Frequency	Percent
0000	271	5.3	991	19.5
0001	71	1.4	120	2.4
0010	90	1.8	162	3.2
0011	63	1.2	77	1.5
0100	106	2.1	299	5.9
0101	83	1.6	130	2.6
0110	158	3.1	214	4.2
0111	263	5.2	269	5.3
1000	38	.7	70	1.4
1001	37	.7	37	.7
1010	52	1.0	61	1.2
1011	80	1.6	81	1.6
1100	59	1.2	125	2.5
1101	109	2.1	130	2.6
1110	294	5.8	435	8.6
1111	3313	65.1	1886	37.1
Total	5087	100.0	5087	100.0

Table 9.4 Number of skills mastered at two cutlines

		Basic proficiency $y = .54$		Instructional $y = .70$	
	Total skills	Mean	SD	Mean	SD
Number	6	5.121	1.506	4.364	1.913
Algebra	7	6.090	1.742	5.398	2.135
Geometry	6	4.813	1.540	4.105	1.769
Data	6	4.461	1.737	3.077	1.917

component were compared to lower bound theta for each student, defined as the point for which 85% of the student's plausible distribution was above. Table 9.4 presents the number of skills that are mastered for each area. It can be seen that the number of mastered skills varies both by area and by the proficiency cutlines. It should be noted that students' mastery in an area do not necessarily not imply mastery of all skills within an area. The difficult skills may not be mastered by students close to the cluster cutline.

9.5.4 Cognitive Processes

Table 9.5 presents descriptive statistics on the component process estimates for the students. It can be seen that moderate to strong empirical reliability was found for individual differences on the three processes.

Table 9.5 Descriptive statistics on cognitive process competencies

	Component trait level		
	Mean	SD	r_{tt}
Translation	.16	.758	.71
Integration	.12	.843	.84
Solution	.13	.923	.88

The relationship of individual differences on the component processes to student variables was examined using a multivariate analysis of variance, using Roy's largest root. The background variables examined included Gender (Male, Female), First Language (English, Spanish and Other), Race-ethnicity (White, Black and Hispanic) and Proficiency Category (Warning, Approaches, Meets, Exceeds and Exemplary). Significant overall effects were found for Gender ($p < .001$, $\eta^2 = .014$), First Language ($p < .001$, $\eta^2 = .038$), Race ($p < .001$, $\eta^2 = .071$) and Proficiency Category ($p < .001$, $\eta^2 = .906$), with varying effect sizes.

The repeated measures analysis of the three cognitive processes were conducted using Huynh-Feldt's correction. Significant interactions of Cognitive Processes (Translation, Integration, Solution) were not observed with Race ($p = .053$, $\eta^2 = .001$) or with First Language ($p < .716$, $\eta^2 < .001$). However, Cognitive Processes did interact significantly with Gender ($p = .005$, $\eta^2 = .001$), but with a very small effect size, and with Proficiency Category ($p < .001$, $\eta^2 = .887$), which had a very large effect size. Figure 9.4 shows that Translation is relatively higher than Integration and Solution within the lower three categories. In the highest category, Translation is relatively lower than the other two categories.

9.5.5 Discussion

This application of MLTM-D to a mathematical achievement test provides an example of the diverse findings that are available. Competency levels and mastery for broad skills in four areas were examined. Individual differences in overall competency levels were observed between the skill areas, and they were assessed with moderately strong reliability. Further, the mastery assessments, obtained by applying cutlines, had high levels of decision confidence. Also, distinct patterns of mastery for the four areas were observed. These results are potentially important for individualizing remedial instruction for the areas. For narrow skills, strong alignment of skill difficulty on the component dimensions was observed. Thus, individual skill mastery could be assessed with relatively strong levels of decision confidence. These results provide further information that is relevant to individualizing instruction.

Individual differences in the cognitive processes that are involved in item solving were also assessed with MLTM-D. Competency levels in the three assessed cognitive processes differed significantly between students with different backgrounds.

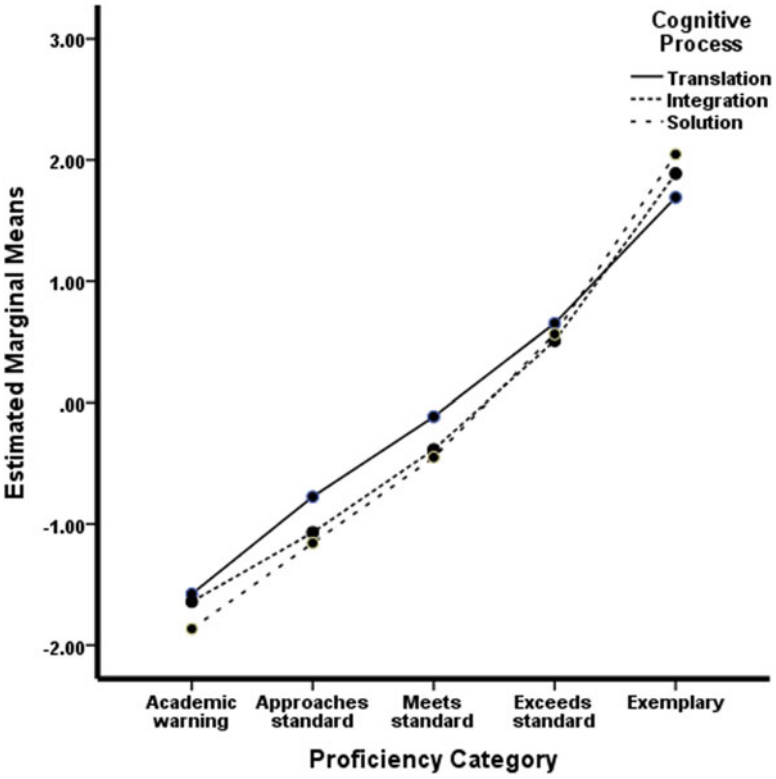


Fig. 9.4 Cognitive process means by proficiency category

These results suggest that instruction can be further individualized to focus on relevant cognitive processes involved in item solving.

It should be noted that MLTM-D is applicable to tests with complex and somewhat heterogeneous content with respect to required skills and cognitive processes. MLTM-D would not be applicable to tests with narrow and homogeneous content. Many items in the achievement test domain tend to be heterogeneous. For mathematical achievement, currently items may be increasing in heterogeneity due to increased emphasis on real world problems. In contrast, ability test items are typically somewhat less heterogeneous than achievement test items as indicated by higher internal consistency indices. However, heterogeneity in cognitive processing could be designed in these items to allow differential diagnosis by models such as MLTM-D.

9.6 Summary

The purpose of this chapter was to provide an overview of the multicomponent latent trait model for diagnosis (MLTM-D). Alternative variants of the model were formulated, along with procedures for diagnosis. Estimation procedures and model fit indices were also presented, with special emphasis on reliability for assessing competency levels and mastery for both broad and narrow skills. Cognitive process assessment was also described throughout the chapter. The application to a broad test of mathematical achievement was presented to illustrate the full scope of MLTM-D applications to broad skills, narrow skills and cognitive processes.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: *Application of an EM algorithm. Psychometrika*, 46, 443–459.

Daniel, R. C., & Embretson, S. E. (2010). Designing cognitive complexity in mathematical problem solving items. *Applied Psychological Measurement, XX*, 348–364.

du Toit, M. (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International, Inc.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–322). New York, NY: Springer.

Embretson, S. E. (2015). The multicomponent latent trait model for diagnosis: Applications to heterogeneous test domains. Invited paper for *Applied Psychological Measurement*, *XX*, 6–30.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50*, 328–344.

Embretson, S. E., Morrison, K., & Jun, H. W. (2015). Reliability of diagnosing broad and narrow skills with the multicomponent latent trait model: A study of middle school mathematics. In L. Andries van der Ark, D. M. Bolt, W.-C. Wang, J. M. Douglas, & S.-M. Chow (Eds.), *Quantitative psychology research*. New York, NY: Springer.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*, 14–36.

Hensen, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 1919–1210.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.

Mayer, R. (2003). Mathematical problem solving. In J. M. Royer (Ed.), *Mathematical cognition: A volume in current perspectives on cognition, learning, and instruction* (pp. 69–92). Cambridge, MA: Information Age Publishing, Inc.

Morrison, K., & Embretson, S. E. (2014). Using cognitive complexity to measure psychometric properties of mathematics assessment items. *Multivariate Behavior Research, 49*, 1–2.

National Assessment Governing Board (NAGB). 2017. *Mathematics framework for the 2015 National Assessment of Educational Progress*. U.S. Department of Education.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation, 10*, 1–4.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

Chapter 10
Explanatory Cognitive Diagnostic Models

Yoon Soo Park and Young-Sun Lee

Abstract Student- and school-level information from large-scale educational data have been shown to explain trends in test taker performance and to inform factors that can enhance the learning environment. This study presents methods to specify and model predictive relationships of latent and observed explanatory variables within a cognitive diagnostic model, referred to as the Explanatory Cognitive Diagnostic Model (ECDM) framework. Explanatory factors can be incorporated simultaneously as observed covariates or latent variables (estimated using item response theory) that can explain patterns of attribute mastery. This chapter is divided into two studies that demonstrate real-world application using large-scale international testing data and simulation studies, which examine parameter recovery and classification for varying sample sizes and number of attributes. Simultaneous estimation of multiple observed and latent (using dichotomous and polytomous items as indicators for the latent construct) predictors show consistency in attribute classification and parameter recovery. Extensions of the ECDM framework are discussed.

10.1 Introduction of the Model

10.1.1 Background

Large-scale data collected from surveys and assessments often contain relational information that can *explain* associations between educational or psychological outcomes and background variables. *Explanatory variables* refer to either observed

Y. S. Park (✉)
Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, IL, USA
e-mail: yspark2@uic.edu

Y.-S. Lee
Teachers College, Columbia University, New York, NY, USA
e-mail: sly2003@columbia.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_10

or latent variables that can be used as predictors to model relationships in cognitive diagnostic models (CDMs). While explanatory approaches for CDMs have been limited in the literature, explanatory models in item response theory (IRT) have been used widely and include applications that incorporate both observed and latent variables (De Boeck & Wilson, 2004; Fox & Glas, 2003). Prior studies within the CDM framework have used explanatory variables in the form of covariates by using a latent regression modeling approach (Dayton & Macready, 1988; Huang & Bandeen-Roche, 2004; Mislevy et al., 1992; Park, Xing, & Lee, 2018). For example, access to educational resources such as computer or calculator can enhance the quality of learning, thereby increasing the mastery of knowledge and skills (Park, Lawson, & Williams, 2012). In this manner, identifying variables collected from students, schools, or types of learning environment such as access to educational tools (e.g., calculator) or examining the impact of student's psychological behavior or attitude (e.g., affect or confidence in subject area) can serve to explain the performance of students and also facilitate answering important substantive questions that can yield better student outcomes in the form of attribute mastery (Park & Lee, 2014).

Traditionally, analyses of assessment data have focused on two areas: (1) *descriptive measurement*, where the aim of the psychometric approach is to increase the precision of examinee performance measuring an underlying construct, or (2) *explanatory models*, which examine the association between item response data and related factors (De Boeck & Wilson, 2004). This chapter focuses on the explanatory tradition of psychometric models, targeting the relationship between test taker performance, as modeled using CDMs and explanatory variables.

In Chap. 10, a generalized explanatory approach to analyze relational data for CDMs is presented, following the framework used in Park et al. (2018). In the explanatory CDM framework, both observed and latent variables are modeled simultaneously – combining both structural and measurement components for CDMs – following a structural equation model (SEM) approach for CDMs. In the explanatory cognitive diagnostic model (ECDM) framework, observed and latent predictors can be specified to explain (1) mastery of attributes, (2) item response probabilities, or (3) higher-order latent trait of the CDM. When observed and latent predictors are specified on attributes, one can study how the explanatory variables affect attribute mastery. When they are specified on items, such predictors can provide information on possible differential item functioning (DIF). Moreover, when explanatory variables are specified on the higher-order latent trait, this can be used to examine how the predictor affects the overall ability measured in the assessment. The observed and latent variables can be either continuous or discrete, where latent explanatory factors can be estimated using IRT. We use the deterministic inputs noisy "and" gate (DINA; Junker & Sijtsma, 2001) model to demonstrate the framework, which can be applied to other CDM families and generalizations of DINA models (e.g., de la Torre, 2011; von Davier, 2008).

This chapter summarizes models presented in Park and Lee (2014) and in Park et al. (2018), as the foundation to presenting the explanatory framework for

CDMs. The focus of models and examples presented in this chapter are based on explanatory factors affecting attribute classification.

10.1.2 Explanatory Cognitive Diagnostic Models (ECDM)

CDMs are semi-parametric models that specify the classification of examinees into profiles of mastery that reflect fine-grained areas of skill (de la Torre, 2009; Rupp, Templin, & Henson, 2010; von Davier, 2008). CDMs include a latent class component that specifies skills as discrete latent variables with a parametric item response function (DiBello, Roussos, & Stout, 2007). To date, various CDMs have been developed and proposed depending on a variety of situations that range in the types of constructs, responses, and dimensionality of the data. Attributes have been modeled as latent classes (e.g., Haertel, 1989; Henson, Templin, & Willse, 2009) and also polytomous levels of mastery (von Davier, 2008). More generalized forms of CDMs include the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), general diagnostic model (GDM; von Davier, 2008), and the generalized DINA model (G-DINA; de la Torre, 2011).

The explanatory cognitive diagnostic model (ECDM) presented in this chapter uses the reparameterized DINA (RDINA; DeCarlo, 2011) to specify the explanatory factors. Prior studies have proposed alternative approaches to incorporating covariates to CDMs including discrete forms and multiple-group extensions (von Davier, Xu, & Carstensen, 2011; Xu & von Davier, 2008). Unobserved covariates can be examined using mixtures of diagnostic models, as presented in von Davier (2007, 2008).

ECDMs are motivated from explanatory IRT models that include latent or observed explanatory factors (see De Boeck & Wilson, 2004) and are extended here for latent class models that subsume CDMs. In RDINA, the DINA model is reparameterized using a logit transformation, where the probability of examinee i's response to item j (Y_{ij}) is modeled using item parameters f_j and d_j and binary latent variable η_{ij}. The parameter f_j indicates the log odds of guessing (g_j); the parameter d_j provides a measure of how well the item discriminates an examinee with or without the mastery of required attribute; and the binary latent variable η_{ij} indicates whether the examinee has mastered all required attributes (α_k) specified for the item in the Q-matrix:

$$\text{logit}\, p\left(Y_{ij} = 1 | \eta_{ij}\right) = f_j + d_j \eta_{ij} \tag{10.1}$$

Covariate Approach The simplest case of ECDM is a covariate model (Park & Lee, 2014). From Eq. (10.1), when an observed covariate, $\underline{Z}$, is introduced, the examinee's response probability conditioned on the covariate can be represented into two parts. As presented in Eq. (10.2), the response probability conditioning on the covariate can be partitioned into the response probabilities, $p(Y_{ij} \mid \underline{\alpha}, \underline{Z})$, or on the attribute probability, $p(\underline{\alpha} \mid \underline{Z})$ (see DeCarlo, 2011; Park & Lee, 2014). As such,

the equation for the observed variable affecting the response probability is shown in Eq. (10.3), and Eq. (10.4) represents the observed variable affecting the attributes (assuming independence):

$$p\left(Y_{i1}, Y_{i2}, \ldots, Y_{iJ}|\underline{Z}\right) = \sum\nolimits_{\underline{\alpha}} p\left(\underline{\alpha}|\underline{Z}\right) \prod\nolimits_{j} p\left(Y_{ij}|\underline{\alpha}, \underline{Z}\right) \tag{10.2}$$

$$\text{logit } p\ \left(Y_{ij}|\underline{\alpha}, \underline{Z}\right) = f_j + d_j \eta_{ij} + \underline{l_j} \underline{Z} \tag{10.3}$$

$$\text{logit } p\ \left(\underline{\alpha_k}|\underline{Z}, \underline{\xi}\right) = b_k + \underline{h_k} \underline{Z} \tag{10.4}$$

When the covariate $\underline{Z}$ is conditioned on the distribution of an attribute, parameters for item j are adjusted by $\underline{l_j}$, which represents the magnitude for which the guessing and slip rates shift (see Eq. (10.3)). In the DINA model, the guessing parameter indicates the probability that an examinee gets an answer correct without having mastered all required attributes; the slip parameter indicates the probability that an examinee gets an answer wrong, even if they mastered all required attributes. When the covariate is conditioned on in the distribution of an attribute, it adjusts the attribute parameters by $\underline{h_k}$ (see Eq. (10.4)).

In the covariate-only model, only observed variables are specified as predictors; that is, latent variables are not included. The covariate-only model (Park & Lee, 2014) includes the structural component, meaning that only observed covariates can be specified. This means that if a latent variable consisting of item response data is to be incorporated into the analysis, a two-step approach is required. For example, the latent variable will need to be estimated using IRT first, then the estimated latent variable can be used as an observed variable in either Eq. (10.3) or (10.4), following a covariate-only model.

Explanatory Cognitive Diagnostic Model: Incorporating Both Observed and Latent Predictors The ECDM approach presented in this chapter allows for simultaneous estimation of both observed and latent predictors, such that they can be specified within a single model (rather than a two-step process), as described in Park et al. (2018). An illustration is presented in Fig. 10.1 using graphic notation from Rupp et al. (2010) to demonstrate the attribute- and item-level explanatory RDINA, which allows estimation of factors that explain mastery of attributes and item responses.

In Fig. 10.1, the ECDM is presented in generalized form to include *both* observed and latent explanatory variables, where a latent variable (simultaneously estimated with items as indicators of the latent variable using IRT) is also specified as a predictor, in addition to the observed covariate. The shaded area indicates the explanatory latent variable ($\underline{\xi}$) that is simultaneously estimated using item response data from X_M (for M items) as indicators. The right side of the figures represents the CDM where attributes are linked to items using the Q-matrix, following standard

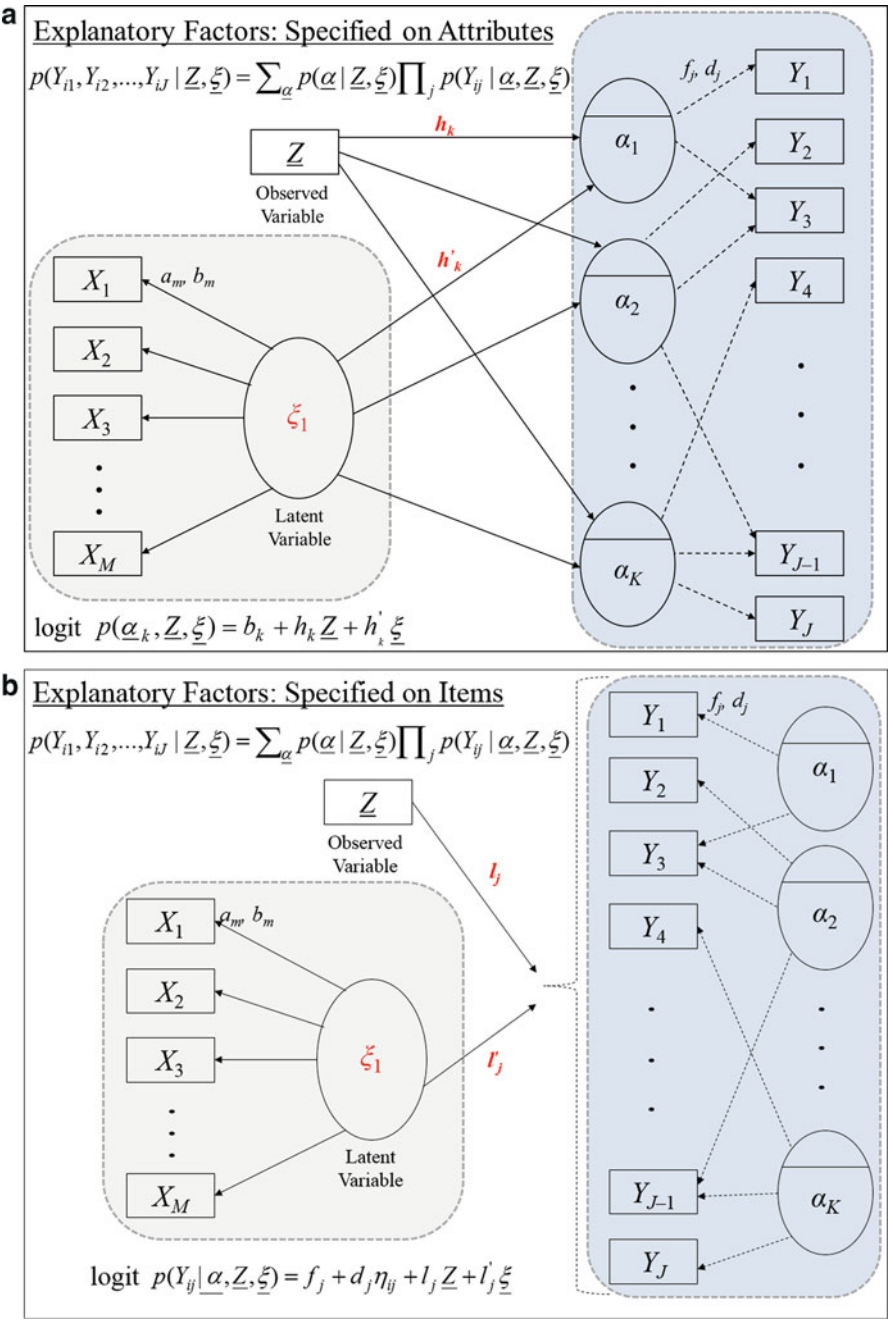


Fig. 10.1 Graphical representation of ECDM. (**a**) Explanatory factors (observed or latent) specified on attributes. (**b**) Explanatory factors (observed or latent) specified on items

specification in CDMs; the left side represents the explanatory variables. The top figure represents the attribute-level explanatory RDINA, with predictors specified at the attribute; the bottom figure shows the item-level explanatory RDINA.

In the ECDM which includes the explanatory latent variable ($\underline{\xi}$), the conditional response probability becomes modified from Eq. (10.2), as represented in Eq. (10.5). Moreover, the structural components linking the explanatory variables are expressed in Eqs. (10.6) and (10.7) for the item- and attribute-levels, respectively.

$$p\left(Y_{i1}, Y_{i2}, \ldots, Y_{iJ}|\underline{Z}, \underline{\xi}\right) = \sum\nolimits_{\underline{\alpha}} p\left(\underline{\alpha}|\underline{Z}, \underline{\xi}\right) \prod\nolimits_j p\left(Y_{ij}|\underline{\alpha}, \underline{Z}, \underline{\xi}\right) \tag{10.5}$$

$$\text{logit } p\left(Y_{ij}|\underline{\alpha}, \underline{Z}, \underline{\xi}\right) = f_j + d_j\eta_{ij} + \underline{l}_j\underline{Z} + \underline{l}_j{}'\underline{\xi} \tag{10.6}$$

$$\text{logit } p\left(\underline{\alpha}_k|\underline{Z}, \underline{\xi}\right) = b_k + \underline{h}_k\underline{Z} + \underline{h}_k{}'\underline{\xi} \tag{10.7}$$

Parameters in Eqs. (10.6) and (10.7) show the regression parameters for item and attribute levels, respectively, where parameters with a prime ($'$) denote the explanatory regression effects corresponding to latent variable predictors. The parameter $\underline{l}_j{}'$ indicates shift in the item parameter due to the latent variable; likewise, the parameter $\underline{h}_k{}'$ represents shift in the attribute parameter due to the latent variable.

The measurement model associated with the latent explanatory variable ($\underline{\xi}$) is estimated using IRT. For dichotomous item response X_M, the 2PL-IRT model is used (see Eq. 10.8), where parameters a_m and b_m represent discrimination and difficulty parameters for item m. For ordinal item response of N categories, the graded response model (GRM; Samejima, 1969) can be used (see Eq. 10.9), where a_m is the discrimination parameter and b_{nm} is the category threshold parameter.

$$\text{logit } p\left(X_{im} = 1|\xi_i\right) = a_m\left(\xi_i - b_m\right) \tag{10.8}$$

$$\log\left[\frac{p\left(X_{im} \geq n|\xi_i\right)}{p\left(X_{im} < n|\xi_i\right)}\right] = a_m\left(\xi_i - b_{nm}\right) \tag{10.9}$$

Estimation of $\underline{\xi}$ can incorporate other variants of IRT models as justified based on theoretical considerations.

10.1.3 Relationship Between RDINA and General Diagnostic Model (GDM)

The parameterization used in the ECDM can be extended and reparameterized as special cases of the general diagnostic model (GDM; von Davier, 2005). In the

GDM, the observed response X is modeled for i items, x response categories, and j respondents as follows:

$$P\left(X = x|i, j\right) = \exp\left[f\left(\lambda_{xi}, \theta_j\right)\right] / \left\{1 + \Sigma_m \exp\left[f\left(\lambda_{xi}, \theta_j\right)\right]\right\} \tag{10.10}$$

GDM item parameters are the $\lambda_{xi} = (\beta_{xi}, \boldsymbol{q}_i, \gamma_{xi})$, which include slope parameters and the Q-matrix specification, $\underline{q_i}$. In the DINA where attributes are binary, the skill vector for examinee j, $\theta_j = (a_{j1}, \ldots, a_{jk})$, are binary values. As shown in von Davier (2014, p.58), the DINA can be parameterized as a special case of the GDM as follows:

$$P\left(X_{\upsilon i} = 1|\boldsymbol{q}_i{}^*, \boldsymbol{a}^*\right) = \frac{\exp\left(\beta_i + \Sigma_k \gamma_{ik} a^*{}_k q^*{}_{ik}\right)}{1 + \exp\left(\beta_i + \Sigma_k \gamma_{ik} a^*{}_k q^*{}_{ik}\right)} \tag{10.11}$$

When a covariate $\underline{Z}$ or latent variable ($\underline{\xi}$) is introduced to Eq. (10.11), the following h_k and $h^{'}{}_k$ parameters are added:

$$P\left(X_{\upsilon i} = 1|\boldsymbol{q}_i{}^*, \boldsymbol{a}^*, \boldsymbol{Z}, \boldsymbol{\xi}\right) = \frac{\exp\left(\beta_i + \Sigma_k \gamma_{ik} a^*{}_k q^*{}_{ik} + h_k Z + h_k{}'\xi\right)}{1 + \exp\left(\beta_i + \Sigma_k \gamma_{ik} a^*{}_k q^*{}_{ik} + h_k Z + h_k{}'\xi\right)} \tag{10.12}$$

Taking the logit simplifies the model to the item-level ECDM as presented in Fig. 10.1.

10.2 Estimation

10.2.1 Estimation

Estimation of ECDM used the expectation-maximization (EM) algorithm, followed by the Newton-Raphson (NR) to obtain maximum likelihood (ML) or posterior mode (PM) estimates in case the maximum likelihood estimation does not exist (Park et al., 2018). For identification, the mean and variance of ξ was fixed to (0, 1). The observed Fisher information matrix was examined to be of full rank for local identification (Huang & Bandeen-Roche, 2004). Estimation for ECDMs in Fig. 10.1 was fit using Latent GOLD 5.0 (Vermunt & Magidson, 2013). Syntax for fitting ECDM using Latent Gold is available from the authors and also presented in Park et al. (2018; see Online Supplemental Material for syntax to fit model).

10.2.2 Parameter Recovery

This section summarizes the estimation results based on the following simulation conditions (Park et al., 2018):

1. Explanatory RDINA with 1 latent predictor (3 dichotomous items fit using 2PL),
2. Explanatory RDINA with 1 latent predictor (4 polytomous items fit using GRM),
3. Explanatory RDINA with 1 latent (GRM) and 1 observed dichotomous predictor, and
4. Explanatory RDINA with 2 latent predictors (GRM and 2PL) and 1 observed dichotomous predictor.

The simulation studies examined the conditions above for two sample size conditions (1000 and 2000 examinees) and two attribute sizes (3 attributes and 5 attributes) using the Q-matrix in Park and Lee (2014) and in Park et al. (2018).

When a single latent or observed predictor was specified in the ECDM, parameter recovery was consistent across conditions, number of attributes, and sample sizes (Conditions 1, 2, and 3). For these conditions, the recovery of the measurement component (recovery of IRT parameters) was particularly notable, as % bias was all less than 2.3%. In addition, RDINA item parameters had % bias less than 2.3%. For attribute parameters, when only a single predictor was specified at the attribute level, the bias for attribute difficulty (b_k) was modest. When multiple predictors were specified (Condition 4), bias increased. For the 3-attribute condition, % bias was all less than 8.4%. However, for the 5-attribute condition, % bias was 22.2% for the observed variable, while the latent variables estimated using IRT had % bias less than 27.5%. Overall, bias in estimation was modest for single-predictor ECDM or when predictors were either a single observed or a single latent explanatory variable. When more than one latent explanatory variable was included, bias was noticeably larger. For example, in the condition with 5 attributes (see Table 10.1), bias associated with covariate effects ranged between 19.0% and 27.4% for sample size of 2000. For parameter recovery of latent class sizes, bias was all less than 1.5%, indicating excellent recovery of the attribute distribution.

10.3 Assessment of Fit

10.3.1 Evaluating Model Fit

For the ECDM, model fit can be evaluated using likelihood-based information criteria such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). For classification, the proportion correctly classified (P_c; Clogg, 1995; de la Torre & Douglas, 2004; Park & Lee, 2014) statistic can be used, which is based on the accuracy in the recovery of latent classes using the maximum posterior probability for each attribute.

Table 10.1 Parameter recovery for condition with two latent variables (estimated using 2PL and GRM) and one observed covariate

Attrib	Model	Parameter	$n = 1000$			$n = 2000$		
			Bias	% Bias	MSE	Bias	% Bias	MSE
3	RDINA attribute	b_k	.022	12.1%	.132	−.011	8.3%	.074
		h_{k1} (observed)	−.005	6.4%	.769	−.022	3.4%	.375
		h'_{k2} (latent 1: 2PL)	.129	6.1%	.535	.013	3.9%	.311
		h'_{k3} (latent 2: GRM)	−.095	5.7%	.526	−.022	3.1%	.323
	RDINA item	f_j	.002	2.1%	.018	−.007	2.2%	.009
		d_j	.008	1.0%	.033	.013	.8%	.017
	2PL	b_j	−.010	1.2%	.021	−.008	1.6%	.010
		a_j	.050	1.9%	.173	.019	1.9%	.091
	GRM	b_{nm}	.011	.9%	.026	.005	.3%	.011
		a_m	.015	1.0%	.016	.007	.5%	.007
5	RDINA attribute	b_k	.225	28.4%	.186	.178	9.1%	.130
		h_{k1} (observed)	−.020	36.9%	.772	.071	22.2%	.568
		h'_{k2} (latent 1: 2PL)	−.173	31.0%	.656	−.152	19.0%	.401
		h'_{k3} (latent 2: GRM)	.303	41.5%	.910	.261	27.4%	.609
	RDINA item	f_j	−.017	3.5%	.023	−.007	2.6%	.010
		d_j	.026	1.4%	.040	.013	1.1%	.019
	2PL	b_m	.225	28.7%	.023	.209	29.0%	.013
		a_m	.012	.9%	.014	.012	1.0%	.007
	GRM	b_{nm}	.001	1.7%	.018	−.009	.7%	.012
		a_m	.001	2.2%	.166	.027	1.3%	.108

Note: Results pertain to parameter recovery from Condition 4

10.3.2 Simulation Studies Comparing ECDM with Two-Step Covariate CDM

Simulated data were fit comparing the ECDM to a two-step covariate RDINA model, where latent explanatory factors were estimated first and subsequently used to fit the covariate model. When comparing these results to a covariate RDINA with only observed predictors, the largest % bias for the structural attribute parameters (h and h' parameters in the 3- and 5-attribute conditions based on sample size of 2000) was 3.9% and 27.4% in the ECDM, whereas they were 14.2% and 36.3% in the covariate RDINA model, respectively. This reflects larger bias when a two-step covariate model is used, rather than the ECDM.

For the four simulation conditions, classification based on P_c was greater than .92, regardless of condition, number of attributes, and sample size. Moreover, bias in latent class sizes was all less than 1.5% across conditions.

10.4 Exemplary Application

10.4.1 Methods

To demonstrate an application of the ECDM, real data from Booklet 14 of the 2007 Trends in International Mathematics and Science Study (TIMSS) 4th grade mathematics were used ($n = 1975$ examinees).[1] The Q-matrix was adopted from Park and Lee (2014) for 25 items, measuring 5 attributes: α_1: Whole Numbers (Number Domain), α_2: Fractions & Number Patterns (Number Domain), α_3: Lines & Angles (Geometric Shapes and Measures Domain), α_4: Dimensions & Locations (Geometric Shapes and Measures Domain), and α_5: Data Display (Data Domain).

Data included item responses from 25 items and three explanatory variables: confidence (latent variable measured using 4 items), affect (latent variable measured using 3 items), and calculator (observed covariate). The explanatory variables were selected based on their association (correlation) with the mathematics score ($r_{\text{calculator}} = .14$, $r_{\text{affect}} = .13$, and $r_{\text{confidence}} = .33$, all $p < .001$).

1. Confidence (latent variable): 4-item self-reported measure of self-confidence in mathematics rated on a 4-point scale, ranging from "1: Disagree", "2: Agree a little", "3: Somewhat agree", and "4: Agree a lot". The items are: (1) I usually do well in mathematics, (2) Mathematics is easier for me than for many of my classmates, (3) I am good at mathematics, and (4) I learn things quickly in mathematics (Cronbach's alpha = .70; Mean = 2.02, SD = .71).
2. Affect (latent variable): 3-item self-reported measure of affect in mathematics, reported as "Yes = 1" or "No = 0". The items are as follows: (1) I enjoy learning mathematics, (2) Mathematics is fun, and (3) I like mathematics (Cronbach's alpha = .80; Mean = .76, SD = .34).
3. Calculator (observed variable): Self-reported status of calculator ownership (82% own calculators).

The attribute-level ECDM was fitted to the data. GRM and 2PL IRT were used to estimate "confidence" and "affect", respectively; "calculator" was included as an observed dichotomous variable. In addition, to compare the attribute-level effects of the predictors, a covariate RDINA model was also fitted, using a two-step approach by separately estimating "confidence" and "affect" via IRT and using the predicted values as covariates. The variance of "confidence" and "affect" was fixed to 1.0 for identification.

[1] Available at https://timssandpirls.bc.edu/TIMSS2007/idb_ug.html

Table 10.2 Explanatory measurement model parameters: GRM and 2PL IRT models

Construct	Item	Step (b_m)			Discrimination (a_m)		R^2
Confidence: Polytomous (Graded response)	1. I usually do well in mathematics	b_1	−4.47	(.18)	1.50	(.09)	.73
		b_2	−2.42	(.10)			
		b_3	.49	(.07)			
	2. Mathematics is easier for me than for many of my classmates	b_1	−2.35	(.09)	1.32	(.08)	
		b_2	−.73	(.06)			
		b_3	.78	(.06)			
	3. I am good at mathematics	b_1	−2.67	(.11)	1.54	(.09)	
		b_2	−1.05	(.07)			
		b_3	.32	(.07)			
	4. I learn things quickly in mathematics	b_1	−3.15	(.12)	1.26	(.08)	
		b_2	−1.34	(.07)			
		b_3	.85	(.06)			
Affect: Dichotomous (2 PL)	1. I enjoy learning mathematics		−1.30	(.09)	2.53	(.22)	.65
	2. Mathematics is fun		−.64	(.04)	1.28	(.10)	
	3. I like mathematics		−1.44	(.15)	3.44	(.45)	

Note:
1. Items measuring "confidence" are polytomous, scored using a 4-point scale ("Disagree", "Agree a little", "Somewhat agree", "Agree a lot"). GRM was used to fit the model; variance for θ was fixed to 1 for identification
2. Items measuring "affect" are dichotomous ("No", "Yes"). The 2PL was used to fit the model; variance for θ was fixed to 1 for identification
3. Values in parenthesis are standard errors
4. R^2 is the proportion of reduction in error between total error and predicted error (see Vermunt & Magidson, 2013)

10.4.2 Results

The attribute-level model converged (tolerance criteria $= 1.0 \times 10^{-8}$; $-2LL = 77{,}572.76$). The mean guessing and slip item parameter estimates were .28 and .27, respectively.

IRT Parameters IRT parameters for "confidence" and "affect" are presented in Table 10.2 and in Fig. 10.2.

Note that items that relate to "affect" and "confidence" are described in Table 10.2. Figure 10.2 provides a graphical illustration of parameters based on the ECDM

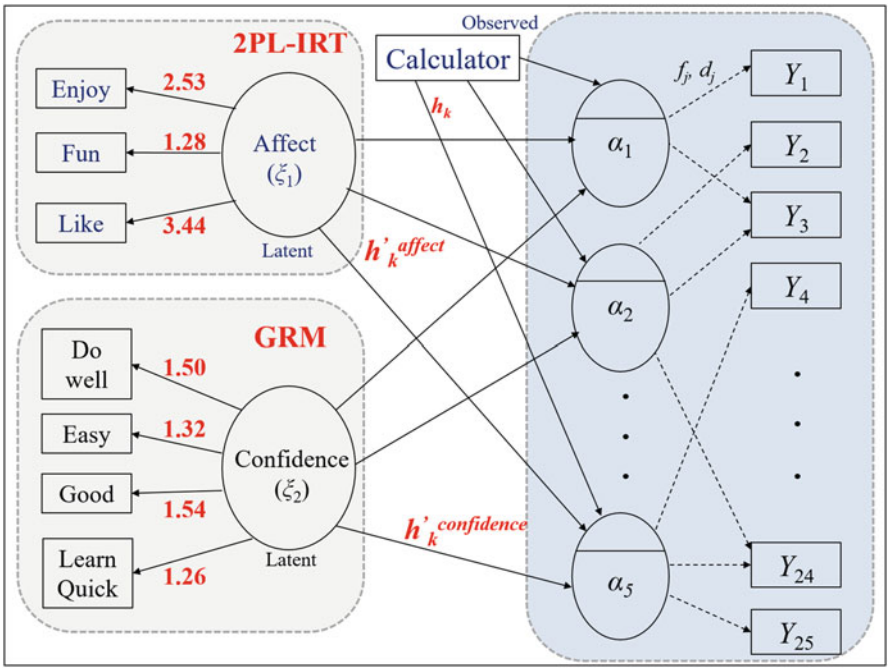


Fig. 10.2 IRT Parameter Estimates for the ECDM: Affect (ξ_1) and Confidence (ξ_2). Note: See Table 10.2 for descriptors of items (e.g., "Enjoy", "Fun", and "Like" for items that contribute to "affect")

estimated. The item discrimination for "confidence" ranged between 1.26 and 1.54; for "affect", item discrimination ranged between 1.28 and 3.44.

Attribute Parameters The attribute-level ECDM results are presented in Table 10.3, where parameters indicate significant effects as predictors shifting attribute difficulty (b_k). Adjacent to the ECDM results, the two-step covariate RDINA results are also presented. To compare the simultaneous estimation provided by the explanatory RDINA and the two-step covariate RDINA models, parameter estimates are presented together.

Estimates varied by the type of model (ECDM versus two-step covariate RDINA) used. Differences between the ECDM and the covariate models were larger when the predictor was based on a latent variable ("affect" and "confidence"). The largest difference in attribute difficulty was in attribute 2 (Fractions & Number Patterns). The structural parameters remained similar for h_{k1} (calculator) and for h'_{k3} (confidence), while the estimates themselves varied. Estimates for the observed predictor (owning a calculator) and for the latent variable (having higher self confidence in mathematics) had significant effects for attributes 1 (Whole Numbers), 2 (Fractions & Number Patterns), and 4 (Dimensions & Locations). However,

Table 10.3 Attribute difficulty (b_k) and predictors (h_{kq})

Attribute	Parameter	Explanatory RDINA			Covariate RDINA		
		Estimate		p-value	Estimate		p-value
1. Whole numbers	b_1	−1.26	(.48)		−.71	(.13)	
	h_{11} (calculator)	3.82	(.80)	< .001	1.36	(.14)	< .001
	h'_{12} (affect)	−5.30	(1.55)	.001	−.27	(.08)	.001
	h'_{13} (confidence)	6.73	(2.15)	.002	.60	(.07)	< .001
2. Fractions and number patterns	b_2	−.20	(.54)		.78	(.25)	
	h_{21} (calculator)	1.74	(.58)	.003	1.05	(.24)	< .001
	h'_{22} (affect)	−2.21	(.52)	< .001	−.08	(.14)	.600
	h'_{23} (confidence)	2.96	(.71)	< .001	.60	(.14)	< .001
3. Lines and angles	b_3	5.65	(2.23)		5.34	(2.72)	
	h_{31} (calculator)	−1.07	(2.08)	.610	−.70	(2.84)	.810
	h'_{32} (affect)	.95	(.82)	.250	.24	(.84)	.770
	h'_{33} (confidence)	.41	(.89)	.640	.68	(.86)	.430
4. Dimensions and locations	b_4	−1.38	(.44)		−.63	(.15)	
	h_{41} (calculator)	3.25	(.50)	< .001	1.55	(.15)	< .001
	h'_{42} (affect)	−3.45	(.44)	< .001	−.14	(.09)	.100
	h'_{43} (confidence)	4.17	(.54)	< .001	.47	(.08)	< .001
5. Data display	b_5	1.41	(.53)		2.00	(.46)	
	h_{51} (calculator)	.38	(.45)	.400	.09	(.47)	.850
	h'_{52} (affect)	−.52	(.51)	.310	−.23	(.21)	.280
	h'_{53} (confidence)	.60	(.46)	.190	.28	(.18)	.110

Note:
1. "Explanatory RDINA" used item-level affect and confidence measures as indicators (using GRM and 2PL) to predict α_k. "Covariate RDINA" used estimated covariates to predict α_k
2. "Calculator" is dichotomous (1 = "Yes", 0 = "No"); "affect" and "confidence" are continuous (Mean = 0, SD = 1)
3. Values in parenthesis are standard errors
4. Given that h'_{k2} (affect) estimates were negative, a separate analysis treating the three affect items to directly regress on the attributes was conducted. Results showed that item 3 ("I like mathematics") had negative effects on attributes for all five attributes, while the effects for the two remaining items were either positive or negative depending on the attribute

estimates for "affect" had negative effects on attribute mastery for attributes 1, 2, and 4.

With respect to differences between models, the latent measure of affect, as estimated using the explanatory model, was significant for attributes 2 (Fraction & Number Patterns; $h'_{22} = -2.21$, $p < .001$) and 4 (Dimensions & Location; $h'_{42} = -3.45$, $p < .001$), whereas it was not significant for the covariate model (Fraction & Number Patterns: $h_{22} = -.08$, $p = .600$; Dimensions & Location: $h_{42} = -.14$, $p = .100$), indicating differential attribute difficulty due to the inclusion of the measurement model, as opposed to only a structural component to the model.

Classification and Latent Class Sizes P_c estimates based on posterior probabilities ranged between .86 and .98, with similar estimates between the explanatory and

covariate RDINA models. Latent class sizes for single attributes were similar, except for attribute 2 (Fractions & Number Patterns), which was .62 for the explanatory and .82 for the covariate RDINA models.

10.5 Discussion

Incorporating observed and latent predictors into an ECDM addresses both methodological and substantive issues that arise when analyzing large-scale data. Educational data often contain rich array of variables that can provide relational information and can promote answering important substantive questions for applied educational and psychological research. For example, access to educational technology (e.g., computer and Internet access) has been shown previously to be positively associated with computational ability of students in elementary and middle school mathematics and science performance (e.g., Chang & Kim, 2009; Tienken & Wilson, 2007). Therefore, linking explanatory variables with CDMs can inform instructional needs and facilitate identifying the effectiveness of educational resources and interventions that can be aligned with the fine-grained mastery of skills.

This chapter presents an overview of the ECDM as well as implications for estimation, model fit, and analysis of real-world data. While this chapter focused on explanatory factors that are specified on attributes to examine attribute mastery, predictors can be specified on items to study DIF or on the higher-order latent trait, depending on the context and substantive needs of the analysis. As such, additional studies on explanatory factors specified on items or higher-order latent trait would provide additional understanding of ECDMs.

Estimation results as noted in the simulation study showed consistency in the recovery of model parameters even with sample sizes of 1000. Moreover, only modest differences were found between the 3- and 5-attribute conditions, and attribute-level estimates were stable. Generally, results from the simulation study indicate that incorporating both latent and observed predictors at the attribute level of the RDINA did not affect estimation greatly. The simulation results did also highlight caution when multiple predictors are estimated in the ECDM, particularly when more than one latent explanatory factor is used. Classification was also consistent, regardless of conditions examined. These simulation results demonstrating larger bias in parameter recovery when multiple latent variables are simultaneously estimated, indicate a need to investigate the effect of multiple latent predictors when large number of attributes are used in CDMs.

In the real-world data analysis application, two sets of items (dichotomous and polytomous) were used as indicators of the latent construct, by fitting IRT models (2PL and GRM), and as such, were simultaneously specified to predict attribute mastery. Results revealed differences in effect sizes by attribute. In addition, real-world data example demonstrated varied results when compared to a two-step

approach (covariate model), where latent variables were first estimated and then used to regress as covariates on the attributes. In particular, estimates from the two-step approach were different for some attributes. These findings underscore the value of simultaneously estimating latent predictors, as specified in the ECDM. These findings shed light on related results from the SEM literature (e.g., Bedeian, Day, & Kelloway, 1997), where the ECDM can provide meaningful implications for simultaneously estimating latent factors.

Further extensions of the ECDM proposed in this chapter are also viable. For example, extensions of the ECDM using more generalized CDMs could yield more flexible and generalized estimation of skill profiles to allow specification of explanatory predictors. Other extensions to models such as the von Davier (2010) multilevel diagnostic model or other higher-order CDMs can also be made to incorporate the ECDM framework.

Additional studies may be needed to examine whether the coefficient estimates may have collinearity issues, potentially leading to a model with weakly identified predictors. As such, additional studies examining specific conditions for identification are needed; for example, the effect of centering to enhance estimation may warrant additional study. Such extensions may promote answering other substantive questions and contribute to the development of other explanatory models for CDMs.

References

Bedeian, A. G., Day, D. V., & Kelloway, E. K. (1997). Correcting for measurement error attenuation in structural equation models: Some important reminders. *Educational and Psychological Measurement, 57*, 785–799.

Chang, M., & Kim, S. (2009). Computer access and computer use for science performance of racial and linguistic minority students. *Journal of Educational Computing Research, 40*, 469–501.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). New York, NY: Plenum Press.

Dayton, C. M., & MacReady, G. B. (1988). A latent class covariate model with applications to criterion-referenced testing. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 129–143). New York, NY: Plenum Publications Inc.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

Fox, J.-P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*(2), 169–191.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Huang, G. H., & Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika, 69*, 5–32.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics, 17*(2), 131–154.

Park, H., Lawson, D., & Williams, H. E. (2012). Relationship between technology, parent, education, self-confidence, and academic aspiration of Hispanic immigrant students. *Journal of Educational Computing Research, 46*(3), 255–265.

Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates: Examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38*, 376–390.

Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. *Applied Psychological Measurement, 42*(5), 376–392.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph Supplement, No. 17).

Tienken, C. H., & Wilson, M. J. (2007). The impact of computer assisted instruction on seventh-grade students' mathematics achievement. *Planning and Changing, 38*, 181–190.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for latent GOLD 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations, Inc.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2007). *Mixture general diagnostic models* (Research Report, RR-07-32). Princeton, NJ: ETS.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical & Statistical Psychology, 61*, 287–307.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*(1), 8–28.

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large scale assessment with a general latent variable model. *Psychometrika, 76*(2), 318–336.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (RR-08-27, ETS Research Report).

Chapter 11
Insights from Reparameterized DINA and Beyond

Lawrence T. DeCarlo

Abstract The purpose of cognitive diagnosis is to obtain information about the set of skills or attributes that examinees have or do not have. A cognitive diagnostic model (CDM) attempts to extract this information from the pattern of responses of examinees to test items. A number of general CDMs have been proposed, such as the general diagnostic model (GDM; von Davier M, Brit J Math Stat Psychol 61:287–307, 2008), the generalized DINA model (GDINA; de la Torre J, Psychometrika 76:179–199, 2011), and the log-linear cognitive diagnostic model (LCDM; Henson RA, Templin JL, Willse JT, Psychometrika 74:191–210, 2009). These general models can be shown to include well-known models that are often used in cognitive diagnosis, such as the deterministic inputs noisy and gate model (DINA; Junker BW, Sijtsma K, Appl Psychol Meas 25:258–272, 2001), the deterministic inputs noise or gate model (DINO; Templin JL, Henson RA, Psychol Methods 11:287–305, 2006), the additive cognitive diagnosis model (ACDM; de la Torre J, Psychometrika 76:179–199, 2011), the linear logistic model (LLM; Maris E, Psychometrika 64:187–212, 1999), and the reduced reparameterized unified model (rRUM; Hartz SM, A Bayesian framework for the unified model for assessing cognitive abilities. Unpublished doctoral dissertation, 2002).

This chapter starts with a simple reparameterized version of the DINA model and builds up to other models; all of the models are shown to be extensions or variations of the basic model. Working up to more general models from a simple form helps to illustrate basic aspects of the models and associated concepts, such as the meaning of model parameters, issues of estimation, monotonicity, duality, and the relation of the models to each other and more general forms. In addition, reparameterizing CDMs as latent class models allows one to use standard software for latent class analysis (LCA), which offers a connection to latent class analysis and also allows one to take advantage of recent advances in LCA.

L. T. DeCarlo (✉)
Department of Human Development, Teachers College, Columbia University, New York, NY, USA
e-mail: decarlo@tc.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_11

11.1 The Reparameterized DINA Model

A well-known CDM, the DINA model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977), provides a useful starting point. Let Y_{ij} be a binary variable that indicates whether the response of the ith examinee to the jth item is correct or incorrect (1 or 0) and let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K)$' denote the vector of K attributes that are needed to solve the items. The Q-matrix consists of elements q_{jk} that specify which of the K attributes are needed to solve the jth item. Thus, the Q-matrix elements consist of zeroes and ones, with a value of zero indicating that the kth attribute is *not* needed, and a value of one indicating that the attribute is needed. For the DINA model, the probability that an examinee gets an item correct is

$$p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}},$$

with

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}.$$

Note that η_{ij} is simply a binary indicator with a value of one indicating that an examinee has all of the required attributes and a value of zero indicating that an examinee is lacking one or more of the required attributes. Thus, if an examinee has all of the required attributes, then $\eta_{ij} = 1$ and

$$p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = \left(1 - s_j\right),$$

where the parameter s_j is the *slip* rate for examinee j. If an examinee is missing one or more of the required attributes, then $\eta_{ij} = 0$ and

$$p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = g_{j,}$$

where the parameter g_j is the *guess* rate.

Although 'slipping' and 'guessing' were suggested as useful mnemonics by Junker and Sijtsma (2001), the relation of the concepts to basic ideas of *signal detection theory* (SDT) is also informative (Green & Swets, 1966; Macmillan & Creelman, 2005; Wickens, 2002). In SDT, $(1 - s_j)$ is the *hit* rate – the examinee has the requisite attributes and gets the item correct, whereas s_j is the *miss* rate – the examinee has the requisite attributes, but gets the item incorrect. If an examinee does not have the requisite attributes yet gets the item correct, then that's a *false alarm*; note that guessing is an interpretation of false alarms.

It has previously been shown that a simple re-parameterization of the DINA model can be obtained by re-writing the false alarm rate, g_j, as

$$g_j = \frac{\exp\left(f_j\right)}{1 + \exp\left(f_j\right)},$$

where *exp* is the exponential function and the parameter f_j is the transformed false-alarm rate. Similarly, one minus the slip rate, the hit rate, can be re-written as

$$1 - s_j = \frac{\exp\left(f_j + d_j\right)}{1 + \exp\left(f_j + d_j\right)},$$

where d_j is a discrimination parameter. The DINA model can then be re-written as

$$\text{logit } p\left(Y_{ij} = 1 | \boldsymbol{\alpha}\right) = f_j + d_j \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \tag{11.1}$$

which has been referred to as the reparameterized DINA model (rDINA; DeCarlo, 2011). The model is a special case of the general diagnostic model of von Davier (2008), with a change in notation to emphasize the signal detection aspects of the model. In particular, f_j is a transformed *false alarm rate* whereas d_j is a difference of transformed hit and transformed false alarm rates that indicates the level of *discrimination* between having and not having the attributes (i.e., $\eta_{ij} = 1$ versus $\eta_{ij} = 0$). Note that the discrimination parameter in SDT is a function of both hits and false alarms, in contrast to looking at slips and guesses separately, in that it follows from the theory that one needs to examine the hit rate relative to the false alarm rate (to get the distance measure d); other variations of the discrimination parameter have also been considered (e.g., differences, ratios, etc.). From a statistical point of view, the intercept f_j is the *log odds* of a correct response given $\eta_{ij} = 0$, and the slope d_j is the *log odds ratio* of a correct response given $\eta_{ij} = 1$ versus $\eta_{ij} = 0$. The rDINA model is equivalent to the DINA model and so estimates of f_j and d_j can be transformed to get estimates of g_j and s_j.

11.2 Monotonicity

Note that, when fitting the model, a constraint must be used so that *monotonicity* is satisfied. In terms of the DINA model, monotonicity holds if

$$0 < g_j < 1 - s_j < 1.$$

Without the monotonicity constraint, examinees who have a required attribute could have a lower probability of getting an item correct than if they did not have the attribute (although this could be appropriate in certain situations). It is informative to interpret the monotonicity constraint in terms of signal detection theory, in that it simply implies that the hit rate $(1 - s_j)$ must be greater than the false alarm rate (g_j), in which case the receiver operating characteristic curve (ROC; see Macmillan & Creelman, 2005) – a plot of hits versus false alarms – will lie above the diagonal

(the diagonal represents zero discrimination) and the discrimination parameter d will be greater than zero. In terms of the rDINA model, monotonicity holds if

$$0 < \frac{\exp\left(f_j\right)}{1+\exp\left(f_j\right)} < \frac{\exp\left(f_j+d_j\right)}{1+\exp\left(f_j+d_j\right)} < 1,$$

and so monotonicity holds if the discrimination parameter d_j is greater than zero (for finite f_j and d_j). This constraint can be implemented in many software packages; the Appendix, for example, provides Latent Gold programs that show how to implement monotonicity by using the (+) command, which constrains the parameter to non-negative values.

11.3 Estimation

A benefit of the rDINA model as written in Eq. (11.1) is that it is simple to fit with standard software for latent class analysis, such as Latent Gold (LG; Vermunt & Magidson, 2016) or the freely available LEM (Vermunt, 1997), given that Eq. (11.1) is simply a logistic latent class model with latent dichotomous interaction terms. The program provided in the Appendix shows that it is straightforward to fit the rDINA model in LG by specifying interaction terms for the latent dichotomous attributes; the Q-matrix being used is also made transparent in the program.

Note that the rDINA model discussed so far is the 'examinee-level' part of the model, whereas the complete model also includes a higher-level model for the attributes, that is, an 'attribute-level' model. This is a model for the probabilities of the various skill combinations, that is, $p(\alpha_1, \alpha_2, \ldots, \alpha_K)$. LG and LEM programs to fit the rDINA model with either an independence structure or a higher-order structure (with a continuous latent variable) for the attribute-level model have previously been provided (DeCarlo, 2011). Here it is shown how to specify an *unstructured* attribute-level model (not previously shown) in LG or other latent class software. Note that when CDM researchers refer to fitting 'the DINA model', they usually mean the DINA model with an unstructured attribute-level model.

To implement the unstructured attribute model one merely needs to use a *saturated model* as the higher-level model. A simple way to do this in LG is to specify a saturated *association model* (Agresti, 2002) for the attribute-level model, with one parameter set to zero, so that the model includes $2^k - 1$ parameters for the 2^k attribute patterns (and so it is saturated). An example of this approach is given by the first rDINA LG program provided in the Appendix. In this case, the first cell is restricted to be zero, using the command r[1,1] = 0, for identification. Estimates of the class sizes for each attribute pattern, and their standard errors, are given in the section of LG output that is labeled as "Profile", along with estimates of the marginal class sizes, that is, estimates of $p(\alpha_k)$.

Another option is to specify a sequence of path models, which gives the same results as when an association model is used, given that both models are saturated. The LG program in the Appendix also illustrates this approach (in the comments, which are specified in LG by the symbol '//'); running the program shows that the results are the same as those obtained using the saturated association model. Depending on the software that one uses, one or the other of the approaches for the attribute-level model might be simpler to implement.

Philipp, Strobl, de la Torre, and Zeilis (2018) recently noted that there is a problem with respect to estimation of the standard errors in CDMs. In particular, "it is common to compute the standard errors only for the parameters that are used to specify the item response function while ignoring the parameters used to specify the joint distribution of the attributes." (p. 2). They note that this common approach can lead to underestimation of the standard errors in both parts of the model (also see von Davier, 2014). Note that, with the LCA approach, the standard errors are estimated for both the examinee-level parameters (i.e., the item response function) and the attribute-level parameters. LG also offers a robust (sandwich) estimator of the SEs, as well as others, details of which are given in the technical manual (Vermunt & Magidson, 2016).

Using software for LCA also makes available a wide array of tools and output for CDMs. For example, with LG, one obtains estimates of the parameters for both the examinee-level and attribute-level models, along with their standard errors, absolute and relative fit statistics (e.g., Chi-square goodness of fit, information criteria such BIC and AIC, etc.), bivariate residuals, various classification statistics and tables, different types of plots, output files with posterior classifications for each examinee, as well as details about the iterations and any convergence or identification problems. In addition, different algorithms are available, such as versions of the Newton-Raphson (NR) and Expectation-Maximization algorithms (LG starts with the EM and moves to NR when in the vicinity of the solution), as well as the option to use posterior mode estimation with different Bayes constants, which controls the degree of smoothing, along with many other options.

11.4 Boundary Problems

A well-known problem that often arises in latent class analysis is known as a *boundary problem* (Clogg & Eliason, 1987; DeCarlo, 2011; Maris, 1999). Boundary problems occur when parameter estimates and SEs are large or indeterminate, or probability estimates are close to zero or one, which is also related to identification problems (such as weak identification). This problem has been somewhat neglected in CDMs (some exceptions are noted below), partly because the SEs are sometimes not reported, and partly because, in the original probability version of the model, finding slipping or guessing parameters close to zero, for example, is not in and of itself cause for alarm (not having slipping or guessing can be viewed as a good thing), whereas it could actually be reflecting an overlooked boundary problem.

The reparameterized model is useful in this regards because the model transforms the zero-one probability scale (for g_j and s_j) to a minus infinity to positive infinity scale for f_j and greater than zero to infinity scale for d_j (because of the monotonicity constraint), and so boundary problems or weak identifiability will tend to be more obvious, in that they will appear as overly large or infinite parameter estimates and/or estimated standard errors. For example, Table 4 of de la Torre (2009) shows parameter estimates for a fit of the DINA model to a subset of 15 items (out of 20) of the well-known fraction subtraction data. The estimate of g_1 for the first item is shown as 0.00 with a standard error of 0.05 (which is the largest standard error in the table) and the estimate of s_1 is 0.28 (the highest in the table) with a standard error of 0.013. A fit of the rDINA model with LG to this data (with unstructured attributes) gives an estimate of f_1 of about -24 (whereas the lowest f_j for all the other items is around -4.5) with an SE of 0.12, and an estimate of d_1 of 25 with an SE of 1000 (i.e., infinite), and so there are clearly identification problems for this item.

In addition to boundary problems appearing in the examinee-level part of the model, they can also appear in the attribute-level part of the model, particularly when the unstructured attribute model is used. For the fraction subtraction example with 15 items just discussed, LG shows that there are 22 boundary problems in the unstructured attribute model (with 22 SEs appearing as 1000). It is interesting to note that if one fits the original 20 item version of the fraction subtraction data with an unstructured attribute model, as has been widely used in many studies, then LG reports that there are 198 non-identified parameters (note that the unstructured model has $2^8 - 1 = 255$ parameters; problems also appear for a higher-order model). Estimates of the standard errors for the attribute-level model parameters are also all excessively large (>20), again reflecting identification problems. Thus, even though the unstructured attributes model has been widely used, identification problems with the attribute-level model, and the possible effects of this on estimation for the examinee-level model, have generally not been considered.

Boundary problems for fits of the DINA model, ACDM, and GDINA model to real-world data given in the R package *pks* (Heller & Wickelmaier, 2013) were recently noted by Philipp et al. (2018); they noted boundary problems in both the examinee-level and attribute-level parts of the model (p. 20). von Davier (2014) discussed identification problems for the well-known Examination for the Certificate of Proficiency in English (ECPE) data; he noted that, even with constraints, weak identifiability still appeared for the LCDM used by Templin and Bradshaw (2014).

11.5 Posterior Mode Estimation and Bayesian Estimation

A number of authors have discussed the use of *posterior mode estimation* (PME) to deal with boundary problems (e.g., DeCarlo, 2011; DeCarlo, Kim, & Johnson, 2011; Maris, 1999; Vermunt & Magidson, 2016). PME is less computationally intense than

a full Bayesian analysis in that it does not require that the full posterior distribution be generated, but rather only the mode needs to be found. As a result, PME has a computational speed advantage over a full Bayesian analysis, which is useful when performing computer simulations. In addition, standard algorithms that implement maximum likelihood estimation can often easily be modified to implement PME. The approach basically smooths infinite or large parameter estimates and/or estimates of the standard errors. The use of PME in CDMs is a topic for future research; this option is currently available in LG. The use of PME for a simple latent class signal detection model (DeCarlo, 2002, 2005), which is the same as a CDM with a single latent dichotomous attribute (the latent signal), has been examined in simulations presented in DeCarlo (2008, 2010); the use of PME for a hierarchical rater signal detection model with ordinal latent classes has been examined in simulations presented in Kim (2009), and the use of PME with real-world data was examined in DeCarlo et al. (2011).

Another option is to use a full Bayesian analysis to fit CDMs (e.g., Culpepper, 2015; de la Torre & Douglas, 2004; DeCarlo, 2012; Henson et al., 2009). For example, an OpenBugs program (Spiegelhalter, Thomas, Best, & Lunn, 2014) to fit the rDINA model with Bayesian estimation was given in DeCarlo (2012; with a monotonicity constraint implemented by restricting d_j to be greater than zero). The approach also generalized the model by allowing for uncertainty about some elements of the Q-matrix, and simulations suggested adequate recovery of those elements using posterior distributions. The Bayesian approach allows for interesting extensions; for example one can extend the model with a few uncertain Q-matrix elements to allow all of the Q-matrix elements to be uncertain; this approach was examined by DeCarlo and Kinghorn (DeCarlo & Kinghorn, 2016; with monotonicity restrictions) and by Culpepper (2015; with completeness restrictions).

11.6 Classification

Classification in latent class analysis is typically done using the modal posterior probabilities (e.g., Clogg, 1995; Dayton, 1998). For example, one approach, *maximum a posteriori* (MAP) classification, is to simply classify each examinee into the attribute set with the largest posterior probability. Another option is to use marginal probabilities to classify examinees for each skill separately, as in *expected a posteriori* (EAP) classification. In maximum likelihood estimation (MLE), classification is accomplished by finding attribute patterns that maximize the posterior. The various approaches were compared in the context of CDMs by Huebner and Wang (2011).

11.7 Identifiability

Identifiability is concerned with whether one can obtain unique estimates of the model parameters. Xu and Zhang (2016) gave necessary and sufficient conditions for identifiability of the model parameters for the DINA model (also see Chen, Liu, Xu, & Ying, 2015). They also noted that their results could be extended to the DINO model, because of the duality of the models (see below). The issue of boundary problems discussed above is also related to the issue of identifiability, with large standard errors often indicating 'weak identification', in which case the data (or model) are not informative about the parameters.

The effect of identifiability on classification has also been discussed. Chiu, Douglas, and Li (2009) noted that *completeness* of a Q-matrix is generally needed for identification of all possible attribute patterns. For the DINA and DINO models, for example, completeness is satisfied if, for each attribute, there is an item that measures that attribute alone. Köhn and Chiu (2017) noted that the conditions for completeness depend on the model and examined completeness for several CDMs.

For the fraction subtraction data, DeCarlo (2011) noted that, because of incompleteness of the Q-matrix, some of the posterior classifications from the DINA model depend solely on the priors, and so the data offers no additional information over the priors. Zhang, DeCarlo, and Ying (2013) noted that, although certain attribute patterns are in the same equivalence class for the fraction-subtraction data and so are not identifiable, individual attributes within an equivalence class may still be identifiable. They proposed a measure of the *marginal identifiability rate*, which is the proportion of the population for which each attribute is marginally identifiable, and suggested that it can be viewed as a measure of test (and model) quality. Zhang et al. also proposed classification algorithms that took into account the effects of marginal identifiability.

11.8 Reparameterized DINO

Here it is shown that a reparameterized version of the DINO model (Templin & Henson, 2006) clarifies issues about the relations between the DINO and DINA models (duality) and their parameters. The DINO model is similar to the DINA model with the exception that, instead of requiring that all of the skills be present in order to solve an item, only one or more of the skills need to be present. The DINO condensation rule is usually referred to as being *disjunctive*, whereas the DINA condensation rule is *conjunctive* (Rupp, Templin, & Henson, 2010). The model is

$$p\left(Y_{ij}=1|\boldsymbol{\alpha}\right)=\left(1-s_j'\right)^{\omega_{ij}}{g'}_j^{1-\omega_{ij}},$$

where $\omega_{ij}=1-\prod_{k=1}^{K}\left(1-\alpha_{ik}\right)^{q_{jk}}$ equals one if any required skill is present, and zero only if all the required skills are absent.

It is important to note that slips and guesses, s'_j and g'_j, are defined differently in the DINO model as compared to the DINA model. In the DINO model, the hit rate $1-s'_j$ is the probability of a correct response given that an examinee has *at least one* of the attributes, whereas in the DINA model, the hit rate $1-s_j$ is the probability of a correct response given that an examinee has *all* of the attributes. Similarly, the false alarm rate g'_j in the DINO model is the probability of a correct response given that an examinee has *none* of the attributes, whereas the false alarm rate g_j in the DINA model is the probability of a correct response given that an examinee is *missing at least one* attribute.

The DINO model can be reparameterized using the same approach used above for the DINA model, which gives,

$$\text{logit}\, p\left(Y_{ij}=1|\boldsymbol{\alpha}\right)=f'_j+d'_j\left[1-\prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}}\right], \tag{11.2}$$

which will be referred to as the *rDINO* model; the model was also recently derived in terms of the GDM by Köhn and Chiu (2016). Once again, monotonicity is satisfied if d'_j is greater than zero.

As for the rDINA model, the model in this form is straightforward to fit using software for latent class analysis. Suppose, for example, that Item 1 requires the first three skills. The model for the first item is then

$$\text{logit}\, p\left(Y_{i1}=1|\boldsymbol{\alpha}\right)=f'_1+d'_1\left(\alpha_{i1}+\alpha_{i2}+\alpha_{i3}-\alpha_{i1}\alpha_{i2}-\alpha_{i1}\alpha_{i3}-\alpha_{i2}\alpha_{i3}+\alpha_{i1}\alpha_{i2}\alpha_{i3}\right),$$

which is a logistic model with all main effects and higher order interaction terms. Further, the coefficients $\left(d'_j\right)$ are restricted to be equal across all terms and have alternating signs across the two and three way interactions, as was also noted by de la Torre (2011) for the G-DINA model. A sample rDINO program that shows how to implement the parameter constraints of Eq. (11.2) in LG is given in the Appendix.

11.9 DINO/DINA Duality

The rDINO model of Eq. (11.2) can be re-written as.

$$\text{logit}\, p\left(Y_{ij}=1|\boldsymbol{\alpha}\right)=\left(f'_j+d'_j\right)-d'_j\prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}}. \tag{11.3}$$

Note that, if one replaces $1-\alpha_{ik}$ in the above with reverse coded $\alpha^*_{ik}=1-\alpha_{ik}$, then Eq. (11.3) has the same form as the rDINA model of Eq. (11.1), with a redefined intercept and a negative slope. This was also shown by Köhn and Chiu

(2016, see Section 3.3) by replacing $1 - \alpha_{ik}$ with α^*_{ik} and by reverse coding the data, so that $Y^*_{ij} = 1 - Y_{ij}$ which merely reverses the signs of Eq. (11.3), given that logit $p = -$logit $(1 - p)$, and so.

$$\text{logit } p\left(Y^*_{ij} = 1|\boldsymbol{\alpha}^*\right) = \left(-f'_j - d'_j\right) + d'_j \prod_{k=1}^{K} \left(\alpha^*_{ik}\right)^{q_{jk}}. \tag{11.3a}$$

Equation (11.3a) is clearly related in form to the rDINA model of Eq. (11.1), with a redefined intercept (and different parameters). In this respect, there is a *duality* between the rDINA and rDINO models (and so between the DINA and DINO models as well; Chen et al., 2015; Köhn & Chiu, 2016; Liu, Xu, & Ying, 2011).

An important consequence of the duality between the DINA and DINO models (and rDINA and rDINO) is that theoretical results developed for one model can be applied to the other model (Liu et al., 2011). For example, Köhn and Chiu (2016) used this duality to determine the conditions necessary for completeness of the Q-matrix for both the DINA and DINO models.

Köhn and Chiu (2016) noted another interesting consequence of duality, which is that it implies that the DINO model can be fit by using a DINA program. The simple reparameterized versions of the models presented here are helpful in that they suggest more than one way that this can be done. To start, note that the rDINO model can be fit directly as given in Eq. (11.2), as shown by the rDINO program given in the Appendix. In this case, the program is a little more involved than the rDINA program because of the parameter restrictions implied by the rDINO model (i.e., equal d'_j and alternating signs).

Equation (11.3a) suggests another option, also suggested by Köhn and Chiu (2016), which is to use a DINA program to fit the DINO model. This can be done if one can fit the DINA model with reverse coded α^*_{ik} in lieu of α_{ik}, which is the key to the difference between the models. Köhn and Chiu (2016) accomplished this by reverse coding the data and maintaining the monotonicity constraint. Note that, for symmetric links such as the logit, reverse coding the data simply reverses the parameter signs. However, because the monotonicity constraint is also maintained, the model cannot account for the reversed Y with a negative sign for the discrimination parameter, but rather with a reversed α (i.e., α^*). Thus, reverse coding the data and maintaining monotonicity is simply a way to induce the use α^* in the model in lieu of α.

A practical advantage of the above approach is that one can then fit the rDINO model using the simpler rDINA program given in the Appendix by reverse coding the data and keeping the monotonicity constraint (i.e., positive values of d'_j). Note that if the monotonicity constraint is removed, then fitting the reverse coded data will simply give results for the rDINA model with reversed parameter signs, and not the rDINO model, as the reader can verify.

Another interesting option is suggested by Eq. (11.3) – fit the rDINO model with an rDINA program, but impose a *negative monotonicity* constraint, that is, restrict d'_j in Eq. (11.3) to be less than zero. An interesting aspect of this approach is that it again allows one to fit the rDINO model with an rDINA program, but

there is *no need to reverse code the data*. That is, one can fit the original data, again using an rDINA program, by simply replacing (+) in the LG program given in the Appendix with (−), to give a negative monotonicity constraint. Specifying a negative monotonicity constraint will tend to lead to α_{ik}^{*} being used in the model in lieu of α_{ik}, in which case the rDINO model of Eq. (11.3) is fit (and not the rDINA model).

It is apparent that the simplest approach in LG is the third one – simply use the original data and impose a negative monotonicity constraint in an rDINA program to get the rDINO model of Eq. (11.3). It would be interesting in future research to see if there are any differences across the three approaches to fitting the rDINO model, in terms of estimation advantages or disadvantages.

It should be noted that using a negative monotonicity constraint or reverse coding the data and using a positive monotonicity constraint may not be sufficient to lead to α_{ik}^{*} being used in the model in lieu of α_{ik} (this is also related to 'label switching' issues discussed in latent class analysis, although it is not simply label switching in this case in that the likelihood differs, but this is beyond the scope of the current chapter), this needs to be considered more closely in future research. For example, if Eq. (11.3a) is used, then one must check that all of the d'_j estimates have positive signs, so that monotonicity holds. Another useful check is to compare the latent class size estimates, that is, the estimates of $p(\alpha_k)$ for the rDINO model, to those obtained for a fit of the rDINA model – the class size estimates will usually differ (beyond a simple reversal in categories). If they are the same, apart from a category reversal, then it is likely that the rDINA model was fit, not the rDINO model.

Equations (11.2), (11.3), and (11.3a) are useful in that they also show exactly what estimates are obtained with each of the three approaches. That is, if one fits the rDINO model as specified in Eq. (11.2), then the intercept and slope of the logistic model will give direct estimates of f'_j and d'_j respectively; the estimates of $p(\alpha_k)$, the latent class sizes, will be given in LG as Class Size 2 (i.e., the class size for having the attribute). If the data are reverse coded and Köhn and Chiu's (2016) approach is used, then the intercept gives an estimate of $-f'_j - d'_j$ and the slope gives an estimate of d'_j (see Eq. (11.3a)). Thus, one must add the estimate of d'_j to the intercept and then reverse the sign to get an estimate of f'_j. The latent class sizes will also be reversed, so that Class Size 1 in LG will give the class size for having the attribute. Finally, if one fits rDINO with Eq. (11.3), then the intercept will give an estimate of $f'_j + d'_j$ and the slope will give an estimate of $-d'_j$. Thus, one simply adds the slope estimate to the intercept estimate to get an estimate of f'_j and reverses the sign of the slope to get d'_j. The latent class sizes will again be reversed, and so Class Size 1 again gives the desired estimate. It is instructive to use the three approaches and compare the results; simulated DINO data is available at the author's website, along with LG programs, so that the three approaches can be implemented and compared.

The equations also help to clarify similarities and differences between the DINA and DINO models. First, the relation between Eqs. (11.1) and (11.3a) does not imply that the rDINA model and the rDINO model are *equivalent*; they can and will give

different log likelihoods when fit to the same data (and different log posteriors, in a Bayesian approach), given that they impose a different structure on the data (for items that involve two or more attributes). Note that Köhn and Chiu (2016; Section 3.3) showed that the expected item response function for DINO with Y^* and α^* is equivalent to that for DINA with Y and α, as also shown by Eq. (11.3a). This does not mean however that the DINO model is equivalent to the DINA model. Rather, it should be recognized that if one uses a DINA program to fit the model with α^*_{ik}, regardless of whether or not Y is reverse coded, then the DINO model is being fit, as shown by Eqs. (11.3) and (11.3a), and not the DINA model. That is, using notation suggested by a reviewer, Eqs. (11.3) and (11.3a), respectively, show that.

$$\text{rDINO}\left(Y, \alpha\right) = \text{rDINA}\left(Y^*, \alpha^*\right) = \text{rDINA}\left(Y, \alpha^*\right),$$

with the parameters related as shown above. This is what allows one to use a DINA program to fit the DINO model – use α^* in the DINA model in lieu of α, irrespective of whether Y or Y^* is used, and the DINO model is being fit. However, this does not mean that the DINO model is equivalent to the DINA model, that is,

$$\text{rDINO}\left(Y, \alpha\right) \neq \text{rDINA}\left(Y, \alpha\right)$$

as can be seen by comparing Eqs. (11.1) and (11.2) (for items that load on two or more attributes). Thus, rDINO and rDINA (and DINO and DINA) are structurally different models, the above just shows that using α^* in place of α in a DINA program, and maintaining monotonicity, results in the DINO model being fit. A useful exercise is to use the rDINA LG program given in the Appendix with reverse coded data, but remove the monotonicity constraint; the result is that the rDINA model is fit, not the rDINO model, and the parameter signs are simply reversed from those obtained for a fit of the rDINA model to the original data. On the other hand, if the monotonicity constraint is enforced for the reverse coded data, then the rDINO model will be fit, not the rDINA model, and the parameter estimates and likelihood will differ.

Another point of clarification is that the models also differ with respect to the parameter estimates, that is, the DINO parameter estimates are not transformations of the DINA parameter estimates (contrary to some claims). The models are structurally different and involve different parameters. Table 11.1 shows a simple example with two attributes. The third column shows the condensation rule η_j for the DINA model and the fourth column shows the condensation rule ω_j for the DINO model. The fifth column shows the DINA parameters for a correct response. The column shows that, for the DINA model, false alarms, g_j, occur for the first three rows, whereas the last row represents hits, $1 - s_j$. The sixth column shows the DINO parameters; in this case, only the first row represents false alarms, g'_j, whereas the other rows all correspond to hits, $1 - s'_j$.

A comparison of the second and third rows of Table 11.1, where only one of the skills is present, helps to highlight differences between the models. For DINA, the

Table 11.1 Relation between terms and parameters for the DINA and DINO models

α_1	α_2	η_j	ω_j	DINA	DINO
0	0	0	0	g_j	g'_j
1	0	0	1	g_j	$1 - s'_j$
0	1	0	1	g_j	$1 - s'_j$
1	1	1	1	$1 - s_j$	$1 - s'_j$

Table notes: for DINA, $\eta_j = \prod_{k=1}^{K} \alpha_k^{q_{jk}}$; for DINO, $\omega_j = 1 - \prod_{k=1}^{K} (1-\alpha_k)^{q_{jk}}$; the DINA and DINO columns show the parameters that correspond to a correct response

second and third row parameters are the same as the first row, that is, they are all false alarms g_j. In contrast, for DINO, the second and third row parameters are the same as the fourth row, the hit rate $1 - s'_j$. Thus, different parameter estimates will generally be obtained for fits of the two models (for items that involve two or more attributes) and one set of parameters are not simply transformations of the other set, that is, $g'_j \neq g_j$ and $1 - s'_j \neq 1 - s_j$.

To summarize, if one uses a DINA program and induces the use of α^* in place of α, either by reverse coding the data or enforcing negative monotonicity, then one is fitting the DINO model and not the DINA model.

11.10 A General Reparameterized Model

The rDINA and rDINO models show a clear and simple pattern. Consider the rDINA model for an item that requires three skills,

$$\text{logit } p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = f_j + d_j\, \alpha_{i1}\alpha_{i2}\alpha_{i3},$$

whereas the corresponding rDINO model is

$$\text{logit } p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = f'_j + d'_j\, (\alpha_{i1} + \alpha_{i2} + \alpha_{i3} - \alpha_{i1}\alpha_{i2} - \alpha_{i1}\alpha_{i3} - \alpha_{i2}\alpha_{i3} + \alpha_{i1}\alpha_{i2}\alpha_{i3})\,,$$

and similarly for the other items. It is clear that the rDINA model only includes the highest-order interaction term whereas the rDINO model also includes main effects and lower order interaction terms. Further, rDINO restricts the coefficients of all the terms to be equal and the signs to be alternating.

It is immediately obvious that both the rDINA and rDINO models, as well as others, are simply special cases of a more general model that includes all main effects and higher order interactions. For example, for the above item with three attributes, a general reparameterized model is.

$$\begin{aligned} \text{g}\left[\, p\left(Y_{ij} = 1|\boldsymbol{\alpha}\right)\right] = f_j + d_{j1}\alpha_{i1} + d_{j2}\alpha_{i2} + d_{j3}\alpha_{i3} + d_{j,12}\alpha_{i1}\alpha_{i2} + d_{j,13}\alpha_{i1}\alpha_{i3} \\ + d_{j,23}\alpha_{i2}\alpha_{i3} + d_{j,123}\alpha_{i1}\alpha_{i2}\alpha_{i3}, \end{aligned}$$

where g is a link function, such as the logit, probit, or complementary log-log link (all available in LG, along with others). Note that the discrimination parameters are now attribute-specific, that is, d_j in the rDINA and rDINO models is replaced with the attribute-specific d_{jk}, for the first-order terms. For the interaction terms, the discrimination parameter subscripts indicate which attributes are involved; for example, for a three-way interaction term the discrimination parameter is $d_{j,kk'k''}$, where the j indicates the item, as before, and $kk'k''$ indicates the three attributes involved in the interaction (giving $d_{j,\,123}$ in the example above). Applying the model to every item, according to the Q-matrix structure, gives a *general reparameterized model* (GRM), which is simple to fit with software such as LG. The notation makes clear that the added parameters are discrimination parameters that show how (transformed) hits increase compared to (transformed) false alarms.

The GRM with logit link gives a saturated version of the GDM of von Davier (2008) and the LCDM of Henson et al. (2009); with an identity link it is a saturated version of the GDINA model of de la Torre (2011); also see von Davier (2013, 2014). With appropriate parameter restrictions, the GRM includes the rDINA and the rDINO models discussed above. Another simplification is to only include main effects, which gives the linear logistic model (LLM) of Maris (1999); using an identity link gives the ACDM of de la Torre (2011). With constraints placed on the coefficients of the higher-order interaction terms, one can obtain the reduced reparameterized unified model (rRUM; Hartz, 2002), given that Chiu and Köhn (2016) recently showed that rRUM is a (non-saturated) logistic model with parameter constraints. The parameter constraints for rRUM, however, are somewhat complex (and apparently cannot be implemented in LG at this time).

Although an unrestricted saturated model is quite simple to fit in software such as LG, one has to pay close attention to parameter restrictions that might need to be imposed. For example, if the monotonicity constraint is to be satisfied, then the coefficients d_{jk} of single attribute terms should be restricted to be greater than zero. Note that if one fits the model (as GDINA) using 'rule = GDINA2' in the CDM package in R (George, Robitzsch, Kiefer, Gross, & Uenlue, 2016; also see Chap. 26 in this volume), for example, then monotonicity is not enforced, as can be verified using the ECPE data – the first item gives a negative d_{jk} for the first attribute, and so monotonicity does not hold for the first item. The GDINA package in R (Ma & de la Torre, 2017; also see Chap. 29 in this volume) allows one to place monotonicity constraints on the parameters. As before, in LG, non-negativity for d_{jk} is implemented by using the monotonicity constraint (+), whereas negative monotonicity is implemented by using (−), as shown by the programs given in the Appendix.

Another consideration has to do with whether or not restrictions should be placed on the coefficients of the interaction terms. For example, if they are left unrestricted, then it is possible that the probability of a correct response can be lower when an examinee has two required attributes as compared to only one of the attributes. If this is viewed as being theoretically undesirable (although in some cases one could possibly argue for an interference effect) then restrictions should

be placed on the interaction parameters. For example, for an item that requires two attributes, the restriction $d_{j,kk'} > -1*\min(d_{jk}, d_{jk'})$ will ensure that the probability of a correct response when an examinee has both attributes will not be lower than when they only have one of the attributes (namely the one that gives the highest probability of a correct response). The restriction can also be written as $d_{j,kk'} > -d_{jk}$ and $d_{j,kk'} > -d_{jk'}$, which is the form used by Templin and Hoffman (2013) for an implementation of the model in Mplus.

With respect to fitting the model in LG, although there is a way to implement order restrictions (as shown below), the multiple restrictions required above cannot currently be implemented simultaneously (to my knowledge). For example, for items that require two attributes, there are four required restrictions: $d_{j1} > 0$, $d_{j2} > 0$, $d_{j,12} > -d_{j1}$, and $d_{j,12} > -d_{j2}$; three of the four restrictions can be implemented in LG, but not all four. A simple work-around is to use a two-step approach: in the first step, fit the GRM with monotonicity constraints on the first order terms but without restrictions on the interaction terms and examine the parameter estimates; in the second step constrain interaction parameters where necessary (i.e., they violate the above order condition) using information gained in the first step. That is, if one finds that the estimate of d_{j2} is greater than the estimate of d_{j1}, then only the restriction $d_{j,12} > -d_{j2}$ is needed, in addition to the two monotonicity restrictions.

To illustrate the suggested approach, consider the well-known ECPE data, where the saturated LCDM with appropriate restrictions has previously been fit (using Mplus; Templin & Bradshaw, 2014). The first step is to fit a saturated model in LG with a monotonicity constraint on the first-order terms, but unrestricted interaction terms; the program in the Appendix shows that this is very simple to do in LG. The results then allow one to see (1) if and where the above restriction on the interaction term is violated and (2) if it is, which discrimination parameter is smallest, which gives one information about $\min(d_{jk}, d_{jk'})$, and so only three of the four restrictions noted above are needed. For example, for the ECPE data, it was apparent from a fit of the GRM that there were problems with Item 7; the coefficient for the second attribute for this item was also clearly smaller than for the first attribute. Thus, for the second step, the saturated model was fit adding the constraint $d_{j,12} > -d_{j2}$ to Item 7, and the results reproduce those shown in Table 1 and Figure 1 of Templin and Bradshaw (2014) to two decimal places; the log-likelihood was also identical to that obtained with Mplus. This is not to say that the two-step approach will work in general, but the point here is to show possible ways to implement more complex restrictions in current software.

The program given in the Appendix shows how to implement the order constraint for Item 7 in LG; a 'trick' is used of adding a positive constant to the coefficient that must be greater than zero by introducing an additional interaction term that is restricted to be greater than zero, which implements the order constraint.

It should be noted that there is also a computational speed advantage of LG for CDMs, which is useful when conducting simulations. For example, on a machine with 8GB RAM, 2.30 GHz Intel Core processor, and 64-bit OS, an Mplus program

to fit the saturated LCDM to the ECPE data (retrieved from https://jonathantemplin.com/dcm-workshop-spring-2012-ncme/) took 47 min to converge, whereas the model fit with LG (using the program provided in the Appendix) took less than 3 s (packages in R also have shorter run times).

Several researchers have suggested starting with a saturated model and attempting to determine which sub-model might be more appropriate (e.g., Rupp et al., 2010). Recent studies have examined this approach using information criteria (Chen, de la Torre, & Zhang, 2013) and the Wald test (de la Torre & Lee, 2013) for the DINA, DINO, and ACDM models. Given the ability to fit the saturated model and the various sub-models in LG, it is straightforward to implement these types of model comparisons.

11.11 Discussion

Reparameterized diagnostic models are useful both for illustrating and understanding basic aspects of CDMs, as well as providing a bridge to latent class models and accompanying software. The importance of recognizing the signal detection nature of the parameters is emphasized. Monotonicity, for example, is seen to be a simple restriction on the discrimination parameter (i.e., that it is greater than zero) which ensures that the corresponding ROC curves lie above the diagonal line (which represents zero discrimination). The reparameterized models also help make concepts such as duality more transparent and are useful for showing how different models are related. For example, the models show that duality leads to a simple way to fit the rDINO model with a program for the rDINA model by using a negative monotonicity constraint. It also clarifies that DINA and DINO are structurally different models with different parameters. All the options for estimation, classification, and other output and tools available in latent class software become immediately available for CDMs. One can also go beyond the GRM, in that one can consider models with nominal or ordinal indicators with more than two categories, models with continuous indicators, models with nominal, ordinal, or continuous latent variables, and models with other link functions besides the logit, all in a very straightforward manner, and all available in current software.

Appendix

Complete Latent Gold program to fit the rDINA model, 15 items, 4 unstructured attributes.

```
model
options
   maxthreads=all;
```

```
   algorithm
      tolerance=1e-008 emtolerance=0.01 emiterations=250
      nriterations=50 ;
   startvalues
      seed=0 sets=16 tolerance=1e-005 iterations=50;
   bayes
      categorical=0 variances=0 latent=0 poisson=0;
   montecarlo
      seed=0 sets=0 replicates=500 tolerance=1e-008;
   quadrature  nodes=10;
   missing  excludeall;
   output
      parameters=effect  betaopts=wl standarderrors
      profile probmeans=posterior
      bivariateresiduals estimatedvalues=model;
variables
   dependent y1 cumlogit, y2 cumlogit, y3 cumlogit,
   y4 cumlogit, y5 cumlogit,
   y6 cumlogit, y7 cumlogit, y8 cumlogit, y9 cumlogit,
   y10 cumlogit, y11 cumlogit,
   y12 cumlogit, y13 cumlogit, y14 cumlogit, y15 cumlogit;
   latent
     a1 ordinal 2 score=(0 1), a2 ordinal 2 score=(0 1),
     a3 ordinal 2 score=(0 1), a4 ordinal 2 score=(0 1);
equations
//next line uses a saturated association model for the
  attribute model//
  (r~full) a1 <-> a2 <-> a3 <-> a4;
//for sequential path approach, replace above with://
//a1 <- 1; a2 <- 1 + a1; a3 <- 1 + a1 + a2 + a1 a2//
//a4 <- 1 + a1 + a2 + a3 + a1 a2 + a1 a3 +
  a2 a3 + a1 a2 a3//
   y1 <- 1 + (+)a1;
   y2 <- 1 + (+)a2;
   y3 <- 1 + (+)a3;
   y4 <- 1 + (+)a4;
   y5 <- 1 + (+)a1 a2;
   y6 <- 1 + (+)a1 a3;
   y7 <- 1 + (+)a1 a4;
   y8 <- 1 + (+)a2 a3;
   y9 <- 1 + (+)a2 a4;
   y10 <- 1 + (+)a3 a4;
   y11 <- 1 + (+)a1 a2 a3;
   y12 <- 1 + (+)a1 a2 a4;
   y13 <- 1 + (+)a1 a3 a4;
   y14 <- 1 + (+)a2 a3 a4;
   y15 <- 1 + (+)a1 a2 a3 a4;
//remove next line for the sequential path approach//
r[1,1]=0;
end model
```

Latent Gold program to fit the rDINO model of Eq. (11.2) (starting from variables statement)

```
variables
   dependent y1 cumlogit, y2 cumlogit, y3 cumlogit,
   y4 cumlogit, y5 cumlogit,
  y6 cumlogit, y7 cumlogit, y8 cumlogit, y9 cumlogit,
   y10 cumlogit, y11 cumlogit,
  y12 cumlogit, y13 cumlogit, y14 cumlogit, y15 cumlogit;
   latent
     a1 ordinal 2 score=(0 1), a2 ordinal 2 score=(0 1),
     a3 ordinal 2 score=(0 1), a4 ordinal 2 score=(0 1);
equations
 (r~full) a1 <-> a2 <-> a3 <-> a4;
  y1 <- 1 + (+)a1;
  y2 <- 1 + (+)a2;
  y3 <- 1 + (+)a3;
  y4 <- 1 + (+)a4;
  y5 <- 1 + (+a)a1 + (+a)a2 + (-a)a1 a2;
  y6 <- 1 + (+b)a1 + (+b)a3 + (-b)a1 a3;
  y7 <- 1 + (+c)a1 + (+c)a4 + (-c)a1 a4;
  y8 <- 1 + (+d)a2 + (+d)a3 + (-d)a2 a3;
  y9 <- 1 + (+e)a2 + (+e)a4 + (-e)a2 a4;
  y10 <- 1 + (+f)a3 + (+f)a4 + (-f)a3 a4;
  y11 <- 1 + (+g)a1 + (+g)a2 + (+g)a3 + (-g)a1 a2
   + (-g)a1 a3 + (-g)a2 a3 + (+g)a1 a2 a3;
  y12 <- 1 + (+h)a1 + (+h)a2 + (+h)a4 + (-h)a1 a2
   + (-h)a1 a4  + (-h)a2 a4 + (+h)a1 a2 a4;
  y13 <- 1 + (+i)a1 + (+i)a3 + (+i)a4 + (-i)a1 a3
   + (-i)a1 a4 + (-i)a3 a4 + (+i)a1 a3 a4;
  y14 <- 1 + (+j)a2 + (+j)a3 + (+j)a4 + (-j)a2 a3
   + (-j)a2 a4 + (-j)a3 a4 + (+j)a2 a3 a4;
  y15 <- 1 + (+k)a1 + (+k)a2 + (+k)a3 + (+k)a4
   + (-k)a1 a2 + (-k)a1 a3 + (-k)a1 a4 + (-k)a2 a3
   + (-k)a2 a4 + (-k)a3 a4  + (+k)a1 a2 a3 +(+k)a1 a2 a4
   + (+k)a1 a3 a4 + (+k)a2 a3 a4 + (-k)a1 a2 a3 a4;
r[1,1]=0;
end model
```

Latent Gold program to fit the restricted (and unrestricted) GRM to the ECPE data

```
variables
   dependent i1 cumlogit, i2 cumlogit, i3 cumlogit, i4 cumlogit,
        i5 cumlogit, i6 cumlogit,
   i7 cumlogit, i8 cumlogit, i9 cumlogit, i10 cumlogit,
        i11 cumlogit, i12 cumlogit, i13 cumlogit,
   i14 cumlogit, i15 cumlogit, i16 cumlogit, i17 cumlogit,
        i18 cumlogit, i19 cumlogit,
   i20 cumlogit, i21 cumlogit, i22 cumlogit, i23 cumlogit,
        i24 cumlogit, i25 cumlogit,
   i26 cumlogit, i27 cumlogit, i28 cumlogit;
   latent
      a1 ordinal 2 score=(0 1), a2 ordinal 2 score=(0 1),
      a3 ordinal 2 score=(0 1);
equations
```

```
    (r~full) a1 <-> a2 <-> a3;
    i1 <- 1 + (+)a1 + (+)a2 + (+)a1 a2;
    i2 <- 1 + (+)a2;
    i3 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;
    i4 <- 1 + (+)a3;
    i5 <- 1 + (+)a3;
    i6 <- 1 + (+)a3;
  //i7 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;//
  //order restriction on Item 7 can be done as follows//
    i7 <- 1 + (+)a1 + (+a)a3 + (-a)a1 a3 + (+)a1 a3;
    i8 <- 1 + (+)a2;
    i9 <- 1 + (+)a3;
    i10 <- 1 + (+)a1;
    i11 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;
    i12 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;
    i13 <- 1 + (+)a1;
    i14 <- 1 + (+)a1;
    i15 <- 1 + (+)a3;
    i16 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;
    i17 <- 1 + (+)a2 + (+)a3 + (+)a2 a3;
    i18 <- 1 + (+)a3;
    i19 <- 1 + (+)a3;
    i20 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;
    i21 <- 1 + (+)a1 + (+)a3 + (+)a1 a3;
    i22 <- 1 + (+)a3;
    i23 <- 1 + (+)a2;
    i24 <- 1 + (+)a2;
    i25 <- 1 + (+)a1;
    i26 <- 1 + (+)a3;
    i27 <- 1 + (+)a1;
    i28 <- 1 + (+)a3;
   r[1,1]=0;
  end model
```

References

Agresti, A. A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*, 850–866.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

Chiu, C.-Y., & Köhn, H.-F. (2016). The reduced rum as a logit model: Parameterization and constraints. *Psychometrika, 81*(2), 350–370.

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 211–359). New York, NY: Plenum.

Clogg, C. C., & Eliason, S. R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research, 16*, 8–44.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics, 40*(5), 454–476.

Dayton, C. M. (1998). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373.

DeCarlo, L. T. (2002). A latent class extension of signal detection theory, with applications. *Multivariate Behavioral Research, 37*, 423–451.

DeCarlo, L. T. (2005). A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement, 42*, 53–76.

DeCarlo, L. T. (2008). *Studies of a latent-class signal-detection model for constructed response scoring* (ETS Research Report No. RR-08-63). Princeton, NJ: Educational Testing Service.

DeCarlo, L. T. (2010). *Studies of a latent class signal detection model for constructed response scoring II: Incomplete and hierarchical designs* (ETS Research Report No. RR-10-08). Princeton, NJ: Educational Testing Service.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447–468.

DeCarlo, L. T., Kim, Y. K., & Johnson, M. S. (2011). A hierarchical rater model for constructed responses with a signal detection rater model. *Journal of Educational Measurement, 48*, 333–356.

DeCarlo, L. T., & Kinghorn, B. R. C. (2016, April). *An exploratory approach to the Q-matrix via Bayesian estimation*. Paper presented at the 2016 meeting of the National Council on Measurement in Education, Washington, DC.

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*, 1–24.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities*. Unpublished doctoral dissertation.

Heller, J., & Wickelmaier, F. (2013). Minimum discrepancy estimation in probabilistic knowledge structures. *Electronic Notes in Discrete Mathematics, 42*, 49–56.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*, 407–419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kim, Y. K. (2009). *Combining constructed response items and multiple choice items using a hierarchical rater model*. Doctoral dissertation, Teachers College, Columbia University.

Köhn, H.-F., & Chiu, C.-Y. (2016). A proof of the duality of the DINA model and the DINO model. *Journal of Classification, 33*, 171–184.

Köhn, H.-F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests. *Psychometrika, 82*, 112–132.

Liu, J., Xu, G., & Ying, Z. (2011). *Learning item-attribute relationship in Q-matrix based diagnostic classification models* (Report: arXiv:1106.0721v1). New York, NY: Columbia University, Department of Statistics. Retrieved from the website https://arxiv.org/abs/1106.0721v1

Ma, W., & de la Torre, J. (2017). *GDINA: The generalized DINA model framework* (R package version 1.4.2). Retrieved from https://cran.r-project.org/web/packages/GDINA/index.html

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York, NY: Cambridge University Press.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics, 43*, 88–115.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2014). *OpenBUGS version 3.2.3 user manual*. Helsinki, Finland. Retrieved from http://www.openbugs.net/w/Manuals

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317–339.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification and model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*, 37–50.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Vermunt, J. K. (1997). *LEM: A general program for the analysis of categorical data*. Tilburg, The Netherlands: Tilburg University.

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for latent GOLD 5.1: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations Inc.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model – Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (ETS Research Report No. RR-14-40). Princeton, NJ: Educational Testing Service.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York, NY: Oxford University Press.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*, 625–649.

Zhang, S. S., DeCarlo, L. T., & Ying, Z. (2013). *Non-identifiability, equivalence classes, and attribute-specific classification in Q-matrix based cognitive diagnosis models*. Available at: http://arxiv.org/abs/1303.0426v1

Part II
Special Topics

Chapter 12
Q-Matrix Learning via Latent Variable Selection and Identifiability

Jingchen Liu and Hyeon-Ah Kang

Abstract Much of the research and application in cognitive diagnostic assessments to date has been centered on a confirmatory approach where a Q-matrix is pre-identified using content experts' opinion or test developers' knowledge on test items. As opposed to the traditional methods, which require prior knowledge about latent dimensions and underlying structure of test items, the approaches described in this chapter attempt to identify a Q-matrix solely relying on the observed test response data and thus avoid probable decision error. There are several important aspects to consider when estimating a Q-matrix from the observed data. First, a fundamental question of identifiability arises, that is, whether and to what extent Q can be estimated from data. The second aspect to consider in learning Q concerns the computational intensity that arises from estimation. The third aspect pertains to the presence of missing data, more precisely, the latent attributes underlying the observed data. The completeness of a Q-matrix, the other important aspect to consider in identifying Q, is beyond the scope of the present chapter.

12.1 Introduction

Much of the research and application in cognitive diagnostic assessments to date has been centered on a confirmatory approach where a Q-matrix is pre-identified using content experts' opinion or test developers' knowledge about test items. The

Jingchen Liu is supported in part by NSF SES-1323977, SES-1826540, IIS-1633360, and Army Research Office grant W911NF-15-1-0159.

J. Liu (✉)
Department of Statistics, Columbia University, New York, NY, USA
e-mail: jcliu@stat.columbia.edu

H.-A. Kang
Department of Educational Psychology, University of Texas, Austin, TX, USA
e-mail: hkang@austin.utexas.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_12

Q-matrix constructed this way however is often subject to mis-specification, which can consequently have an adverse impact on the estimation of model parameters and assessment accuracy (e.g., de la Torre, 2008; Rupp & Templin, 2008). The focus of the present chapter is on statistical methods that can estimate a Q-matrix objectively from observed data. As opposed to the traditional methods, which require prior knowledge about latent dimensions and underlying structure of test items, the approaches described in this chapter attempt to identify a Q-matrix solely relying on the observed test response data and thus avoid probable decision error.

There are several important aspects to consider when estimating a Q-matrix from the observed data. First, a fundamental question of identifiability arises, that is, whether and to what extent Q can be estimated from data. The identifiability of Q is a nontrivial question especially when a multitude of diagnostic classification models (DCMs) are conceivable for a given data set. In many cases, a Q-matrix cannot be uniquely identified, and it is not uncommon for different Q-matrices to lead to an identical response distribution. The present chapter outlines results suggested in the current literature concerning the extent to which a Q-matrix can be identified from data in the absence of prior information. In particular, the chapter recounts some of the theoretical results derived in Liu, Xu, and Ying (2013) and Chen, Liu, Xu, and Ying (2015) within the framework of the DINA model (Haertel, 1989; Junker & Sijtsma, 2001). The more recent study of Fang, Liu, and Ying (2017b) is also introduced, which discusses identifiability within the general latent class models that subsume most of the DCMs currently in use.

The second aspect to consider in learning Q concerns the computational intensity that arises from estimation. When a Q-matrix is considered as one of the model parameters and estimated via standard inference method such as maximum likelihood estimation, the optimization of the estimation procedure is typically saddled with significant computational overhead. Since Q is a discrete matrix, standard calculus tools are not applicable; instead, one needs to search through all matrices in the possible domain. Suppose a test is designed to measure D distinct attributes with K items. Then, a most likely Q-matrix is found by rummaging through the discrete domain $\{0,\ 1\}^{K\times D}$. Notice that this space grows exponentially with K and D, and the computational intensity is substantial even for a test with small K or D. The present chapter discusses a number of approaches that can cope with this computational issue in estimating Q. The methods are categorized in two schemes: (1) the estimation based on the empirical distribution (e.g., Liu, Xu, & Ying, 2012; Liu et al., 2013), and (2) the estimation via latent variable selection (e.g., Chen et al., 2015; Fang, Liu, & Ying, 2017a).

The third aspect pertains to the presence of missing data, more precisely, the latent attributes underlying the observed data. In DCMs, the relationship between the responses and latent attributes is specified by a nonlinear discrete function, and the particular distributional assumptions are often imposed on the latent attributes. The standard approach to estimating the model parameters in the presence of missing data is to use a marginalized likelihood function where the unobservable latent attributes are integrated out. Such procedure is commonly implemented via well-known expectation-maximization (EM) algorithm. In like manner, the

procedures of estimating Q (e.g., Chen et al., 2015, Fang et al., 2017a) employ the EM algorithm and iteratively update the marginal probability distribution of the latent attributes during parameter estimation. The present chapter discusses the application of the EM algorithm when the Q-matrix is estimated through the latent variable selection problem. We also refer the readers to Friedman, Hastie, and Tibshirani (2010), and Tseng (1988, 2001) for related optimization methods.

Associated with the current issue, the present chapter also briefly introduces a method that estimates a factor loading matrix within the context of multidimensional item response theory (MIRT). The MIRT posits continuous latent trait variables to account for examinees' test performance, and the links between the test items and trait variables are specified by a so-called factor loading matrix. While the nature of the underlying latent space differs between the MIRT models and DCMs, similar approaches can be applied to estimate the factor loading matrix in MIRT. The present chapter introduces one of such procedures that has been proposed in the same spirit with the Q-matrix estimation.

As a final note, it is to be noted that several other strategies that have been proposed in relation to construction of Q are not the focus of this chapter. For example, some studies considered creating an initial matrix by applying the available knowledge and executing statistical algorithms to refine the provisional Q-matrix (e.g., Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016) or to estimate the elements of Q (e.g., DeCarlo, 2012; Templin & Henson, 2006). A Bayesian approach for objectively estimating a Q-matrix has also been proposed by Chen, Culpepper, Chen, and Douglas (2017). In addition, the completeness of a Q-matrix, the other important aspect to consider in identifying Q, is beyond the scope of the present chapter. Interested readers may refer to Chiu, Douglas, and Li (2009), de la Torre and Chiu (2016), and Köhn and Chiu (2017). See also Liu (2017) for a further theoretical discussion.

12.2 Identifiability of Q-Matrix

Cognitively diagnostic assessments hinge on a Q-matrix that specifies loading structure of test items on a set of cognitive attributes being measured. Based on the specification of item-attribute association, cognitive diagnosis attempts to classify individuals into a number of homogeneous groups. The groups are commonly characterized by a two-level latent vector though multi-level latent status is possible, and each dimension of the latent vector indicates mastery (or level of mastery) of the corresponding attribute.

Suppose a test measures D distinct latent attributes with K binary items. For simplicity of discussion, the current chapter focuses on the two-level attributes and binary-response items; most of the conclusions made in the following however can be extended to multi-level attributes and multi-category responses. Given the set of D attributes measured by the test, a Q-matrix consists of $K \times D$ elements, each indicating the relationship between the item and attribute being assessed. An

individual entry of the Q-matrix, q_{kd}, equals 1 if attribute d is required to respond to item k correctly, and 0 otherwise. For a given item k, the kth row vector of Q, $\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kD})$, specifies the set of attributes that are measured by, or relevant to, item k.

A standard definition of identifiability of a parameter θ requires that distinct values of θ correspond to the distinct probability distributions, or equivalently, there exist no $\theta' \neq \theta$ that satisfies $f(X \mid \theta) = f(X \mid \theta')$. In many cases of DCM analyses, however, an exact Q-matrix cannot be uniquely identified because permuting columns of a matrix (or equivalently, relabeling the attributes) can lead to an identical response distribution. That is to say, a Q-matrix can be identified only up to a column permutation when estimated from data. In fact, "up to a column permutation" is the finest identifiability result one can draw regarding the specific meaning of each attribute in the absence of prior knowledge. To indicate that two matrices Q and Q' are identical up to a column permutation, we write $Q \sim Q'$; otherwise, denote as $Q \nsim Q'$. The notion $Q \sim Q'$ implies that Q and Q' have identical column vectors if rearranged in the different orders.

The first study on identifiability of Q was presented in Liu et al. (2013). The study provides sufficient conditions under which a Q-matrix can be estimated consistently up to a column permutation within the DINA model framework. The current chapter introduces some of the essential results from Liu et al. as well as from the ensuing studies. To begin the discussion, some necessary notation is introduced as follows. Continuing the assumption that a test measures D distinct attributes, an underlying attribute profile is denoted by $\boldsymbol{A} = (A_1, \ldots, A_D)$ with a realization $\boldsymbol{a} = (a_1, \ldots, a_D)$, where $a_d \in \{0, 1\}$ indicates the presence or absence of the dth attribute. Among the 2^D possible attribute patterns, the probability distribution for a particular pattern, $\boldsymbol{a}$, is denoted by $p_{\boldsymbol{a}} = P(\boldsymbol{A} = \boldsymbol{a})$. For the population of interest, $\boldsymbol{p} = (p_{\boldsymbol{a}} : \boldsymbol{a} \in \{0, 1\}^D)$ subject to the constraint that $p_{\boldsymbol{a}} \in [0, 1]$ and $\sum_{\boldsymbol{a}} p_{\boldsymbol{a}} = 1$. Suppose that responses of N individuals have been observed and are denoted by $(X_{nk} : n = 1, \ldots, N, \ k = 1, \ldots, K)$. For notational simplicity, the responses of a generic subject will be denoted by $X_1, \ldots, X_K$ without the subscript n.

In the DINA model, the observable response is dictated by an ideal response, denoted by η_k:

$$\eta_k = \prod_{d=1}^{D} (a_d)^{q_{kd}} = \mathbb{I}(a_d \geq q_{kd} : d = 1, \ldots, D), \tag{12.1}$$

where $\mathbb{I}(\cdot)$ is an indicator function, that is, $\mathbb{I}(a_d \geq q_{kd} : d = 1, \ldots, D) = 1$ if all a_d are greater than or equal to q_{kd} across all different values of d. The η_k denotes the individual's ideal response to an item k when the individual has attribute profile $\boldsymbol{a} = (a_1, \ldots, a_D)$. The cognitive process of responding to item is considered conjunctive in the DINA model in the sense that it requires an examinee to master all the attributes assessed by the item for providing a correct response. Given η_k, the DINA model defines the probability of a positive response to item as

$$P(X_k = 1 \mid \boldsymbol{a},\ Q,\ s_k,\ g_k) = (1 - s_k)^{\eta_k} g_k^{1-\eta_k},$$

where s_k and g_k are the slipping and guessing parameters for item k, respectively. The complement of s_k, $1 - s_k$, is often reparameterized as c_k, indicating the probability that masters of item k answer the item correctly.

Assuming local independence of item responses given the latent class $\boldsymbol{a}$, the probability distribution of a set of responses $\boldsymbol{X} = (X_1, \ldots, X_K)$ is obtained as

$$f(\boldsymbol{X} = \boldsymbol{x} \mid Q,\ \boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g}) = \sum_{\boldsymbol{a}} p_{\boldsymbol{a}} \prod_{k=1}^{K} P(X_k = x_k \mid Q,\ \boldsymbol{a},\ \boldsymbol{c},\ \boldsymbol{g}), \tag{12.2}$$

where $\boldsymbol{x} = (x_1, \ldots, x_K)$ denotes a realization of $\boldsymbol{X}$, $\boldsymbol{c} = (c_1, \ldots, c_K)$, and $\boldsymbol{g} = (g_1, \ldots, g_K)$. Equation (12.2) serves as an expected probability distribution of response vector $\boldsymbol{X}$. The empirical distribution is obtained from the observed data as

$$\hat{P}(\boldsymbol{X} = \boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}(\boldsymbol{X}_n = \boldsymbol{x}). \tag{12.3}$$

If the Q-matrix and the other parameters $(\boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g})$ were correctly specified, the empirical distribution $\hat{P}(\boldsymbol{x})$ converges to its theoretical value (12.2) as the sample size grows. An estimator of Q can be then constructed such that it minimizes the discrepancy between the theoretical and empirical probability distributions of the response patterns.

The sufficient conditions under which the Q-matrix can be consistently estimated (subject to a column permutation) are (Liu et al., 2013)

C1. $p_{\boldsymbol{a}} > 0$ for all $\boldsymbol{a} \in \{0,\ 1\}^D$;
C2. $1 - s_k > g_k$ for all k; and
C3. The true matrix Q_0 is complete, that is, $\{\boldsymbol{e}_d : d = 1, \ldots, D\} \subset \{\boldsymbol{q}_k : k = 1, \ldots, K\}$, where $\boldsymbol{e}_d$ is the standard basis vector in the D-dimensional Euclidean space.

C1 states that the examinee population should be fully diversified such that all attribute patterns exist in the population. C2 requires all test items have positive discriminating power. C3 presents a sufficient and necessary condition for a set of items to consistently identify examinees' attribute profiles.

Under these conditions, one can construct a consistent estimator $\hat{Q}$ up to a column permutation. Specifically, the estimator for the DINA model satisfies

$$\lim_{N \to \infty} P(\hat{Q} \sim Q_0) = 1$$

when the guessing parameters are known. The same conclusion can be derived when the guessing parameters are unknown, yet with additional conditions (Chen et al., 2015):

C4. $\sum_{k=1}^{K} q_{kd} \geq 3$ for all d; and
C5. After some row permutation corresponding to the reordering of items, Q_0 is reformulated as

$$Q_0 = \begin{pmatrix} \mathcal{I}_D \\ \mathcal{I}_D \\ Q' \end{pmatrix}, \tag{12.4}$$

where $\mathcal{I}_D$ is the $D \times D$ identity matrix, and Q' denotes the rest of the Q_0-matrix.

C4 requires that each latent attribute be measured by at least three items. C5 states that, for each latent attribute, there exist at least two items measuring the attribute exclusively. Under the conditions C1, C2, C4, and C5, the Q-matrix is identifiable with

$$\lim_{N \to \infty} P(\hat{Q} \sim Q_0) = 1. \tag{12.5}$$

In addition, if the conditions C1, C2, C4, and C5 hold, $\boldsymbol{p}$, $\boldsymbol{c}$, and $\boldsymbol{g}$ are all identifiable.

While the above identifiability results are derived primarily for the DINA model, the established theorems can be readily extended to the DINO model as a result of duality of the two models (Chen et al., 2015; Köhn & Chiu, 2016). To put it concretely, let s_k', g_k', and $\boldsymbol{a}'$ denote the parameters pertaining to the DINO model. Equating

$$1 - s_k = s_k', \quad 1 - g_k = g_k', \quad \text{and} \quad \boldsymbol{a} = 1 - \boldsymbol{a}'$$

results in the identical probability distribution for the observed response pattern $\boldsymbol{x}$ under the DINA and DINO models. Hence, the identifiability results established for the DINA model are also applicable to the DINO model.

As a matter of fact, identifiability results can be further extended to the generic restricted latent class models. Fang et al. (2017b) consider the problem of identifiability within the framework of the general latent class model, which subsumes DCMs with special constraints on the latent attribute space and loading structure. Note that the item response function depends on specific parametrization of a latent space, and hence, it is generally difficult to derive a unified theory applicable to all DCMs. As such, Fang et al. approach the identifiability problem by considering partial information that each item can provide for differentiating the latent classes. The motivation is that a single item in usual cognitive assessments does not provide sufficient information for differentiating all dimensions of attribute profiles; rather, an individual item induces an equivalence relation over the latent classes.

Formally, two distinct latent classes $\boldsymbol{a}$ and $\boldsymbol{a}'$ are said equivalent in view of item k if

$$P(X_k = x_k \mid \boldsymbol{a}) = P(X_k = x_k \mid \boldsymbol{a}')$$

for all x_k. Let $\boldsymbol{a} \overset{k}{\sim} \boldsymbol{a}'$ denote the item-specific equivalence relation and $[\boldsymbol{a}]_k$ represent the equivalence class corresponding to "$\overset{k}{\sim}$". If $[\boldsymbol{a}]_k = [\boldsymbol{a}']_k$, it is said that item k provides no information for distinguishing between $\boldsymbol{a}$ and $\boldsymbol{a}'$, and consequently, item k identifies $\boldsymbol{a}$ only up to the equivalence relation of $[\boldsymbol{a}]_k$. In other words, the equivalence relation $\overset{k}{\sim}$ characterizes the partial information that item k provides on the latent space, and a Q-matrix can be seen as a parameterization of the items' partial information structure.

Among the several identifiability results provided in Fang et al. (2017b), the present chapter introduces the theorem pertinent to the current context below. Under the following conditions,

A1. For each attribute d, there exist at least three non-overlapping sets of items, each of which measures the attribute d only;
A2. Each set of items provides information to identify all levels of a_d; and
A3. $p_{\boldsymbol{a}} > 0$ for all $\boldsymbol{a} \in \{0,\ 1\}^D$.

The partial information of all items are identifiable up to a permutation of the latent classes. That is, one can estimate equivalence class for item k, denoted by $\langle \cdot \rangle_k$, such that there exists a permutation on the latent class space σ that follows

$$\lim_{N \to \infty} P\left(\langle \sigma(\boldsymbol{a}) \rangle_k = [\boldsymbol{a}]_k\right) = 1. \tag{12.6}$$

The results can also be generalized to more complex models that are comprised of multi-category responses and/or multi-class attributes. For other results regarding the partial information identification, see Fang et al. (2017b).

12.3 Q-Matrix Learning

The present section introduces several approaches to estimating a Q-matrix. Most of the methods are generic and applicable to general DCMs.

12.3.1 Maximum Likelihood Estimation

The standard approach to estimating Q is to consider a maximum likelihood (ML) estimator:

$$\hat{Q}_{\mathrm{ML}} = \underset{Q}{\operatorname{argsup}} \sup_{(\boldsymbol{p},\, \boldsymbol{c},\, \boldsymbol{g})} L(\boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g},\ Q), \tag{12.7}$$

where

$$L(\boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g},\ Q) = \prod_{n=1}^{N} \sum_{\boldsymbol{a}_n} p_{\boldsymbol{a}_n} \left[\prod_{k=1}^{K} P_k(\boldsymbol{a}_n)^{X_{nk}} (1 - P_k(\boldsymbol{a}_n))^{1-X_{nk}} \right],$$

and $P_k(\boldsymbol{a}_n) = P(X_k = 1 \,|\, \boldsymbol{a}_n,\ c_k,\ g_k,\ Q)$. Under the identifiability conditions discussed through C1 to C5, the estimator $\hat{Q}_{\mathrm{ML}}$ is consistent in the sense of (12.5). The ML estimators of the other parameters are likewise consistent; see Liu et al. (2012) for more details.

12.3.2 Using an Empirical Distribution

The second approach utilizes the empirical distribution of observed data and compares against the marginal probability distribution under the presumed model. Suppose the marginal distribution of $\boldsymbol{X}$, $f(\boldsymbol{X} = \boldsymbol{x} \,|\, Q,\ \boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g})$, is given by (12.2), and the empirical distribution of $\boldsymbol{X}$ is given by (12.3). An estimator can be then obtained by minimizing the L_2 distance between the marginal and empirical distributions of $\boldsymbol{X}$:

$$\min_{Q,\ \boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g}} \sum_{\boldsymbol{x} \in \{0,1\}^D} \left| \hat{P}(\boldsymbol{X} = \boldsymbol{x}) - f(\boldsymbol{X} = \boldsymbol{x} \,|\, Q,\ \boldsymbol{p},\ \boldsymbol{c},\ \boldsymbol{g}) \right|^2. \tag{12.8}$$

The estimator obtained by (12.8) is consistent and can identify the true Q-matrix up to an equivalence class under some sufficient conditions (Liu et al., 2013).

12.3.3 Latent Variable Selection

While the above estimators have sound theoretical properties, they tend to induce substantial computational overhead. In order to maximize the likelihood function or to minimize the L_2 distance, one needs to search through the entire space of $K \times D$ binary matrices, which is often practically infeasible even for moderate K and D. Although there exist some iterative algorithms that can be used for optimizing the objective function over the large discrete space (e.g., Liu et al., 2012), computation remains a critical challenge for estimating a Q-matrix. Recently, more viable alternatives have been proposed that cast the Q-matrix estimation as a latent variable selection problem. These methods piggyback on well-developed optimization techniques and demonstrate more computational efficiency than the classic inference methods.

To illustrate the connection between the variable selection and Q-matrix estimation, consider a simple example. Suppose that a test measures three attributes, and

a $\boldsymbol{q}$-vector for an item has the form of $\boldsymbol{q} = (1,\ 1,\ 0)$. When it comes to responding to the item, having or not having mastery of the third attribute has no impact on the response distribution, and hence, the following identity holds

$$P(X = x \mid \boldsymbol{a}) = P(X = x \mid a_1,\ a_2).$$

This identity holds for all DCMs although the specific form of the item response distribution may vary by the model. The above example suggests that the problem of estimating $\boldsymbol{q}$ can be seen as identifying an array of attributes associated with the item response. If the attributes were observed, the analysis of $P(X = x \mid \boldsymbol{a})$ simplifies to the variable selection problem in a usual regression model as it attempts to single out the relevant elements in a predictor variable $\boldsymbol{a}$. When estimating the parameters of the DCM, however, the attribute profile is not directly observable and assumed latent. Hence, identifying the pertinent attributes (i.e., estimating $\boldsymbol{q}$) is equivalent to selecting the latent variables related to the item response. The methods described in the following are based on this idea and attempt to estimate Q through the well-developed latent variable selection technique.

12.3.3.1 DCM as a Generalized Linear Model

To cast the Q-matrix estimation problem as a latent variable selection problem, the DCM is reformulated as a generalized linear model (e.g., de la Torre, 2011; Henson, Templin, & Willse, 2009; von Davier, 2005, 2008). The response distribution then belongs to a natural exponential family. For each attribute profile $\boldsymbol{a} = (a_1,\ \ldots,\ a_D)$, let

$$\boldsymbol{\alpha} = (1,\ a_1,\ \ldots,\ a_D,\ a_1a_2,\ \ldots,\ a_1a_2 \cdots a_D)^\top$$

be the 2^D-dimensional binary latent vector containing all interactions among the components in $\boldsymbol{a}$. The corresponding vector of regression coefficients for item k is denoted as

$$\boldsymbol{\beta}_k = (\beta_{k0},\ \beta_{k1},\ \ldots,\ \beta_{kD},\ \beta_{k12},\ \ldots,\ \beta_{k12\cdots D})^\top. \tag{12.9}$$

Using the canonical link function, the response distribution is expressed as

$$P(X_k = x \mid \boldsymbol{\alpha}) = f(x) \exp\{T(x)\boldsymbol{\beta}_k^\top \boldsymbol{\alpha} - \varphi(\boldsymbol{\beta}_k^\top \boldsymbol{\alpha})\}, \tag{12.10}$$

where $T(x)$ is sufficient statistic of x. The $T(x)$ can be of multiple dimensions whereby $\boldsymbol{\beta}_k$ becomes a matrix. In the case that X_k is a binary variable (i.e., $x \in \{0, 1\}$), (12.10) reduced to a logistic model

$$P(X_k = 1 \mid \boldsymbol{\alpha}) = \frac{\exp\{\boldsymbol{\beta}_k^\top \boldsymbol{\alpha} - \varphi(\boldsymbol{\beta}_k^\top \boldsymbol{\alpha})\}}{1 + \exp\{\boldsymbol{\beta}_k^\top \boldsymbol{\alpha} - \varphi(\boldsymbol{\beta}_k^\top \boldsymbol{\alpha})\}}. \tag{12.11}$$

If $D = 2$ and $\boldsymbol{\alpha} = (a_1,\ a_2)$, for instance, the above model is expressed as

$$P(X_k = 1 \mid \boldsymbol{\alpha}) = \frac{\exp\{\beta_{k0} + \beta_{k1}a_1 + \beta_{k2}a_2 + \beta_{k12}a_1a_2\}}{1 + \exp\{\beta_{k0} + \beta_{k1}a_1 + \beta_{k2}a_2 + \beta_{k12}a_1a_2\}}. \tag{12.12}$$

Note that the above models are expressed in a saturated form without any constraints on the response distribution. The specific DCMs may however place additional constraints or parametric forms. For example, a two-attribute DINA model of the form (12.12) requires two elements in each $\boldsymbol{\beta}_k$ be non-zero—one for β_{k0} and the other for the highest interaction among the attributes required by item k. To put it concretely, if $\boldsymbol{q} = (1,\ 0)$, both β_{k0} and β_{k1} are nonzero; if $\boldsymbol{q} = (0,\ 1)$, β_{k0} and β_{k2} are nonzero; and if $\boldsymbol{q} = (1,\ 1)$, β_{k0} and β_{k12} have nonzero values. Hence, it can be seen that the Q-matrix estimation for the DINA model corresponds to identifying the nonzero coefficients in $\boldsymbol{\beta}_k$, which may again be subject to parametric constraint of the specific DCM of concern. Similar correspondence between the Q-matrix and nonzero pattern in $\boldsymbol{\beta}$ can be identified for other DCMs; see Chen et al. (2015) for a discussion.

12.3.3.2 Regularized Likelihood Estimation of Q for Parametric DCMs

A standard way to estimating a Q-matrix within the generalized linear modeling framework similarly involves the evaluation of the likelihood function. Recall that the ML estimator will result in substantial computational overhead in consequence of searching through $2^{K\times D}$ matrices. To relieve the computational intensity, Chen et al. (2015) propose a regularized ML estimator, the method of which has been widely-studied and well-established for variable selection problems in regression models. Specifically, the likelihood function within the generalized linear modeling framework is expressed as follows.

$$L(B,\ \boldsymbol{p};\ \boldsymbol{X}_n,\ n = 1,\ \ldots,\ N) = \prod_{n=1}^{N} \sum_{\boldsymbol{\alpha}_n} p_{\boldsymbol{\alpha}_n} \left[\prod_{k=1}^{K} P_k(\boldsymbol{\alpha}_n)^{X_{nk}} (1 - P_k(\boldsymbol{\alpha}_n))^{1-X_{nk}} \right],$$

where $B = (\boldsymbol{\beta}_k : k = 1,\ \ldots,\ K)$, and $P_k(\boldsymbol{\alpha}_n) = P(X_k = 1 \mid \boldsymbol{\alpha}_n)$. For identifying the nonzero coefficients, Chen et al. (2015) consider maximizing the regularized likelihood:

$$(\hat{B},\ \hat{\boldsymbol{p}}) = \underset{(B,\ \boldsymbol{p})}{\arg\max}\ l(B,\ \boldsymbol{p};\ \boldsymbol{X}_n,\ n = 1,\ \ldots,\ N) - N \sum_{k=1}^{K} p_{\lambda_k}(\boldsymbol{\beta}_k), \tag{12.13}$$

where $l(\cdot) = \log L(\cdot)$, and $p_{\lambda_k}(\cdot)$ is a penalty function with λ_k being a regularization parameter. The above regularized ML estimator naturally incorporates the principle

of parsimony in the sense that the optimization tends to penalize complexity in $\boldsymbol{\beta}_k$. The resulting estimate of $\boldsymbol{\beta}_k$ contains as few nonzero elements as possible while accomplishing the desired level of explanation.

To simplify computation even more in (12.13), one can constrain the value of λ_k such that they are equal across the items, thus resulting in $\lambda_k = \lambda$. Two commonly examples of the penalty function are the L_1 penalty and SCAD penalty (e.g., Fan & Li, 2001; Friedman et al., 2010; Tibshirani, 1996). The L_1 penalty is given by

$$p_\lambda(\boldsymbol{\beta}) = \lambda \sum_l |\beta_l|, \tag{12.14}$$

where l runs through the 2^K elements of $\boldsymbol{\beta}$. The SCAD penalty is defined based on the derivative of the penalty term:

$$\frac{dp_\lambda}{dx}(x) = \lambda \left\{ I(x \leq \lambda) + \frac{\max(0,\ \gamma\lambda - x)}{(\gamma - 1)\lambda} \right\}, \quad \text{for } x > 0$$

where γ is another tuning parameter that is generally suggested to be set at $\gamma = 3.7$ according to a prior simulation study (Fan & Li, 2001). Chen et al. (2015) suggest that the regularized estimator (12.13) admits the oracle property both under the identifiability conditions and other regularity conditions if $\lambda \to 0$ and $\sqrt{N}\lambda \to \infty$ as $N \to \infty$.

12.3.3.3 Q-Matrix Learning for Nonparametric DCMs

The idea of the regularized estimator of Q can be further extended to nonparametric latent class model settings. Fang et al. (2017a) consider a general latent class model with Dirichlet allocation and propose the regularized ML estimator that penalizes complexity of $\boldsymbol{\beta}_k$ at the group level. Revisiting the example of the two-attribute DINA model, suppose a $\boldsymbol{q}$ vector for an item is (1, 0). Then, all the coefficients related to the second attribute will become zero. In such case, it appears natural to group the coefficients of $\boldsymbol{\beta}$ in a way that elements in the same group are either all zero or all nonzero. Specifically, in the simple example (12.12), there are two parameter groups:

$$\theta_1 = (\beta_{k1},\ \beta_{k12}), \quad \text{and} \quad \theta_2 = (\beta_{k2},\ \beta_{k12}),$$

each of which contains all coefficients related to the individual attribute. For each attribute d and item k, we define a vector containing all coefficients related to attribute j as follows:

$$\theta_{kj} = (\beta_{kj},\ \underbrace{\beta_{k1j},\ \ldots,\ \beta_{kjD}}_{\text{2nd order}},\ \underbrace{\beta_{k12j},\ \ldots,\ \beta_{kj(D-1)D}}_{\text{3rd order}},\ \ldots,\ \underbrace{\beta_{k12\ldots D}}_{D\text{th order}}). \tag{12.15}$$

Imposing a group LASSO penalty yields

$$p_\lambda(\boldsymbol{\beta}_k) = \sum_{j=1}^{D} \lambda |\theta_{kj}|$$

where $|\cdot|$ is the usual L_2 norm of a vector. The group LASSO penalty regularizes coefficients such that all elements in the vector θ_{kj} are shrunk toward zero simultaneously. If θ_{kj} is regularized to a zero-vector, then all the coefficients related to attribute j are zero and thus the response distribution of X_k does not depend on the attribute a_j and therefore $q_{kj} = 0$; otherwise, $q_{kj} = 1$. Notice that this approach does not require to specify a particular loading structure and thus is ideal for exploratory study in which the precise model parameterization is undefined.

Once $\hat{\theta}_{kj}$ is obtained as an estimate of θ_{kj}, one can obtain q_{kj} such that $\hat{q}_{kj} = 1$ if $\hat{\theta}_{kj} \neq 0$ and $\hat{q}_{kj} = 0$ otherwise. This map from $\hat{\beta}$ to $\hat{Q}$ is applicable to all DCMs.

12.3.4 Some Related Issues in Latent Variable Selection

The following discusses several issues regarding the implementation of the latent variable selection methods.

12.3.4.1 Selection of a Penalty Parameter λ

To determine λ in the penalty function (12.13), one can consider a model selection criterion. For example, Chen et al. (2015) make use of the solution path that minimizes Bayesian information criterion (Schwarz, 1978)

$$\text{BIC}(\mathscr{M}) = -2l(\hat{\beta}(\mathscr{M})) + |\mathscr{M}| \log N,$$

where $\mathscr{M}$ is the currently fitted model, $l(\hat{\beta}(\mathscr{M}))$ is the maximum log-likelihood under $\mathscr{M}$, and $|\mathscr{M}|$ is the number of free parameters in $\mathscr{M}$. For each λ, the regularized likelihood estimator yields a separate model, denoted by $\mathscr{M}^\lambda$. The tuning parameter λ can be then selected so that it minimizes the BIC:

$$\text{BIC}(\mathscr{M}^\lambda) = -2 \max_{B \in M^\lambda} \{l(B)\} + |\mathscr{M}^\lambda| \log N, \tag{12.16}$$

where $|\mathscr{M}^\lambda|$ equals the number of nonzero coefficients of the regularized estimator corresponding to λ.

12.3.4.2 Computation via EM Algorithm

The optimization of the regularized likelihood function can be implemented applying the standard expectation-maximization (EM) algorithm. Let $l(B;\ \boldsymbol{X}_n,\ \boldsymbol{A}_n,\ n = 1,\ \ldots,\ N)$ denote the compete-data log-likelihood for individual that includes both the observed responses $\boldsymbol{X}_n$ and missing data $\boldsymbol{A}_n$. The EM algorithm is then carried out in two steps: E- and M-step. In the E-step, the expectation of the complete-data log-likelihood is computed with respect to the posterior predictive distribution of the missing data:

$$\mathcal{Q}(B,\ p_{\boldsymbol{\alpha}} \mid p_{\boldsymbol{\alpha}}^{(t)},\ B^{(t)}) = E\big[l(B,\ p_{\boldsymbol{\alpha}};\ \boldsymbol{X}_n,\ \boldsymbol{A}_n,\ n = 1,\ \ldots,\ N) \,\big|\, p_{\boldsymbol{\alpha}}^{(t)},\ B^{(t)},\ \boldsymbol{X}_n,\ n = 1,\ \ldots,\ N\big],$$

where $p_{\boldsymbol{\alpha}}^{(t)}$ and $B^{(t)}$ are the estimated parameter values from step t. Note that the E-step involves a closed-form computation. Due to the local independence assumption of $\boldsymbol{X}_n$ given the latent attribute profile, the log-likelihood functions of the different observations are additive. Furthermore, since the latent attribute profile is defined on a discrete space, the marginal probability distribution is computed explicitly.

The M-step maximizes the $\mathcal{Q}$-function together with the penalty term:

$$(B^{(t+1)},\ p_{\boldsymbol{\alpha}}^{(t+1)}) = \underset{(B,\ p_{\boldsymbol{\alpha}})}{\arg\max}\ \mathcal{Q}(B,\ p_{\boldsymbol{\alpha}} \mid p_{\boldsymbol{\alpha}}^{(t)}, B^{(t)}) - N \sum_{k=1}^{K} p_\lambda(\boldsymbol{\beta}_k).$$

For optimizing the above function with respect to B, the coordinate descent algorithm (Friedman et al., 2010) can be employed that optimizes one parameter at a time. Since the function $\mathcal{Q}$ is convex and differentiable, and p_λ is convex when defined by the L_1 penalty, the coordinate descent algorithm will converge to its maximizer (Tseng, 1988, 2001). Furthermore, each step in the computation is either a closed-form solution or convex optimization.

12.3.4.3 Multi-categorical Responses

The above latent variable selection approach is also applicable to multi-category response data. The basic idea is to make use of the general form of the natural exponential family in (12.10) together with the multidimensional sufficient statistic $T(x)$. For a response taking h different values $\{0,\ 1,\ \ldots,\ h-1\}$, the general form of $T(x)$ is an $(h-1)$-dimensional vector with binary entries. That is, $T(x) = (t_1,\ \ldots,\ t_{h-1})$ and $t_i = 1$ if $x = i$; and $T(x) = \boldsymbol{0}$ if $x = 0$. The corresponding $\boldsymbol{\beta}$ becomes a $2^D \times (h-1)$ matrix, each column of which represents one response type of X. Specifically, $\boldsymbol{\beta} = (\boldsymbol{\beta}^1,\ \ldots,\ \boldsymbol{\beta}^{h-1})$ and each $\boldsymbol{\beta}^i$ is of the same dimension as that in (12.9). Then, the response distribution is

$$P(X = x \mid \boldsymbol{\alpha}) = f(x) \exp\{T(x)\boldsymbol{\beta}^\top \boldsymbol{\alpha} - \varphi(\boldsymbol{\beta}^\top \boldsymbol{\alpha})\}.$$

If one chooses a uniform base function f, the above distribution can be equivalently written in a more friendly form as follows.

$$P(X = i \mid \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{\alpha}^\top \boldsymbol{\beta}^i)}{1 + \sum_{j=1}^{h-1} \exp(\boldsymbol{\alpha}^\top \boldsymbol{\beta}^j)}.$$

With the item response function specified in the above generalized linear form, the regularized likelihood approach can be applied in a similar fashion.

12.4 Latent Variable Selection for MIRT Models

Identifying the item-attribute relationship is also of importance in MIRT models. The key difference between the MIRT models and DCMs is that the former assumes continuous latent traits while the latter assumes discrete latent attributes. Let $\boldsymbol{\theta} \in R^D$ denote the D-dimensional vector consisting of continuous latent trait variables. The item response function of a compensatory two-parameter MIRT model is given by

$$P(X_k = 1 \mid \boldsymbol{\theta},\ \mathbf{a}_k,\ b_k) = \frac{\exp\left(\mathbf{a}_k^\top \boldsymbol{\theta} + b_k\right)}{1 + \exp\left(\mathbf{a}_k^\top \boldsymbol{\theta} + b_k\right)}, \tag{12.17}$$

where $k\ (= 1, \ldots, K)$ indexes the item, $\mathbf{a}_k = (a_{k1}, \ldots, a_{kD})$ is the vector of discrimination parameters, and b_k is known as the easiness parameter (i.e., the complement of the difficulty parameter).

The analysis of an MIRT model largely falls into two categories, exploratory and confirmatory. In exploratory analysis, there is often little to no prior knowledge available, and the parameters are estimated without any constraints. In confirmatory analysis, on the other hand, each item is known to be associated with a distinct subset of latent traits, and this relationship is specified a priori via a factor loading matrix, denoted by here Λ. The confirmatory analysis based on an MIRT model is often considered one of the nonlinear versions of the confirmatory factor analysis. Typical confirmatory analysis requires that Λ be completely specified and that the item parameters be estimated subject to the constraints induced by Λ.

Analogously to the Q-matrix estimation in DCMs, a key question can arise as to how to estimate a Λ-matrix when it is not known or partially known. Sun, Chen, Liu, Ying, and Xin (2016) present a data-driven approach to estimating Λ via a similar regularized estimation method as that for the Q-matrix discussed in the previous section. Let Λ be an incidence matrix that corresponds to the nonzero pattern of the K-by-D factor loading matrix, expressly, $\Lambda = (\lambda_{kd})_{K \times D}$, where $\lambda_{kd} = \mathbb{I}(a_{kd} \neq 0)$ with $\mathbb{I}$ being the indicator function. Assuming conditional independence of item responses given $\boldsymbol{\theta}_n$, the complete-data likelihood for the two-parameter logistic MIRT is given by

$$L(\boldsymbol{A},\, \boldsymbol{b};\, \boldsymbol{X}_n) = \int_{\boldsymbol{\Theta}} p_{\boldsymbol{\theta}_n} \left[\prod_{k=1}^{K} P_k(\boldsymbol{\theta}_n)^{X_{nk}} \big(1 - P_k(\boldsymbol{\theta}_n)\big)^{1-X_{nk}} \right] d\boldsymbol{\theta}_n,$$

where $\boldsymbol{A} = (\boldsymbol{a}_k^\top : k = 1, \ldots, K)$ is the K-by-D matrix of item factor loadings, $\boldsymbol{b}$ is the vector of item intercept parameters with length K, $P_k(\boldsymbol{\theta}_n) = P(X_k = 1 \,|\, \boldsymbol{\theta}_n, \boldsymbol{a}_k, b_k)$, and $p_{\boldsymbol{\theta}_n}$ is the prior distributional weight of $\boldsymbol{\theta}_n$. Analogously with (12.13), the L_1-regularized estimator is obtained as

$$(\hat{\boldsymbol{A}},\, \hat{\boldsymbol{b}})_\eta = \underset{(\boldsymbol{A},\, \boldsymbol{b})}{\arg\max}\; l(\boldsymbol{A},\, \boldsymbol{b};\, \boldsymbol{X}_n,\, \forall n) - N\delta \sum_{k=1}^{K} \sum_{d=1}^{D} |a_{kd}|,$$

where $l(\boldsymbol{A},\, \boldsymbol{b};\, \boldsymbol{X}_n,\, \forall n) = \sum_{n=1}^{N} \log L(\boldsymbol{A},\, \boldsymbol{b};\, \boldsymbol{X}_n)$.

Similar to the previous discussion, the value of δ is obtained as the one that minimizes the BIC. The computation of $(\hat{\boldsymbol{A}},\, \hat{\boldsymbol{b}})$ follows the similar scheme as presented in Sect. 12.3.3 while some modifications are needed to address the issues related to the continuous latent variables (see Sun et al., 2016 for details). Once the model parameter estimates are obtained, the Λ-matrix can be simply obtained as $\hat{\lambda}_{kd} = \mathbb{I}(\hat{a}_{kd} \neq 0)$.

12.5 Discussion

The present chapter has focused on an empirical approach for constructing a Q-matrix. In contrast to traditional methods, which utilize experts' prior knowledge and thus can be subject to decision error, the methods presented in this chapter are driven by observed response data and can result in consistent estimation of Q. The chapter has also introduced theoretical results concerning identifiability of a Q-matrix as well as several estimation methods. The estimation of Q was in particular cast as a variable selection problem, and the corresponding estimator was derived from the regularized likelihood function that penalizes the complexity in Q. While the usual inference methods typically induce considerable computational intensity, the procedures introduced here are computationally workable and have mathematically sound properties.

Throughout the chapter, the estimators are obtained assuming no supplemental prior knowledge regarding the Q-matrix. It is however possible in practice to impart experts' knowledge in the subject matter for drafting a Q-matrix. Such information, if correct, cannot only serve as a guideline for estimating the remaining unknown parameters but also substantially reduces the computational load in that one only needs to search in the vicinity of the pre-constructed Q. Furthermore, the Q-matrix incorporating experts' opinion can oftentimes lead to better interpretation of the estimated parameters. In applications, the introduced procedures can be customized to combine the prior knowledge. If a Q matrix can be completely pre-specified,

one can utilize this matrix as a starting point of the algorithm. If the pre-specified matrix is largely correct, then the algorithm will converge very fast. If the matrix is partially identifiable, one may impose regularization on the portion of parameters that correspond to the unspecified elements of Q. This type of analysis is often case-by-case.

References

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2017). Bayesian estimation of the DINA Q matrix. *Psychometrika*. https://doi.org/10.1007/s11336-017-9579-4

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*, 850–866.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*, 598–618.

Chiu, C.-Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447–468.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association, 96*(456), 1348–1360. Retrieved from http://www.jstor.org/stable/3085904

Fang, G., Liu, J., & Ying, Z. (2017a). Latent variable selection via overlap group LASSO with applications to cognitive assessment. Preprint.

Fang, G., Liu, J., & Ying, Z. (2017b). On the identifiability of diagnostic classification models. ArXiv e-prints. Retrieved from https://arxiv.org/abs/1706.01240

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22.

Haertel, E. H. (1989). Using restricted latent class models to map the attribute structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Köhn, H.-F., & Chiu, C.-Y. (2016). A proof of the duality of the dina model and the dino model. *Journal of Classification, 33*, 171–184.

Köhn, H. F., & Chiu, C.-Y. (2017). A procedure for assessing the completeness of the q-matrices of cognitively diagnostic tests. *Psychometrika, 82*, 112–132.

Liu, J. (2017). On the consistency of Q-matrix estimation: A commentary. *Psychometrika, 82*, 523-527.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli, 19*, 1790–1817.

Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 19*, 461–464.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika, 81*(4), 921-939. https://doi.org/10.1007/s11336-016-9529-6

Templin, J. L., & Henson, R. A. (2006). A Bayesian method for incorporating uncertainty into Q-matrix estimation in skills assessment. In *Symposium Conducted at the Meeting of the American Educational Research Association*, San Diego, CA.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, 58*, 267–288.

Tseng, P. (1988). *Coordinate ascent for maximizing nondifferentiable concave functions* (Technical Report LIDS-P, 1840). Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Tseng, P. (2001). Convergence of block coordinate descent method for nondifferentiable maximization. *Journal of Optimization Theory and Applications, 109*, 474–494.

von Davier, M. (2005, September). *A general diagnostic model applied to language testing data* (Research report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–301.

Chapter 13
Global- and Item-Level Model Fit Indices

Zhuangzhuang Han and Matthew S. Johnson

Abstract One of the primary goals in cognitive diagnosis is to use the item responses from a cognitive diagnostic assessment to make inferences about what skills a test-taker has. Much of the research to date has focused on parametric inference in cognitive diagnosis models (CDMs), which requires that the parametric model used for inference does an adequate job of describing the item response distribution of the population of examinees being studied. Whatever the type of model misspecification or misfit, users of CDMs need tools to investigate model-data misfit from a variety of angles. In this chapter we separate the model fit methods into four categories defined by two aspects of the methods: (1) the level of the fit analysis, i.e., global/test-level versus item-level; and (2) the choice of the alternative model for comparison, i.e., an alternative CDM (relative fit) or a saturated categorical model (absolute fit).

13.1 Introduction

One of the primary goals in cognitive diagnosis is to use the item responses from a cognitive diagnostic assessment to make inferences about what skills a test-taker has. Much of the research to date has focused on parametric inference in cognitive diagnosis models (CDMs), which requires that the parametric model used for inference does an adequate job of describing the item response distribution of the population of examinees being studied. Given the importance of data-model fit

Z. Han
Department of Human Development, Teachers College, Columbia University, New York, NY, USA
e-mail: zh2198@tc.columbia.edu

M. S. Johnson (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: msjohnson@ets.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_13

for inference, it is necessary to have methods for investigating the ability of a model to fit observed data from an assessment.

Misfit for CDMs can come from a variety of sources. Incorrectly specifying the form of item-response model (e.g., DINA versus DINO) and using it for inference is a typical source of misfit. In some cases misfit might stem from some common assumptions of the model. For example, the *local independence* assumption presumes that items on the assessment are conditionally independent given the skills being measured, that is, given a specific latent attribute class. Such assumptions may be too strong to fit the actual data. Also, there are some specific types of misfit for CDMs. For instance, the misspecification of the Q-matrix and the distribution of the latent attribute pattern (e.g., the number of attributes and the hierarchy among skills). Whatever the type of model misspecification or misfit, users of CDMs need tools to investigate model-data misfit from a variety of angles.

In this chapter we separate the model fit methods into four categories defined by two aspects of the methods: (1) the level of the fit analysis, i.e., global/test-level versus item-level; and (2) the choice of the alternative model for comparison, i.e., an alternative CDM (relative fit) or a saturated categorical model (absolute fit).

Global model fit has been a major focus in preceding research (de la Torre & Douglas, 2008; Sinharay & Almond, 2007). In this category, global relative fit utilizes conventional information-based indices to select one from several models. In contrast, global absolute fit tries to find out how exact the model reproduces the observed data by examining squared-residual based statistics (e.g., model-level χ^2, G^2 and root mean square error of approximation (RMSEA)) or non-inferential indices (e.g., mean absolute difference (MAD)). Typically, these measures can serve as general-purpose statistics to test the model assumptions such as specification of the model parametric form, the *local independence*, specification of the Q-matrix and the dimensionality.

Some additional attention should be drawn on the issue of Q-matrix specification. The Q-matrix often constructed by domain experts could be misspecified and then result in model misfit. Q-matrix refinement and validation methods (de la Torre & Chiu, 2016; Chiu, 2013) have shown promising empirical performance in addressing this concern. However, the problem of Q-matrix misspecification and refinement should not be isolated from the Q-matrix learning and identification (see Chaps. 12 and 16 in this volume). An integrated view on these problems is constructive to the understanding of CDMs.

Item-level fit analysis focuses on the local misfit caused by the misspecified parametric form of individual or subsets of items. It allows practitioners to identify these aberrant items and provide guidance about how to refine the measurement instrument. Such analyses have been supported by recent empirical evidence (de la Torre & Lee, 2013; de la Torre, van der Ark, & Rossi, 2015) showing that the assessment with items assumed to follow different models (e.g., including both DINA and DINO items) instead of uniformly having a single form might better fit the real data. To achieve the refinement, item-level relative fit methods offer a way

by comparing nested models via Likelihood Ratio (LR), Wald (W), and Lagrange multiplier (LM) tests. Absolute fit methods can be adapted to the item-level as well. For example, item-level goodness-of-fit statistics (Orlando & Thissen, 2000; Wang, Shu, Shang, & Xu, 2015) are constructed on the basis of the squared residual of observed and expected proportion of correctness that are obtained by grouping respondents. Different grouping strategies lead to various types of fit statistics, which has been a focus in recent studies. Item-level absolute fit statistics can also be extended to detect misfit for item pairs or triplets. It is particularly useful if one is interested in locating the source of misfit and taking remedial action when the global model test identifies the existence of overall misfit and *local dependence* is the potential culprit.

It is also worth mentioning that person-fit analysis, which is not discussed in this chapter, offers another perspective to investigate model-data fit. Person-fit methods focus on the misfit happening in individual response vectors. Theses methods are helpful to detect aberrant test-taking behaviors such as cheating or a speeded test. Several person-fit indices and tests have been proposed particularly for CDMs such as the hierarchy consistency index (Cui & Leighton, 2009) and the generalized LR test (Liu, Douglas, & Henson, 2009). Well-developed person-fit analysis in other latent variable models such as the item response theory (IRT) (Meijer & Sijtsma, 2001) might be adaptable to CDMs.

In this chapter we restrict our focus on relative and absolute model-level and item-level fit. After reviewing several methods that have been introduced in the literature, we illustrate the implementations of some methods by a real-data example. The review is outlined along the categories we defined previously. For each category of methods, pros and cons are discussed based on current simulation studies. Some general guidance on which measures one should use is included as well.

13.2 The Model Framework

Here we use the generalized DINA (G-DINA) model (de la Torre, 2011) as the basic framework to discuss model fit methods. As other general CDMs, such as the general diagnostic model (von Davier, 2008) and the log-linear CDM (LCDM) (Henson, Templin, & Willse, 2009), the G-DINA model relates several CDMs by its flexible parameterization. See Chaps. from 6, 7, and 8 for more details about those general frameworks.

The G-DINA model requires a $K \times D$ Q-matrix (with binary elements $\{q_{kd}\}$). The required number of attributes for item k is D_k^*, where $D_k^* = \sum_{d=1}^{D} q_{kd}$. Such representation efficiently reduces the attribute vector of item k from $\boldsymbol{a}_l = (a_{l1}, a_{l2} \ldots, a_{lD})$ to $\boldsymbol{a}_{lk}^\star = (a_{l1}^\star, a_{l2}^\star, \ldots, a_{lD_i^*}^\star)$, where the number of classes partitioned by item k is reduced from 2^D to $2^{D_k^*}$. For example, if $D = 3$ and the

kth has q-vector $\boldsymbol{q}_k = (1, 1, 0)^\top$, then the full attributes vectors $\boldsymbol{a}_l = (0, 1, 0)$ and $\boldsymbol{a}_{l'} = (1, 1, 0)$ are simplified as reduced vectors $\boldsymbol{a}^\star_{lk} = (0, 1)$ and $\boldsymbol{a}^\star_{l'k} = (1, 1)$. The probability of respondents with latent profile $\boldsymbol{a}^\star_{lk}$ answering item k correctly is denoted by $P(X_k = 1|\boldsymbol{a}^\star_{lk}) = P(\boldsymbol{a}^\star_{lk})$, more specifically,

$$P(\boldsymbol{a}^*_{lk}) = \delta_{k0} + \sum_{d=1}^{D^*_i} \delta_{kd} a^\star_{ld} + \sum_{d=1}^{D^*_k} \sum_{d'=d+1}^{D^*_k} \delta_{kdd'} a^\star_{ld} a^\star_{ld'} + \cdots + \delta_{k12\ldots D^*_k} \prod_{d=1}^{D^*_k} a^\star_{ld}, \quad (13.1)$$

where δ_{k0} is the intercept for item i; δ_{kd} is the main effect due to a_d; $\delta_{kdd'}$ and $\delta_{k12\ldots D^*_k}$ are interactions for the two-way and other higher orders among $a_1, \ldots, a_{D^*_k}$. In some applications, monotonicity constraints are imposed on item parameters to make sure that subjects having more skills answer an item correctly with the probability no less than those having fewer skills. See Chap. 7 for details on the monotonicity constraints. The model representation (13.1) uses a *identity-link* function which can be simply modified and extended through putting other transform functions (e.g., *logistic-* and *log-link*) on $P(\boldsymbol{a}^\star_{lk})$.

It is not hard to tell the flexibility of such a formulation. For example, the DINA model can be obtained by using identity-link function and setting all parameters to 0 except for δ_{k0} and $\delta_{i12\ldots D^*_k}$; in which case the guessing parameter follows $g_k = \delta_{k0}$ and the slipping parameter satisfies $s_k = 1 - \delta_{k0} + \delta_{k12\ldots D^*_k}$. Notice that the flexibility enables us to summarize and estimate the parameters of multiple CDMs by a single parametric framework. It also provides a convenient basis for comparing nested models and allows us to examine one item at a time.

13.3 Relative Fit Indices

Relative fit indices evaluate the fit of a model compared to some competing model. In the following two subsections, we first review the indices suitable for the global-level fit and then look at how some of them can be used at the item-level.

13.3.1 Global Level

One way to evaluate the comparative fit of a model relative to a competing model, when it is a nested model, is the likelihood ratio test (LRT). A nested model is one that can be defined by enforcing some constraints on some of the model parameters. For example, within the G-DINA framework, the DINA model is nested within the G-DINA model because it can be obtained by setting all coefficients other than the intercept and the highest-order interaction term equal to zero. The LRT compares

the fit of the two models by comparing the log-likelihoods ℓ_r and ℓ_f evaluated at the maximum likelihood estimates (MLEs) for the reduced and full models respectively, where the log-likelihood is defined as

$$\ell(\boldsymbol{X}|\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{n=1}^{N} \log \sum_{l=1}^{L} p(\boldsymbol{a}_l|\boldsymbol{\gamma}) \prod_{k=1}^{K} P(\boldsymbol{a}_{lk}^{\star})^{X_{nk}} \left[1 - P(\boldsymbol{a}_{lk}^{\star})\right]^{(1-X_{nk})}, \tag{13.2}$$

where N is the number of participants and $L = 2^D$; $p(\boldsymbol{a}_l|\boldsymbol{\gamma})$ is the prior probability of $\boldsymbol{a}_l$. The item response probability $P(\boldsymbol{a}_{lk}^{\star})$ is obtained by compressing $\boldsymbol{a}_{lk}$ as shown in the previous section. The MLEs of the item parameter vector, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K)$, and the latent class proportion parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{L'})$ ($L' = L$ and $p(\boldsymbol{a}_l|\boldsymbol{\gamma}) = \gamma_l$ if an unrestricted attribute space is assumed) can be estimated with an expectation-maximization (EM) algorithm (de la Torre, 2011; George, Robitzsch, Kiefer, Groß, & Ünlü, 2016) for example.

The likelihood ratio test statistic that is typically used is two times the difference between the log-likelihoods,

$$\lambda = 2\left(\ell_f(\boldsymbol{X}|\boldsymbol{\delta}_f, \boldsymbol{\gamma}_f) - \ell_c(\boldsymbol{X}|\boldsymbol{\delta}_c, \boldsymbol{\gamma}_c)\right),$$

in the case where observations have been randomly sampled, the statistic λ is approximately chi-squared distributed when the reduced model is the correct model; the degrees of freedom of the distribution is equal to the difference in the number of parameters in the two models. For example, if the full model is the G-DINA model and the reduced model is the DINA, the number of parameters are $p_f = \sum_{k=1}^{K} 2^{D_k^*} + L - 1$ and $p_r = 2K + L - 1$ respectively.

The LRT has a couple of limitations. First, according to the old adage, 'all models are wrong', the LRT tends to find evidence against simpler models when the sample size N is large. Second, the likelihood ratio test requires the reduced model to be nested within the full model framework.

Two information-based criteria attempting to address these issues are Akaike's information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz, 1978), which are defined as

$$AIC = -2\ell(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}}) + 2p$$

$$BIC = -2\ell(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\gamma}}) + p\ln(N),$$

To use AIC and/or BIC for model evaluation, the user should estimate multiple competing models. In both cases, the model that should be selected is the one that minimizes the criterion. So, if one is interested in whether the DINA model fit a specific data set, the researchers would fit the DINA model and other candidates from the G-DINA framework and then check if the AIC and/or BIC for the DINA model is the smallest.

The difference between the penalty terms makes the BIC penalize the model with a larger number of parameters more than the AIC does. This is partially due to the purposes of each; AIC attempts to find the model that best predicts future observations, whereas BIC attempts to quantify evidence for a model in model-selection problems. Kunina-Habenicht, Rupp, and Wilhelm (2012) found that AIC and BIC are effective in selecting the model with a correctly specified Q-matrix against those with misspecified Q-matrices within the framework of the log-linear CDM. Chen, de la Torre, and Zhang (2013) showed that AIC and BIC perform well in selecting among nested models within the G-DINA framework.

Another way to compare non-nested model is the log-penalty index (Gilula & Haberman, 1994) which is obtained by dividing the AIC by the number of observations in the sample. It is more like the BIC penalizing the number of parameters while accounting for the sample size. The index has been used in comparing models within the framework of GDM (von Davier, 2008) and shown promise of its implementation.

The LRT, AIC, BIC and log-penalty index all require MLEs for the model parameters, and thus are used in frequentist applications. The deviance information criterion (Spiegelhalter, Best, Carlin, & van Der Linde, 2002) and the Bayes factor (Kass & Raftery, 1995), in contrast, are applicable for global relative fit within the Bayesian modeling framework. The DIC is defined as

$$DIC = \bar{D} + p_D,$$

where $\bar{D}$ is the expectation of $-2\ell(\boldsymbol{\delta}, \boldsymbol{\gamma})$ over the joint posterior distribution of $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ given the observed assessment data. The quantity $p_D = \bar{D} - 2\ell(\bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\gamma}})$, where $(\bar{\boldsymbol{\delta}}, \bar{\boldsymbol{\gamma}})$ are the posterior mean vectors is a measure of the complexity of the Bayesian model.

The Bayes factor is the Bayesian analog to the frequentist LRT.

$$BF_{12} = \frac{P(\boldsymbol{X}|\mathscr{M}_1)}{P(\boldsymbol{X}|\mathscr{M}_2)}$$

where

$$P(\boldsymbol{X}|\mathscr{M}_m) = \int \exp\left[\ell(\boldsymbol{\delta}_m, \boldsymbol{\gamma}_m)\right] p(\boldsymbol{\delta}_m, \boldsymbol{\gamma}_m|\mathscr{M}_m) d\boldsymbol{\delta}_m d\boldsymbol{\gamma}_m$$

and $p(\boldsymbol{\delta}_m, \boldsymbol{\gamma}_m|\mathscr{M}_m)$ is the joint prior density of parameters from the mth model. In most applications exact calculation of the Bayes factor is difficult or impossible. A possible approach for approximating the marginal likelihoods needed to calculate the Bayes factor is with the Laplace-Metropolis estimator as proposed by Raftery (1996).

DIC and Bayes factors have been suggested for use with CDMs. For example, de la Torre and Douglas (2004, 2008), and Sinharay and Almond (2007) showed

the effectiveness of these statistics for model selection. Specifically, de la Torre and Douglas (2004, 2008) implemented the Bayes factor to compare the Higher-order DINA and multiple-strategy DINA models to the traditional DINA model.

13.3.2 Item Level

The G-DINA framework allows us to evaluate the parametric form of an assumed CDM used at the item-level by performing specific hypothesis tests. In these hypothesis tests, the null hypothesis (H_0) assumes the reduced model (e.g., DINA) is correct and the alternative (H_1) states that the general (or full) model (e.g., G-DINA) is correct. Notice that this test does not touch the Q-matrix. Thus, the size of parameter space for the full model is determined by the number of skills required by the item. Let's say, for instance, the Q-matrix specifies up to 3 skills but the item only requires 2 skills. The full model of the item can have up to 4 parameters according to the Eq. (13.1): an "intercept", two "main effect", and an "interaction".

The likelihood ratio (LR) introduced earlier for model-level fit evaluation could be applied to item-level fit by fitting the assumed model as the reduced model, and a second model that assumes a G-DINA structure for that item. To check the fit of all K items would require estimating $K + 1$ models, the reduced model, and a separate "full" model for each item; this issue somewhat limits the use of the likelihood ratio statistic for item-level evaluation.

Unlike the LR statistic and testing procedure, the Lagrange multiplier (LM), or score test only requires estimation of the reduced model, which makes it particularly useful for evaluating item-level fit of a model. The general idea of the score test is that if the null hypothesis is correct, then the first derivative of the full model likelihood evaluated at the reduced model maximum likelihood estimates should be close to zero. If $\hat{\boldsymbol{\delta}}_k^0$ denotes the maximum likelihood estimator of the item parameters for item k under the reduced model, then the LM statistic is

$$LM = \left[\frac{\partial \ell_f(\boldsymbol{\delta}_k)}{\partial \boldsymbol{\delta}_k} \mid_{\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_k^0}\right]^T \boldsymbol{I}^{-1}(\boldsymbol{\delta}_k) \left[\frac{\partial \ell_f(\boldsymbol{\delta}_k)}{\partial \boldsymbol{\delta}_k} \mid_{\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_k^0}\right], \tag{13.3}$$

where $\boldsymbol{I}(\boldsymbol{\delta}_k) = V\left[\frac{\partial \ell_f(\boldsymbol{\delta}_k)}{\partial \boldsymbol{\delta}_k} \mid_{\boldsymbol{\delta}_k = \hat{\boldsymbol{\delta}}_k^0}\right]$ is the information matrix (from the full model) for the item parameter vector $\boldsymbol{\delta}_k$ evaluating at $\hat{\boldsymbol{\delta}}_k^0$; in practice the information matrix is approximated using the observed information matrix $\boldsymbol{I}(\hat{\boldsymbol{\delta}}_k^0)$. Under the null hypothesis the distribution of the LM approaches a chi-squared distribution with degrees of freedom equal to $df = p_f - p_r$, where p_f and p_r, by an abuse of the notation, denote the number of item parameters for the item k in the full and reduced models.

The LRT and the Lagrange multiplier test are asymptotically equivalent to one another, so the results tend to be similar for large sample sizes. A third

asymptotically equivalent test statistic is the Wald test statistic. The Wald test for item-level model fit assessment requires fitting the full model (e.g., G-DINA) in order to evaluate the fit of the reduced model (e.g., DINA). As discussed earlier, the DINA model can be obtained from the G-DINA model by assuming all parameters other than the intercept and the highest-order interaction term are equal to zero. For example, suppose we have an item measuring two skills. Then the full model parameter vector is $\boldsymbol{\delta}_k = (\delta_{k0}, \delta_{k1}, \delta_{k2}, \delta_{k12})^\top$; the test to evaluate fit of the DINA model assumes a null hypothesis of the form $H_0 : \boldsymbol{\delta}_k = (\delta_{k0}, 0, 0, \delta_{k12})^\top$, or equivalently $H_0 : \boldsymbol{R}_k\boldsymbol{\delta}_k = (0, 0)^\top$, where $\boldsymbol{R}_k$ is the restriction matrix

$$\boldsymbol{R}_k = \begin{pmatrix} 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix}.$$

For general models R_k is a $(p_f - p_r) \times p_f$ matrix describing the null model restrictions; see de la Torre (2011) for examples. The Wald test is then defined

$$W = \left[\boldsymbol{R}_k\hat{\boldsymbol{\delta}}_k^1\right]^\top \left[\boldsymbol{R}_k V(\hat{\boldsymbol{\delta}}_k^1)\boldsymbol{R}_k^\top\right]^{-1} \left[\boldsymbol{R}_k\hat{\boldsymbol{\delta}}_k^1\right], \tag{13.4}$$

where $\hat{\boldsymbol{\delta}}_k^1$ is the maximum likelihood estimator under the full model (H_1) and $V(\hat{\boldsymbol{\delta}}_k^1)$. It should be noted that $V(\hat{\boldsymbol{\delta}}_k^1)$ is the sub-matrix of the covariance matrix of the MLEs for all item parameters and latent attribute distribution parameters. The covariance matrix is usually approximated with the inverse of the observed information matrix. The asymptotic distribution under the null hypothesis is also $\chi^2_{(p_f - p_r)}$.

Simulation studies by Torre and Lee (2013) and Sorrel, Abad, Olea, de la Torre, and Barrada (2017) shown the statistics have accurate Type I error rates and high power with large N and small D for typical significance levels. Sorrel et al. (2017) found that the LR and Wald tests perform better than the Lagrange multiplier test in terms of the Type I error and power across cases with $N \leq 1000$, $K \leq 36$ and $D = 4$. However, all statistics are found (Sorrel et al., 2017; Ma, Iaconangelo, & de la Torre, 2016) to be highly affected when items have low discrimination.

13.4 Absolute Fit Indices

In this section we start by reviewing the global-level statistics and then move to introduce the item-level indices. A review of posterior predictive methods assessing model misfit within the Bayesian approach is included as the end of the section.

13.4.1 Global Level

Classical goodness-of-fit (GOF) statistics such as Pearson's χ^2 and the likelihood ratio G^2 are fundamental overall fit indices in categorical data analysis. For a test with K dichotomous items,

$$\chi^2 = N \sum_{c=1}^{2^K} \frac{(p_c - \hat{\pi}_c)^2}{\hat{\pi}_c} \quad \text{and} \quad G^2 = 2N \sum_{c=1}^{2^K} p_c \ln(\frac{p_c}{\hat{\pi}_c})$$

where p_c and $\hat{\pi}_c$ are the observed and model-based expected proportions for one cell c in the 2^K contingency table (for all possible response patterns). The model-based proportions, $\hat{\pi}_c$, is calculated by the marginal likelihood in the right-hand side of (13.2) with estimated parameters. For small K and under the null hypothesis that the assumed CDM is the correct model, the statistics follow the chi-square distribution with *degrees of freedom* (df) $2^K - p - 1$, where p is the total number of model parameters.

These full-information statistics suffer from the problem of sparsity when the number of items, K, is large or when the sample size, N, is small, which leads to unknown asymptotic properties. One way out of this is using the resampling and bootstrapping techniques but the computation is prohibitively expensive. Maydeu-Olivares and Joe (2005) introduced the limited-information family of statistics to address such concerns for item response models in general. Hansen, Cai, Monroe, and Li (2016) and Liu, Tian, and Xin (2016) have implemented statistics in this family to evaluate global fit for CDMs.

The idea is to utilize the up-to-rth-order moments, $\boldsymbol{\pi}_r$, rather than the proportions of all possible response patterns (or referred as all cells in the contingency table, $\boldsymbol{\pi}$), to formulate the fit statistic. For instance,

$$\boldsymbol{\pi_2} = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \dot{\pi}_{12} \\ \dot{\pi}_{13} \\ \dot{\pi}_{23} \end{pmatrix} = \boldsymbol{T}_2\boldsymbol{\pi} = \begin{pmatrix} 0\,1\,0\,0\,1\,1\,0\,1 \\ 0\,0\,1\,0\,1\,0\,1\,1 \\ 0\,0\,0\,1\,0\,1\,1\,1 \\ 0\,0\,0\,0\,1\,0\,0\,1 \\ 0\,0\,0\,0\,0\,1\,0\,1 \\ 0\,0\,0\,0\,0\,0\,1\,1 \end{pmatrix} \begin{pmatrix} \pi_{000} \\ \pi_{100} \\ \pi_{010} \\ \pi_{001} \\ \pi_{110} \\ \pi_{101} \\ \pi_{011} \\ \pi_{111} \end{pmatrix} \tag{13.5}$$

for the case of $K = 3$; $\boldsymbol{T}_2$ is the matrix transforming $\boldsymbol{\pi}$ to $\boldsymbol{\pi}_2$. The limited-information statistic M_r is written as

$$M_r = N(\boldsymbol{p}_r - \hat{\boldsymbol{\pi}}_r)^T \hat{\boldsymbol{C}}_r (\boldsymbol{p}_r - \hat{\boldsymbol{\pi}}_r)$$

on the basis of the up-to-rth moments. Given a specified CDM model being the null model, M_r follows the chi-square distribution with $df = s_r - p$, where $s_r = \sum_{i=1}^{r} \binom{K}{i}$ is the number of elements in $\boldsymbol{\pi}_r$. The detailed derivation of $\hat{\boldsymbol{C}}_r$ is described in Maydeu-Olivares and Joe's (2005) paper.

Hansen et al. (2016) and Liu et al. (2016) examined the limited information statistic for the evaluation of CDMs. Simulations in both studies show that M_2 has more stable performance in detecting misfit simulated from Q-matrix misspecification than χ^2 and G^2 for moderate sample sizes. Hansen et al. (2016) also found that M_2 is sensitive to misfit from item-level model misspecification and to violations of local independence, but insensitive to the misspecification of the higher-order structure of the attributes.

One of the shortcomings of the goodness-of-fit is that they treat the null hypothesis as the desired model, and the alternative model as the saturated model. In practice the true data generation process is likely to be more complex than any assumed model, and therefore will be rejected with a sufficiently large sample size. To deal with this issue, Browne and Cudeck (1992) introduced the root mean squared error of approximation (RMSEA), which attempts to measure the discrepancy between the population ($\boldsymbol{\pi}_T$) and the null model ($\boldsymbol{\pi}_0$) probability vectors.

$$\text{RMSEA} = \sqrt{\max\left(\frac{\hat{\chi}^2 - df}{N \times df},\ 0\right)}$$

where $\hat{\chi}^2$ is the observed χ^2 statistic for the data set. Maydeu-Olivares and Joe (2014) give the limited-formation version is

$$\text{RMSEA}_r = \sqrt{\max\left(\frac{\hat{M}_r - df_r}{N \times df_r},\ 0\right)}.$$

The 90% of confidence interval of RMSEA_r is derived from the non-central chi-square distribution $F_{\chi^2}(\hat{M}_r; df_r)$. Maydeu-Olivares and Joe (2014) show that RMSEA_r ($r \leq 3$) has more accurate confidence intervals than RMSEA when $2^K > 300$ using simulations generated under dichotomous IRT models.

In practice, the cut-off values for RMSEA are suggested to determine the degree of fit. For example, Oliveri and von Davier (2011) suggest using $\text{RMSEA}_1 > 0.1$ as poor fit when they measure the item-level misfit for the PISA (Programme for international Student Assessment) data with the GDM; Liu et al. (2016) recommend the cut-off values (less than) 0.030 and 0.045 for RMSEA_2 as "excellent" and "good" fit under the LCDM.

Using item-level and item-pairwise fit indices to assess overall misfit has been suggested in the item response literature. Two examples are

$$\text{MAD}_k = |\dot{p}_k - \hat{\dot{\pi}}_k|,$$

$$\chi^2_{kk'} = N \sum_{x_k=0}^{1} \sum_{x_{j'}=0}^{1} \frac{(p_{x_k x_{k'}} - \hat{\pi}_{x_k x_{k'}})^2}{\hat{\pi}_{x_k x_{k'}}},$$

where $\hat{\hat{\pi}}_k$ is the model-implied proportion of answering the item k correctly; $\hat{\pi}_{x_k x_{k'}}$ is the expected probability of cell in the bivariate table for item k and k'; $\dot{p}_k$ and $p_{x_k x_{k'}}$ are observed probabilities. Other item-pairwise indices not described in this chapter, such as the Fisher transformation of item-pair correlations and the item-pairwise log-odds ratio, were studied in Chen et al. (2013). Researchers (Chen et al.; Lei & Li, 2016) applied these statistics by simply taking the average or conducting Bonferroni-adjusted multiple comparisons. Both studies show that the pairwise fit indices perform with better power in detecting global misfit than item-level statistics fit do.

13.4.2 Item Level

Squared-residual based statistics play a vital role in item-level fit. To collect the squared residuals, we partition the test-takers into groups by certain criteria. Once the groups are given, we can calculate o_{ks} and e_{ks} denoting the observed and expected proportion of answering the item k right for the test-takers in group s. It's easy to see that different grouping variables lead to different statistics.

Yen (1981) proposed Q_1 by grouping the test-takers according to their latent abilities. In the context of CDMs, the examinees are grouped by their attribute patterns. In practice the assignment of a subject to her latent attribute class is given by the posterior $P(\hat{\boldsymbol{a}}_l|\boldsymbol{x}_n)$ where $\hat{\boldsymbol{a}}_l$ and $\boldsymbol{x}_n$ are the attribute pattern l and response vector for subject n. Yen (1981) approximated the limiting distribution of Q_1 by the chi-square distribution with df $2^D - p_k - 1$, where p_k is the number of parameters for item k. The statistic is criticized for two reasons. First, some latent attribute classes are extremely rare, especially when D is large, which means that almost no test-taker will be assigned in these classes. Some researchers suggested binning the classes to reduce the effect of sparsity. But how to bin them appropriately is still a complex question. Second, the uncertainty of the class assignment is not considered in the approximation of Q_1's limiting distribution.

$S - \chi^2_k$ and $S - G^2_k$ proposed by Orlando and Thissen (2000) address these problems. The statistics are defined as

$$S - \chi^2_k = \sum_{s=1}^{S-1} N_s \frac{(o_{ks} - e_{ks})^2}{e_{ks}(1 - e_{ks})}$$

$$S - G^2_k = 2 \sum_{s=1}^{S-1} N_s \left[o_{ks} \log\left(\frac{o_{ks}}{e_{ks}}\right) + (1 - o_{ks}) \log\left(\frac{1 - o_{ks}}{1 - e_{ks}}\right) \right]$$

where s indicates the group of test-takers who score s; N_s is the number of examinees in group s; o_{ks} and e_{ks} are what we define before; e_{ks} is calculated as

$$e_{ks} = \frac{\sum_{l=1}^{2^D} P(X_{ik} = 1|\boldsymbol{a}_l) P\left(S^{(-k)} = s - 1|\boldsymbol{a}_l\right) p(\boldsymbol{a}_l)}{\sum_{l=1}^{2^D} P\left(S = s|\boldsymbol{a}_l\right) p(\boldsymbol{a}_l)}.$$

$P\left(S^{(-k)} = s - 1|\boldsymbol{a}_l\right)$ is recursively computed via the algorithm developed by Lord and Wingersky (1984) as described in detail by Orlando and Thissen (2000).

Orlando and Thissen (2000) approximated the distribution of $S - \chi_k^2$ and $S - G_k^2$ by the chi-square with $df = K - 1 - p_k$, where p_k is the number of item parameters for item k. Notice that the squared residuals are grouped by (or conditioned on) raw scores rather than by estimated latent ability groups. Orlando and Thissen (2000) in simulation studies showed that their statistics have more sensible Type-I error than Q_1 does.

Wang et al. (2015) suggested applying Stone's method (2000) in Q_1 to take the uncertainty of $\hat{\boldsymbol{a}}_l$ into consideration. Instead of using observed counts grouped by point estimated $\hat{\boldsymbol{a}}_l$ to create squared residuals in Q_1, Stone (2000) computed

$$O_{kl}^* = \sum_{n=1}^{N} x_{nk}\, p(\hat{\boldsymbol{a}}_l|\boldsymbol{x}_n)$$

to account for the uncertainty by the posterior distribution of $\hat{\boldsymbol{a}}_l$. The chi-square is no longer a good approximation to the limiting distribution of the new statistic given the dependence among examinees introduced from $p(\hat{\boldsymbol{a}}_l|\boldsymbol{x}_n)$. A Monte Carlo resampling technique is suggested to construct the distribution for inference.

Simulation studies conducted by Wang et al. (2015) show that Stone's Q_1 has more promising power and Type I error than its original counterpart to detect Q-matrix and model-type misspecification under the DINA model. One of the drawbacks of Stone's methods is that they are computationally expensive. Sorrel et al. (2017) noted that $S - \chi_k^2$ avoids inflated Type I error when detecting the item-level misfit within the G-DINA. But they also remarked that the power of $S - \chi_k^2$ is quite unacceptable in many cases.

13.4.3 Posterior Predictive Assessment

The posterior predictive model-checking (PPMC) method (Rubin, 1984) is one of the popular approach within the Bayesian paradigm, not because of its intuitive appeal and ease of implementation, but more importantly, its strong theoretical basis.

Sinharay (2006a) argued that $S - \chi_k^2$ and $S - G_k^2$ do not have the assumed limiting distribution due to the use of item parameters estimated from ungrouped observations. Sinharay (2006a) suggested using the PPCM method, working along with Markov chain Monte Carlo (MCMC) sampling technique, to sample the empirical distributions for $S - \chi_k^2$ and $S - G_k^2$ that approximate their actual posterior distributions.

The idea behind the PPMC is to compare the observed data $\boldsymbol{x}$ with *replicated data* $\boldsymbol{x}^{rep}$ generated from the *posterior predictive distribution*

$$p\left(\boldsymbol{x}^{rep}|\boldsymbol{x}\right) = \int p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}. \tag{13.6}$$

$\boldsymbol{\theta}$ contains $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$, or hyper-parameters according to the prior assumption; $p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}\right)$ is the joint likelihood function and $p(\boldsymbol{\theta}|\boldsymbol{x})$ is the posterior distribution given the observed data.

In practice *test quantities/discrepancy measures*, $D(\boldsymbol{x}\boldsymbol{\theta})$, are defined (Gelman, Meng, & Stern, 1996) to evaluate the adequacy of a model; the lack-of-fit can be summarized by the *posterior predictive p-value* (ppp)

$$ppp = \int_{\boldsymbol{\theta}} \int_{\boldsymbol{x}^{rep}} I_{[D(\boldsymbol{x},\boldsymbol{\theta}) \leq D(\boldsymbol{x}^{rep},\boldsymbol{\theta})]} p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{x}^{rep} d\boldsymbol{\theta}, \tag{13.7}$$

where $I[\cdot]$ is the indicator function. The analytical difficulty in (13.6) and (13.7) can be reduced by numerically carrying these out along with the MCMC steps. Model parameters $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, ..., $\boldsymbol{\theta}^{(M)}$ are simulated from the (approximate) posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ within the converged MCMC algorithm. The replicated data, $\boldsymbol{x}^{rep(m)}$, is generated from the likelihood $p\left(\boldsymbol{x}^{rep}|\boldsymbol{\theta}^{(m)}\right)$ for $m = 1, \ldots, M$. This process leads to M draws from the joint distribution $p(\boldsymbol{x}^{rep}, \boldsymbol{\theta}|\boldsymbol{x})$, which can then be used to approximate the ppp by calculating the proportion of times the replicated data has a larger discrepancy than the observed data to approximate the integral in (13.7).

The choice of $D(\boldsymbol{x}, \boldsymbol{\theta})$ is vital but also flexible for the PPMC method. Sinharay and Almond (2007) suggested examining the item-fit by Q_1. Wang et al. (2015) employed the power-divergence (PD; a more general statistic family including Q_1) and Stone-type PD to check item-level fit. Sinharay and Almond (2007) assessed the overall fit by looking at the residual between individual raw score and expected score. GOF statistics and RMSEA mentioned above could be chosen as the discrepancy measure for detecting overall misfit.

Robins, van der Vaart, and Ventura (2000) show that the ppp tends to be conservative for some choices of discrepancy measure. Similar issues were found in Wang et al. (2015) that the ppp is more conservative than its classic GOF counterparts. However, as many researchers argued, a conservative diagnostic with reasonable power is better than tests with unknown properties or poor Type I error rates.

Other posterior predictive based methods, such as the direct display (for overall fit) and the odds ratios (for item association/pairs fit), are not covered in this chapter. We refer readers to Sinharay (2006b) for more details about these methods which have been used in model diagnostics for Bayesian networks.

13.5 Empirical Illustration

A real dataset for the 28-item Examination for the Certificate of Proficiency in English (ECPE) is analyzed in this section as an example. ECPE is developed and scored by the English Language Institute of the University of Michigan. The data has been used to investigate multidimensional cognitive attributes (Buck & Tatsuoka, 1998; Templin & Hoffman, 2013) and to examine attribute hierarchy (Templin & Bradshaw, 2014).

Current discussions on the attribute hierarchy is worthy of mention. Von Davier and Haberman (2014) point out that the hierarchical diagnostic classification models (HDCMs; Templin & Bradshaw, 2014) are equivalent to an ordered latent class model. Also, Templin and Hoffman (2013) argue that HDCM and G-DINA models perform not substantially better than the unidimensional two-parameter IRT model. Von Davier and Haberman (2014) suggest to start with the simplest possible model rather than with a potentially overly complex model.

In this illustrative example, the hierarchy among attributes is not be considered for the model assumption. But we still focus on the CDM framework. Global fit for several common CDMs are compared by information criteria and the absolute overall fit is also examined. Item-level fit is checked when the DINA framework is assumed to fit the data.

In specific three attributes in ECPE dataset (Buck & Tatsuoka, 1998) are measured: morphosyntactic rules, cohesive rules, and lexical rules. The data with a sample of 2,922 test-takers and its Q-matrix have been included in R packages such as G-DINA (Ma, de la Torre, & Sorrel, 2018) and CDM (Robitzsch, Kiefer, George, & Ünlü, 2018).

13.5.1 Global Model Fit Results

13.5.1.1 Relative Fit

Table 13.1 presents the performance of AIC, BIC and sample-size adjusted BIC across the saturated G-DINA, the Additive-CDM (ACDM) and a mixed form (MIX) of G-DINA and ACDM. ACDM only contains terms in (13.1) up-to main effects. For the mixed form, items 3, 11, 12, 17 and 21 are set as the ACDM since their estimated second-order interaction coefficients are not significantly different from

Table 13.1 Global relative fit indices

	p	AIC	BIC	sBIC
DINA	63	85,813.98	86,190.72	86,191.24
G-DINA	81	85,642.67	86,127.05	86,127.71
NC-GDINA	81	85,639.19	86,123.57	86,124.24
ACDM	72	85,639.01	86,069.57	86,070.16
MIX	76	85,642.17	86,096.65	86,097.27

Table 13.2 Global absolute fit indices

	M_2	df	RMSEA_2	$\max(\chi^2_{kk'})$
ACDM	474.557 (0.000)	325	0.013 (0.010, 0.015)	38.712 (0.000)
MIX	500.841 (0.000)	330	0.013 (0.010, 0.016)	39.639 (0.000)
DINA	515.707 (0.000)	343	0.013 (0.011, 0.015)	26.608 (0.000)

0 under the G-DINA model. Non-constrained G-DINA (NC-GDINA) denotes the saturated G-DINA without monotonicity constraints.

The information criterion in Table 13.1 picks out the ACDM. It also shows that G-DINA and NC-GDINA are different models, which should be noted when choosing a model. Notice that the NC-GDINA model is probably not identified in this case. The general discussions on the identification issue related to monotonicity constraints are considered in von Davier (2014). Here we just use this model to emphasize that the monotonicity constraints should not be ignored in model fitting and selection. The standard AIC and BIC statistics examined here do not capture the identification issue for the unconstrained model, and so are must be interpreted when comparing the constrained G-DINA model to the unconstrained NC-GDINA model.

13.5.1.2 Absolute Fit

Following the relative fit, Table 13.2 provides the absolute fit of ACDM, MIX, and DINA. The statistics M_2 and RMSEA_2 are limited-information based statistics as mentioned previously. The p-values for the test statistics and the 95% confidence intervals for the RMSEA are given in parentheses following the various statistics. The final column, $\max(\chi^2_{kk'})$, is the largest $\chi^2_{kk'}$ among all pairs of items; the p-value for the statistic is obtained by the Holm-Bonferroni procedure.

Both limited-information and item-pairwise test statistics suggest that none of the three models provide adequate fit to the data. A possible reason is the misspecification (underspecification) of the Q-matrix, which would lead to local dependence among the items. In contrast the RMSEA suggests that the models produce an adequate and similar fit to the data. The different results provided by RMSEA and other absolute fit support what we mention in the review section. Absolute fit statistics such as limited-information M_2 tend to reject the null model when sample size is large, whereas RMSEA takes the effect of sample size into consideration.

13.5.2 Item-Level Fit Results

13.5.2.1 Relative Fit

Table 13.3 lists the chi-square statistics based on the Wald test. The Wald test, given in the first column of the table, examines the null hypothesis that the item is DINA against its alternative that is the G-DINA. The second column is for the ACDM case.

The table lists the items rejected under the DINA null. Among them, items 3, 7, and 21 are not rejected under the ACDM null. The df is 2 for the DINA null and 1 for the ACDM null since there are only 2 attributes required by these items.

13.5.2.2 Absolute Fit

The Wald test is useful to select the parametric form for each item but it relies heavily on the nested model structure. The most saturated case is G-DINA within this framework. Absolute fit methods offer a way to check if items fit the data when the G-DINA is assumed.

Table 13.4 shows the absolute fit results. RMSEA_k (Oliveri & von Davier, 2011) is the item-level RMSEA based on RMSEA_1. $S-\chi^2$ is the raw-score based Pearson's chi-square statistic from Orlando and Thissen (2000). $S-RR-\chi^2$ and $S-DN-\chi^2$ are Rao-Robson (RR) and Dzhaparidze-Nikulin (DN) adjusted versions for $S-\chi^2$ described in more detail below.

Chernoff and Lehmann (1954) have shown that a χ^2 statistic computed from the cells of probabilities (e.g., e_{ks} in $S-\chi^2$) based on grouped individual observations while its estimates (e.g., item parameters $\hat{\boldsymbol{\delta}}_k$) are from ungrouped observations does not have the expected limiting distribution.

To address the issue, Rao and Robson (1975) modified the squared-residual based statistics, in the item-level case $\boldsymbol{v}_k = (v_{k,1}, v_{k,2}, \ldots, v_{k,K-1})^T$, as

$$RR-\chi^2 = \boldsymbol{v}_k^T \left(\boldsymbol{I}_{K-1} - \boldsymbol{B}\boldsymbol{J}^{-1}\boldsymbol{B}^T\right)^{-1} \boldsymbol{v}_k,$$

Table 13.3 Item-level relative fit indices

	DINA χ^2_{Wald}	ACDM χ^2_{Wald}
Item 1	39.823 (0.000)	26.342 (0.000)
Item 3	23.871 (0.000)	0.102 (0.750)
Item 7	213.444 (0.000)	36.029 (0.000)
Item 11	98.963 (0.000)	1.173 (0.279)
Item 12	201.990 (0.000)	201.607 (0.000)
Item 16	106.427 (0.000)	5.966 (0.015)
Item 17	27.508 (0.000)	4.194 (0.041)
Item 20	76.782 (0.000)	37.586 (0.000)
Item 21	130.965 (0.000)	2.399 (0.121)

Table 13.4 Item-level absolute fit indices

	RMSEA	$S-\chi^2$	$S-RR-\chi^2$	$S-DN-\chi^2$
Item 2	0.012	46.723 (0.000)	46.727 (0.000)	39.465 (0.002)
Item 10	0.032	54.763 (0.000)	54.791 (0.000)	29.236 (0.032)
Item 15	0.026	49.838 (0.000)	49.854 (0.000)	33.857 (0.009)
Item 19	0.033	51.656 (0.000)	51.689 (0.000)	28.647 (0.038)
Item 22	0.042	61.712 (0.000)	61.754 (0.000)	27.957 (0.045)
Item 23	0.016	59.212 (0.000)	59.225 (0.000)	38.331 (0.002)
Item 24	0.029	75.482 (0.000)	75.521 (0.000)	45.462 (0.000)

where

$$v_{k,s}(\hat{\boldsymbol{\delta}}_k) = \frac{\sqrt{N_k}(o_{ks} - e_{ks}(\hat{\boldsymbol{\delta}}_k))}{\sqrt{e_{ks}(\hat{\boldsymbol{\delta}}_k)(1 - e_{ks}(\hat{\boldsymbol{\delta}}_k))}};$$

$\boldsymbol{J}$ is the information matrix w.r.t the kth item parameters $\hat{\boldsymbol{\delta}}_k$ and $\boldsymbol{B}$ is the Jacobian matrix of $\boldsymbol{e}_k = (e_{k,1}, \ldots, e_{k,K-1})^T$ w.r.t $\hat{\boldsymbol{\delta}}_k$. The statistic is essentially

$$\boldsymbol{v}_k^T Cov\,(\boldsymbol{v}_k)^{-1}\,\boldsymbol{v}_k$$

which follows χ^2_{K-1} instead of $\chi^2_{K-1-p_k}$. Dzaparidze and Nikulin (1975) proposed a similar statistic

$$DN - \chi^2 = \boldsymbol{v}_k^T \left(\boldsymbol{I}_{K-1} - \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\right)^{-1} \boldsymbol{v}_k,$$

which follows $\chi^2_{K-1-p_k}$. The connection between the statistics has been discussed by McCulloch (1985). Simply put, the idea is to approximate the actual covariance matrix for residual $\boldsymbol{v}_k$ based on the MLEs calculated from ungrouped data.

Table 13.4 presents the significant items across the three statistics under the saturated G-DINA model. The dfs are $20-1-2=17$, $20-1=19$, and $20-1-2=17$ for each column respectively. Notice that $K=20$ since we merge the cells with observed counts less than 20. For the item-level fit detection, parameters for the other items and the size of latent classes are assumed to be invariant; plus, all flagged items are DINA items. Therefore, $p_k=2$. The result suggests that a more flexible parametric form or a more sophisticated Q-matrix is needed to account for the misfit.

13.6 Discussion

While this chapter has attempted to review some of the most common measures and methods for evaluating model fit, it is by no means complete. New methods are appearing quite regularly for the evaluation of fit.

An example is the study by Chalmers and Ng (2017), who modify the square-residual based statistics by using plausible value imputation to generate and account for the uncertainty coming from the use of latent trait estimates. The idea is very similar to the resampling-based and the PPMC methods.

Also, for item-level absolute fit methods within the Bayesian approach, residual-based display techniques are not covered in this chapter. Display methods in model diagnosis for Bayesian networks (Sinharay, 2006b) can be employed to examine item fit in CDM. Developing more intuitive display methods to visualize and measure the item-level misfit for CDM is a potential topic.

Further generalizations are needed as well. Current studies focus on CDMs with binary skills and binary item scores. Extending current methods to address the polytomous variants is certainly an area need more research, as is evaluating those methods. More comprehensive simulation and empirical studies on comparable methods could offer researchers further guidance and insights.

It is necessary to consider and assess the practical significance and consequence of model misfit since no model is perfect. The issue has been stressed in the context of IRT framework by Hambleton and Han (2005) and Sinharay and Haberman (2014). Whereas Sinharay and Haberman (2014) focus on the item misfit significance for high stakes tests, van Rijn, Sinharay, Haberman, and Johnson (2016) investigate low-stakes assessment. Two studies both find that the misfit hardly impacts the test outcome. To the best of the authors' knowledge, the methodologies or guidance for this topic have not been comprehensively studied in CDMs, which implies a promising research direction.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactionson Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing*. https://doi.org/10.1177/026553229801500201

Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement, 41*(5), 372–387. Retrieved from https://doi.org/10.1177/0146621617692079

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Chernoff, H., & Lehmann, E. L. (1954). The use of maximum likelihood estimates in $\chi 2$ tests for goodness of fit. *The Annals of Mathematical Statistics, 25*(3), 579–586. https://doi.org/10.1214/aoms/1177728726

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618. Retrieved from https://doi.org/10.1177/0146621613488436

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429–449. Retrieved from https://doi.org/10.1111/j.1745-3984.2009.00091.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Chiu, C.-Y. (2016, June). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273. Retrieved from https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Douglas, J. A. (2008, Mar). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595. Retrieved from https://doi.org/10.1007/s11336-008-9063-2

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355–373. Retrieved from https://doi.org/10.1111/jedm.12022

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. Retrieved from https://doi.org/10.1177/0748175615569110

Dzaparidze, K. O., & Nikulin, M. S. (1975). On a modification of the standard statistics of pearson. *Theory of Probability & Its Applications, 19*(4), 851–853. Retrieved from https://doi.org/10.1137/1119098

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model finess via realized discrepancies. *Statistica Sinica, 6*, 733–807. Retrieved from http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.142.9951

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2). Retrieved from http://www.jstatsoft.org/v74/i02/, https://doi.org/10.18637/jss.v074.i02

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*(426), 645–656. Retrieved from http://www.jstor.org/stable/2290867.

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington, DC: Degnon Associates.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology, 69*(3), 225–252. Retrieved from https://doi.org/10.1111/bmsp.12074

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*(1), 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Lei, P.-W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and q-matrices. *Applied Psychological Measurement, 40*(6), 405–417. Retrieved from https://doi.org/10.1177/0146621616647954

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*(8), 579–598. Retrieved from https://doi.org/10.1177/0146621609331960

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*(1), 3–26. Retrieved from https://doi.org/10.3102/1076998615621293

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "Equatings". *Applied Psychological Measurement, 8*(4), 453–461. https://doi.org/10.1177/014662168400800409

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200–217. https://doi.org/10.1177/0146621615621717

Ma, W., de la Torre, J., & Sorrel, M. A. (2018). *GDINA: The generalized DINA model framework.* Retrieved from http://cran.r-project.org/package=GDINA

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2n contingency tables. *Journal of the American Statistical Association, 100*(471), 1009–1020. Retrieved from http://pubs.amstat.org/doi/abs/10.1198/016214504000002069

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

McCulloch, C. E. (1985). Relationships among some chi-square goodness of fit statistics. *Communications in Statistics – Theory and Methods, 14*(3), 593–603. Retrieved from https://doi.org/10.1080/03610928508828936

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315–333. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. Retrieved from https://doi.org/10.1177/01466216000241003

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika, 83*(2), 251–266. Retrieved from http://biomet.oxfordjournals.org/content/83/2/251.abstract, https://doi.org/10.1093/biomet/83.2.251

Rao, K. C., & Robson, D. S. (1975). A chi-square statistic for goodness-of-fit tests within the exponential family. *Communications in Statistics, 3*, 1139–1153. https://doi.org/10.1080/03610927408827216

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2000.10474310

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2018). CDM: Cognitive diagnosis modeling. R package version 7.1–20. https://cran.r-project.org/package=CDM

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151–1172. http://projecteuclid.org/euclid.aos/1176346785, https://doi.org/10.1214/aos/1176346785

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. Retrieved from http://projecteuclid.org/euclid.aos/1176344136, https://doi.org/10.1214/aos/1176344136

Sinharay, S. (2006a). Bayesian item fit analysis for unidimensional item response theory models. *The British Journal of Mathematical and Statistical Psychology, 59*(Pt 2), 429–

49. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17067420, https://doi.org/10.1348/000711005X66888

Sinharay, S. (2006b). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986031001001

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic Models: A case study. *Educational and Psychological Measurement, 67*(2), 239–257. Retrieved from http://journals.sagepub.com/doi/10.1177/0013164406292025, https://doi.org/10.1177/0013164406292025

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12024

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*(8), 614–631. Retrieved from https://doi.org/10.1177/0146621617707510

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology, 64*(4), 583–616. https://doi.org/10.1111/1467-9868.00353

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit statistic in IRT models. *Journal of Educational Measurement*, 37, 58–75.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*. https://doi.org/10.1007/s11336-013-9362-0

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12010

van Rijn, P. W., Sinharay, S., Haberman, S. J., & Johnson, M. S. (2016). Assessment of fit of item response theory models used in large-scale educational survey assessments. *Large-Scale Assessments in Education, 4*(1), 10. Retrieved from https://doi.org/10.1186/s40536-016-0025-3

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*. https://doi.org/10.1002/ets2.12043

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'Diagnostic' classification Models-A commentary. *Psychometrika*. https://doi.org/10.1007/s11336-013-9363-z

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA model. *Applied Psychological Measurement, 39*(7), 525–538. Retrieved from https://doi.org/10.1177/0146621615583050

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Chapter 14
Exploratory Data Analysis for Cognitive Diagnosis: Stochastic Co-blockmodel and Spectral Co-clustering

Yunxiao Chen and Xiaoou Li

Abstract Exploratory data analysis (EDA) is an essential stage in statistical analysis that extracts information from data to assist confirmatory statistical modeling. Diagnostic classification models (DCMs) are a confirmatory approach to cognitive diagnosis, for which EDA tools need to be developed to assist the design of DCM-based tests. In this chapter, we propose a stochastic co-blockmodel that approximates the structure of many DCMs and an efficient spectral co-clustering algorithm for fitting the model. The proposed approach explores the structure of assessment data by clustering students and items into latent classes and analyzing the relationship between the student classes and the item classes. The performance of the proposed algorithms is evaluated through simulation studies. A real data example is provided to illustrate the use of the proposed method.

14.1 Introduction

In educational testing research, diagnostic classification models (DCMs) for cognitive diagnosis have been developed to classify students' mastery or nonmastery of multiple skill variables (e.g., Rupp & Templin (2008); Rupp et al. (2010); von Davier (2008)). The DCMs are restricted latent class models (Haertel, 1989), which assume each student comes from a latent class defined by the profile of mastery or nonmastery of multiple attributes. Due to the confirmatory nature of DCMs, the design of a DCM-based test depends on domain knowledge and test development can be labor-intensive, requiring the specification of a set of attributes, and a Q-matrix that provides a formal description of the item-attribute relationship, and some

Y. Chen (✉)
London School of Economics and Political Science, London, UK
e-mail: y.chen186@lse.ac.uk

X. Li
School of Statistics, University of Minnesota, Minneapolis, MN, USA
e-mail: lixx1766@umn.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_14

rationale whether skills are functioning in a compensatory, conjunctive, or some other way.

Exploratory data analysis (EDA) is a set of quantitative approaches for data analysis that extracts information from data beyond confirmatory statistical modeling. It has become an essential stage in statistical data analysis. In psychometrics, the idea of EDA has been fully implemented through the extensive use of exploratory factor analysis, principal component analysis, and cluster analysis. In cognitive diagnosis modeling, not many exploratory data analysis approaches are available for simultaneous exploration of data structure among both students and items, except for data-driven approaches for learning the Q-matrix in cognitive diagnosis modeling (Chen, Liu, Xu, & Ying, 2015; Liu, Xu, & Ying, 2012, 2013) and regularized latent class analysis (Chen, Li, Liu, & Ying, 2017).

In this chapter, we propose a *stochastic co-blockmodel* as an EDA model for cognitive diagnosis item response data. This model is closely related to the stochastic blockmodel (Holland, Laskey, & Leinhardt, 1983), which is widely used for analyzing network data. The proposed model imposes latent class structure among both students and items to capture the key features of cognitive diagnosis modeling. Specifically, a student latent class may represent a group of students who share the same cognitive attribute profile and an item latent class may represent a set of items which measure the same set of attributes. As will be further explained, this model can be viewed as an approximation to many DCMs.

We develop an efficient spectral co-clustering algorithm for fitting the proposed model. Using this algorithm, the latent class memberships of both the students and items are estimated, resulting in non-overlapping homogeneous groups of students and items. Such a task is typically known as co-clustering or bi-clustering (Choi & Wolfe, 2014; Dhillon, 2001; Hartigan, 1972). This algorithm can be viewed as an extension of the spectral bi-clustering algorithm proposed in Dhillon (2001) for co-cluster analysis of documents and words. The proposed algorithm is also closely related to the spectral clustering method in Chen et al. (2017) for unsupervised item classification. The proposed model, together with the algorithm, can be a useful and easy-to-implement EDA tool for examining a pool of items before going through the potentially costly work of developing a confirmatory DCM, in particular, when having rich data with many students and items. It may provide the researchers a better understanding of the item pool and the student population, which further facilitates the development of a DCM-empowered cognitive diagnosis test.

The rest of this chapter is organized as follows. The stochastic co-blockmodel is proposed in Sect. 14.2 and an efficient spectral co-clustering algorithm is described in Sect. 14.3. Section 14.4 reports results of simulation studies designed to evaluate the proposed methods, followed by a real data example in Sect. 14.5. We conclude with discussions in Sect. 14.6.

14.2 Stochastic Co-blockmodel

14.2.1 Proposed Model

Consider N students answering K assessment items. Let $X_{nk} \in \{0, 1\}$ be a random variable, denoting student n's response to item k, where 0 and 1 code the incorrect and correct answers, respectively. In addition, let x_{nk} be the realization of X_{nk}. For ease of exposition, we further denote $\mathbf{X} = (X_{nk})_{N\times K}$ as the response matrix and $\mathbf{x} = (x_{nk})_{N\times K}$ be its realization. The proposed stochastic co-blockmodel assumes that there exist S latent classes among the students and T latent classes among items. We use $A_n \in \{1, 2, \ldots, S\}$ and $\Lambda_k \in \{1, 2, \ldots, T\}$ to denote the latent class memberships of students and items, respectively. Our model assumes that students and items from the same latent class are stochastically equivalent, in the sense that

$$P(X_{nk} = 1|A_n = s, \Lambda_k = t) = b_{st}, \tag{14.1}$$

that is, the distribution of X_{nk} only depends on the latent class memberships of student n and item k. We further use the matrix $\mathbf{B} = (b_{st})_{S\times T}$ to denote all such item response probabilities, characterizing the performance of each student latent class on each item latent class. We call this matrix the *relation matrix*. Moreover, $P(A_n = s) = p_s,\ s = 1, \ldots, S$ and $P(\Lambda_k = t) = \pi_t, t = 1, \ldots, T$. Given S and T, the model parameters include $\mathbf{B}$, $\mathbf{p} = (p_1, \ldots, p_S)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$. In practice, S and T are unknown and also estimated from data. This model is closely related to Holland et al. (1983)'s stochastic blockmodel for undirected networks and Rohe, Qin, and Yu (2016)'s stochastic co-blockmodel for directed networks, both of which are widely used statistical models for network data analysis.

The proposed model describes homogeneous groups of students and items using latent classes. By doing so, the dimensionality of the data is substantially reduced from an N by K data matrix to an S by T relation matrix. In the context of cognitive diagnosis, a student latent class may represent a group of students with the same proficiency levels on a set of skills being measured, and an item latent class may represent a set of equally difficult items that measure the same set of skills. It simplifies reality by assuming that the distribution of X_{nk} only depends on the latent class memberships of student n and item k, but not any other student or item specific information. In other words, under this model, two students/items within the same latent class are not distinguishable based on the item response data. Although possibly over-simplified, the proposed model provides a statistical framework for identifying homogeneous student and item groups from item response data and for analyzing the inter-group relationship. Results from fitting this model may provide education researchers insights into the latent structures of students and items, facilitating the design of the measurement and learning of cognitive abilities.

Like many other latent class models, the proposed model is also invariant under "label swapping", that is both the student and item latent classes can be freely relabeled without affecting the distribution of response data. Consequently,

parameter identifiability can only be established up to label swapping. Allman, Matias, and Rhodes (2009) have established the "generic identifiability" for a wide class of latent class models, which says that up to label swapping the set of nonidentifiable parameters has Lebesgue measure zero in the parameter space. Following Allman et al. (2009)'s method, it is not difficult to establish the generic identifiability of the proposed model.

14.2.2 Connection with DCMs

We explain the connection between the proposed model and DCMs, using the Deterministic Input, Noisy-And gate (DINA) model (Junker & Sijtsma, 2001) as an illustrative example. Consider the setting that D binary attributes are measured. The DINA model assumes that each student n is represented by his/her attribute profile, denoted by $\mathbf{a}_n = (a_{n1}, \ldots, a_{nD})$, where $a_{nd} = 0$ and 1 represent the nonmastery and mastery of the dth attribute, respectively. Moreover, each item k is characterized by a D-dimensional vector $\mathbf{q}_k = (q_{k1}, \ldots, q_{kD})$, where $q_{kd} = 1$ if item k measures attribute d and 0 otherwise. The DINA model assumes that

$$P(X_{nk} = 1|\mathbf{a}_n, \mathbf{q}_k) = \begin{cases} 1 - s_k & \text{if } \mathbf{a}_n \succcurlyeq \mathbf{q}_k, \\ g_k & \text{otherwise,} \end{cases} \tag{14.2}$$

where $\mathbf{a}_n \succcurlyeq \mathbf{q}_k$ denotes $a_{nd} \geq q_{kd}$, for all $d = 1, \ldots, D$. Equation (14.2) implies that if student n has mastered all the attributes measured by the item, the probability of correctly answering is $1 - s_k$ and if at least one necessary attribute has not been mastered, the probability of correctly answering is g_k, where s_k and g_k are known as the slipping and guessing parameters, respectively, which typically take small values (e.g., less than 0.3). The indicator $1_{\{\mathbf{a}_n \succcurlyeq \mathbf{q}_k\}}$ is known as the ideal response, which is the response that student n supposed to provide when both the slipping and guessing parameters are zero. When the slipping and guessing parameters are nonzero, the observed response can be viewed as a perturbation to the ideal response.

In panel (a) of Fig. 14.1, we plot a realization of $\mathbf{X} = (X_{nk})_{N \times K}$ from a DINA model using a heat map, where $N = 400$, $K = 60$, and $D = 2$. For this simulated data set, the students and items are ordered a priori, so that students 1–100, 101–200, 201–300, and 301–400 have attribute profiles (0, 0), (1,0), (0,1), and (1,1), respectively, and items 1–20, 21–40, and 41–60 have q-vectors (1, 0), (0,1), and (1,1), respectively. Moreover, the slipping and guessing parameters are randomly generated from the uniform distribution over interval [0.1, 0.3]. The heat map provides a visualization of the data matrix, where the entries of the matrix are represented as colors. In Fig. 14.1, we use white and black colors for $x_{nk} = 0$ and 1, respectively. A clear block structure is observed, which is due to the combination of the latent classes of students labeled by the attribute profiles and latent classes of items labeled by the q-vectors.

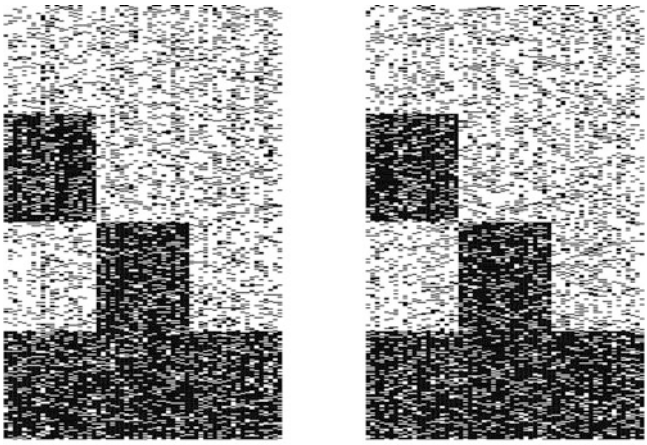


Fig. 14.1 (**a**) Heat map of a realization of response data **X** under a DINA model. (**b**) Heat map of a realization of response data **X** under a stochastic co-blockmodel

This block structure of the DINA model can be approximated by a stochastic co-blockmodel with $S = 4$ student classes and $T = 3$ item classes. In particular, panel (b) of Fig. 14.1 presents a realization from such a stochastic co-blockmodel, where the relation matrix **B** is specified as

$$\mathbf{B} = \begin{pmatrix} 0.2 & 0.2 & 0.2 \\ 0.8 & 0.2 & 0.2 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.8 & 0.8 \end{pmatrix}. \tag{14.3}$$

In fact, by visual inspection, one can hardly find systematic differences in the global structures of the two heat maps in Fig. 14.1. In this example, the stochastic co-blockmodel is an approximation to the generating DINA model, where the four student classes correspond to the attribute profiles $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$, respectively, and the three item classes correspond to the q-vectors $(1, 0)$, $(0, 1)$, and $(1, 1)$, respectively. This stochastic co-blockmodel simplifies the DINA model by assuming that $P(X_{nk} = 1) = P(X_{nk'} = 1)$ as long as k and k' belong to the same item class, while the DINA model allows for item-specific slipping and guessing parameters.

Not specific to the DINA model, the stochastic co-blockmodel can provide good approximations to other DCMs as well, including the Deterministic Input, Noisy-Or gate model (DINO; Templin & Henson, 2006), Noisy-Input Deterministic-And-gate (NIDA) and Noisy-Input Deterministic-Or-gate (NIDO) models (e.g., Rupp et al., 2010), and the Reduced Reparameterized Unified Model (RRUM; e.g., Rupp et al.). For example, the RRUM can be viewed as an extension of the DINA model, with

the following item response function:

$$P(X_{nk} = 1|\mathbf{a}_n, \mathbf{q}_k) = \beta_{k0} \prod_{d=1}^{D} \beta_{kd}^{q_{kd}(1-a_{nd})}, \tag{14.4}$$

where β_{k0} is the correct response probability for subjects who possess all required attributes for item k, and $\beta_{kd} \in (0, 1)$ is the penalty parameter for not possessing the dth attribute. When $\beta_{k0} = 1$ and $\beta_{kd} = 0, d \neq 0$, the item response function (14.4) is consistent with the DINA model with the corresponding slipping and guessing parameters being zero. Therefore, one would expect the proposed stochastic co-blockmodel to approximate the RRUM model well, at least when β_{k0} is close to 1 and the penalty parameters β_{kd} are close to 0.

14.2.3 Limitations

The proposed model has its limitations due to its simple form. First, the proposed model does not approximate all the DCMs, in particular, the general families of DCMs, such as the general diagnostic model (GDM; von Davier, 2008) and Loglinear Cognitive Diagnosis Model (LCDM; Henson, Templin, & Willse, 2009).

Second, results from fitting the proposed model do not directly suggest what diagnostic classification model should be used or suggest the Q-matrix for the items. Instead, analysis based on this model only suggests homogeneous groups of items which may measure similar content areas, homogeneous groups of students sharing similar skills, and the relationship between the student and the item groups. Domain experts may find practical interpretations of the item and student groups based on these results, by examining the contents of the items, possibly assisted by additional information about the items and students from sources other than the item response data. This deeper understanding of the item pool and the student population may further lead to a DCM-empowered cognitive diagnosis test, given additional efforts, such as fitting and comparing different confirmatory DCMs and collecting additional data for model validation.

14.3 Spectral Co-clustering Algorithm

14.3.1 Spectral Co-clustering Algorithm

Stochastic co-blockmodel based analysis aims at learning the latent class memberships of both students and items. A traditional method is an empirical Bayes approach which first fits the model by maximizing the marginal likelihood function of item responses and then infers the latent class memberships of students and items

via a maximum a posteriori probability (MAP) estimator. The marginal likelihood function can be written as

$$L(\mathbf{B}, \mathbf{p}, \boldsymbol{\pi}) = \sum_{s=1}^{S} \sum_{t=1}^{T} \left\{ \prod_{n=1}^{N} \prod_{k=1}^{K} b_{st}^{x_{nk}} (1 - b_{st})^{1 - x_{nk}} p_s \pi_t \right\}, \tag{14.5}$$

which is difficult to optimize using the classical expectation-maximization algorithm (EM; Dempster, Laird, & Rubin, 1977) due to an intractable E-step.

In this chapter, we provide an alternative, the spectral co-clustering algorithm (Algorithm 1), for fitting the model. This algorithm extends the spectral bi-clustering algorithm (Dhillon, 2001) by introducing regularization to the Laplacian matrix, a key quantity of the algorithm. As shown via both empirical studies and theoretical results (Amini, Chen, Bickel, & Levina, 2013; Joseph & Yu, 2016; Qin & Rohe, 2013), regularizing the Laplacian matrix can lead to better performance in cluster analysis. Unlike the empirical Bayes approach, this algorithm does not explicitly optimize an objective function. Instead, this algorithm first embeds students and items into a low dimensional Euclidean space, such that students/items from the same latent class tend to be close to each other after the embedding. As described in Algorithm 1, the dimension of the embedding space is set to be $\min\{S, T\}$, the minimum value of S and T. Then a K-means algorithm is applied to the embedded data for the clustering of both the students and items. Given the classification recorded by $\hat{A}_n$s and $\hat{\Lambda}_k$s, we then estimate the relation matrix $\mathbf{B}$, the population proportions of both students and items, $\mathbf{p} = (p_1, \ldots, p_S)$ and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$, by a moment based method. That is,

$$\begin{aligned} \hat{b}_{st} &= \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} x_{nk} 1_{\{\hat{A}_n = s, \hat{\Lambda}_k = t\}}}{\sum_{n=1}^{N} \sum_{k=1}^{K} 1_{\{\hat{A}_n = s, \hat{\Lambda}_k = t\}}}, \\ \hat{p}_s &= \frac{\sum_{n=1}^{N} 1_{\{\hat{A}_n = s\}}}{N}, \\ \hat{\pi}_t &= \frac{\sum_{k=1}^{K} 1_{\{\hat{\Lambda}_k = t\}}}{K}. \end{aligned} \tag{14.6}$$

The proposed spectral co-clustering algorithm is described as follows and the rationale behind the approach is discussed in Sect. 14.3.2.

Algorithm 1 (Spectral Co-clustering)
Input: *response data matrix* $\mathbf{x} = (x_{nk})_{N \times K}$, *regularization parameter* $\tau \geq 0$, *the number of student clusters* S, *and the number of item clusters* T.

(1) *Compute diagonal matrices*

$$\begin{aligned} \mathbf{D}^{\tau} &= diag(d_{nn})_{N \times N} + \tau \mathbf{I}_{N \times N} \\ \mathbf{O}^{\tau} &= diag(o_{kk})_{K \times K} + \tau \mathbf{I}_{K \times K}, \end{aligned} \tag{14.7}$$

where

$$d_{nn} = \sum_{k=1}^{K} x_{nk},$$
$$o_{kk} = \sum_{n=1}^{N} x_{nk}, \tag{14.8}$$

$diag(d_{nn})_{N\times N}$ and $diag(o_{kk})_{K\times K}$ denote the diagonal matrices with diagonal entries $(d_{11}, \ldots d_{NN})$ and $(o_{11}, \ldots, o_{KK})$, respectively, and $\mathbf{I}_{N\times N}$ and $\mathbf{I}_{K\times K}$ are identity matrices.

(2) *Compute the regularized Laplacian matrix*

$$\mathbf{L}^{\tau} = (\mathbf{D}^{\tau})^{-\frac{1}{2}}\mathbf{x}(\mathbf{O}^{\tau})^{-\frac{1}{2}}. \tag{14.9}$$

(3) *Apply singular value decomposition to the matrix $\mathbf{L}^{\tau}$ and compute the top C left and right singular vectors $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_C) \in \mathbb{R}^{N\times C}$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_C) \in \mathbb{R}^{K\times C}$, where $C = \min\{S, T\}$.*

(4) *Cluster the N rows of $\mathbf{U}$ into S clusters and cluster the K rows of $\mathbf{V}$ into T clusters via a K-means algorithm.*

***Output**: The student latent class memberships $\hat{A}_n \in \{1, 2, \ldots, S\}$ and item latent class memberships $\hat{\Lambda}_k \in \{1, 2, \ldots, T\}$ from step (4).*

A default value of the regularization parameter in Algorithm 1 is chosen as

$$\tau = 2\sum_{n=1}^{N}\sum_{k=1}^{K} x_{nk}/(N + K), \tag{14.10}$$

which follows from a similar spectral clustering algorithm for analyzing undirected network data (Qin & Rohe, 2013). We illustrate the use of Algorithm 1 via its application to the simulated data set from the DINA model in the left panel of Fig. 14.1, for which $S = 4$ and $T = 3$ are assumed known. Consequently, the dimension of the embedding space is $C = \min\{S, T\} = 3$. Figures 14.2 and 14.3 show the embedding of students and items into three-dimensional spaces, where the panels (a), (b), and (c) in both figures correspond to the pairwise scatter plots of column vectors of $\mathbf{U}$ and $\mathbf{V}$, respectively, and the true attribute profiles of students and the true q-vector of items are indicated by different point symbols. As we can observe from these two figures, the true latent classes of the students and the items are well distinguished geometrically under this three dimensional embedding. Consequently, applying the K-means algorithm to $\mathbf{U}$ and $\mathbf{V}$ yields desirable results, that is, only 22 out of 400 students are misclassified and no item is misclassified. Moreover, an estimate of the $\mathbf{B}$-matrix is obtained as given in (14.11), which provides a simple but informative summary of the original data. It implies that the

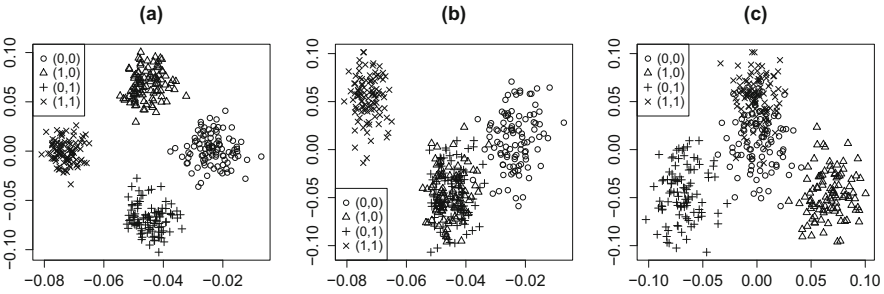


Fig. 14.2 Scatter plots of (**a**) $\mathbf{u}_1$ versus $\mathbf{u}_2$, (**b**) $\mathbf{u}_1$ versus $\mathbf{u}_3$, and (**c**) $\mathbf{u}_2$ versus $\mathbf{u}_3$

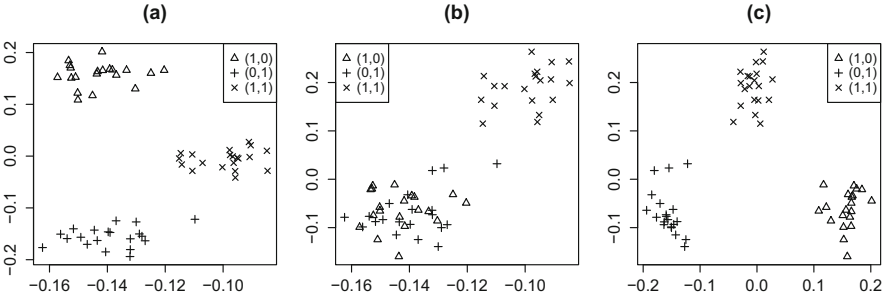


Fig. 14.3 Scatter plots of (**a**) $\mathbf{v}_1$ versus $\mathbf{v}_2$, (**b**) $\mathbf{v}_1$ versus $\mathbf{v}_3$, and (**c**) $\mathbf{v}_2$ versus $\mathbf{v}_3$

first class of students is not good at any type of items, the second and third classes are only good at the first and second types of items, respectively, and the fourth one is good at all types of items.

$$\hat{\mathbf{B}} = \begin{pmatrix} 0.21 & 0.24 & 0.20 \\ 0.79 & 0.20 & 0.18 \\ 0.19 & 0.80 & 0.17 \\ 0.79 & 0.79 & 0.80 \end{pmatrix}. \tag{14.11}$$

14.3.2 Discussion of the Proposed Algorithm

An ideal case We first point out that if data are generated from a DINA model with all the slipping and guessing parameters being zero ($s_k = g_k = 0, k = 1, \ldots, K$), the proposed algorithm can exactly recover the student and item latent classes. Under this situation, only the top $C = \min\{S, T\}$ singular values of L^τ are nonzero, where L^τ is defined in (14.9). Moreover, it is straightforward to show via linear

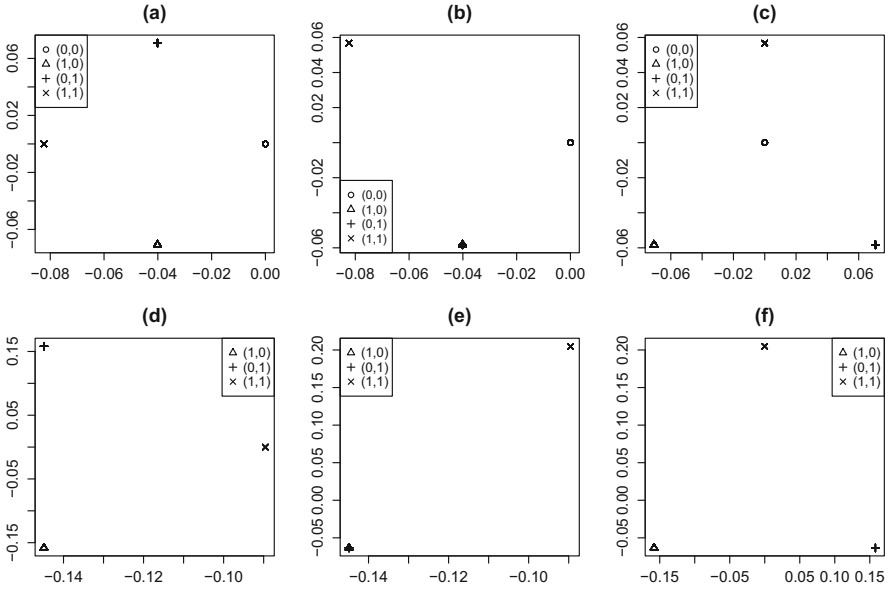


Fig. 14.4 Scatter plots of (**a**) $\mathbf{u}_1$ versus $\mathbf{u}_2$, (**b**) $\mathbf{u}_1$ versus $\mathbf{u}_3$, (**c**) $\mathbf{u}_2$ versus $\mathbf{u}_3$, (**d**) $\mathbf{v}_1$ versus $\mathbf{v}_2$, (**e**) $\mathbf{v}_1$ versus $\mathbf{v}_3$, and (**f**) $\mathbf{v}_2$ versus $\mathbf{v}_3$

algebra that $\mathbf{u}_{n\cdot} = \mathbf{u}_{n'\cdot}$ ($\mathbf{v}_{k\cdot} = \mathbf{v}_{k'\cdot}$) if and only if n and n' have the same attribute profile (k and k' have the same q-vector), where $\mathbf{u}_{n\cdot}$ and $\mathbf{v}_{k\cdot}$ denote the nth row of U and kth row of V, respectively. This is because, under this ideal setting with no measurement error, two students n and n' have the same response pattern and thus the corresponding rows in L^τ are the same, if and only if they share the same attribute profile. This further leads to $\mathbf{u}_{n\cdot} = \mathbf{u}_{n'\cdot}$ if and only if $\mathbf{a}_n = \mathbf{a}_{n'}$, based on the properties of singular value decomposition (e.g., Banerjee & Roy, 2014). For a similar reason, $\mathbf{v}_{k\cdot} = \mathbf{v}_{k'\cdot}$, if and only items k and k' have the same q-vector.

For example, Fig. 14.4 shows the embedding of a data set generated from the DINA model that has the same setting as the one in Sect. 14.2 except that the slipping and guessing parameters are set to be zero, where panels (a)–(c) and panels (d)–(f) display the embedding of students and items, respectively. In these plots, students/items from the same class are mapped to a single point. This ideal situation provides some intuition on the use of the proposed algorithm. Moreover, we point out that such an ideal case can be constructed under other DCMs, such as the DINO, NIDA, NIDO, and RRUM models.

Choosing S and T via cross-validation Algorithm 1 requires the numbers of student classes and item classes as inputs, which are typically not available a priori. We suggest to try different values of S and T for a comprehensive exploration of data and then choose the combination of S and T that best describes the data structure. In addition, a Monte Carlo cross-validation approach is proposed for choosing S and T,

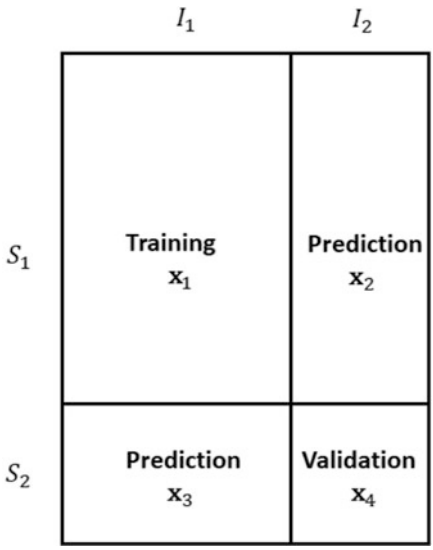


Fig. 14.5 Data splitting in the cross-validation for the selection of S and T

as described in Algorithm 2. We remark that as cluster analysis is an unsupervised learning approach, its cross-validation is more complicated than that for supervised learning (e.g. under a regression setting).

Algorithm 2 (Monte Carlo cross-validation for choosing S and T)
***Input**: Candidate sets of S and T, denoted by $\mathcal{S}$ and $\mathcal{T}$, and the number of Monte Carlo replications M.*
For each Monte Carlo replication m:

1. *As illustrated in Fig. 14.5, in each step of the cross-validation, randomly select 80% of the students and 80% of the items to compose a training data set $\mathbf{x}_1$. Denote the set of selected students as S_1 and the rest as S_2 and denote the set of the selected items as I_1 and the rest as I_2.*
2. *For each $S \in \mathcal{S}$ and $T \in \mathcal{T}$, apply Algorithm 1 to $\mathbf{x}_1$, which gives an estimate of $\mathbf{B}$, $\mathbf{p}$, and $\boldsymbol{\pi}$, denoted by $\hat{\mathbf{B}}$, $\hat{\mathbf{p}}$, and $\hat{\boldsymbol{\pi}}$. We also obtain the estimated latent class memberships of students and items in S_1 and I_1, respectively. Denote them by $\hat{A}_n$, $n \in S_1$ and $\hat{A}_k$, $k \in I_1$.*
3. *Based on $\hat{A}_n$, $n \in S_1$, and the estimated stochastic co-blockmodel, predict the latent class membership of items in I_2 using a Bayesian classifier, based on the prediction data set $\mathbf{x}_2$. More precisely,*

$$\hat{\Lambda}_k = \underset{t\in\{1,\ldots,T\}}{\arg\max} \prod_{n\in S_1} \left(\hat{b}_{\hat{A}_n,t}\right)^{x_{nk}} \left(1-\hat{b}_{\hat{A}_n,t}\right)^{1-x_{nk}} \hat{\pi}_t, \tag{14.12}$$

   for $k \in I_2$.
4. *Similarly, predict the latent class membership of students in S_2 based on the prediction data set $\mathbf{x}_3$, by*

$$\hat{A}_n = \underset{s\in\{1,\ldots,S\}}{\arg\max} \prod_{k\in I_1} \left(\hat{b}_{s,\hat{\Lambda}_k}\right)^{x_{nk}} \left(1-\hat{b}_{s,\hat{\Lambda}_k}\right)^{1-x_{nk}} \hat{p}_s, \tag{14.13}$$

for $n \in S_2$.

5. *We then use the validation set* $\mathbf{x}_4$ *to evaluate the predictions above, by calculating the log-likelihood given the fitted stochastic co-blockmodel and the predicted latent class membership of students and items in* S_2 *and* I_2*, respectively. That is,*

$$l_m(S,T) = \sum_{n\in S_2}\sum_{k\in I_2} x_{nk} \log\left(\hat{b}_{\hat{A}_n,\hat{\Lambda}_k}\right) + (1-x_{nk}) \log\left(1-\hat{b}_{\hat{A}_n,\hat{\Lambda}_k}\right). \tag{14.14}$$

Aggregate the M Monte Carlo replications by

$$\bar{l}(S,T) = \frac{1}{M}\sum_{m=1}^{M} l_m(S,T). \tag{14.15}$$

and its standard error

$$SE(S,T) = \frac{1}{\sqrt{(M-1)M}} \sqrt{\sum_{m=1}^{M} \left(l_m(S,T)-\bar{l}(S,T)\right)^2}. \tag{14.16}$$

We then select S and T using the one standard error rule (e.g., Hastie, Tibshirani, & Friedman, 2009). That is,

$$\hat{S} = \min\left\{S\in\mathcal{S} : \max_{T\in\mathcal{T}} \bar{l}(S,T) \geq \bar{l}(\tilde{S},\tilde{T}) - SE(\tilde{S},\tilde{T})\right\}, \tag{14.17}$$

$$\hat{T} = \min\left\{T\in\mathcal{T} : \max_{S\in\mathcal{S}} \bar{l}(S,T) \geq \bar{l}(\tilde{S},\tilde{T}) - SE(\tilde{S},\tilde{T})\right\}, \tag{14.18}$$

where $(\tilde{S},\tilde{T}) = \arg\max_{S\in\mathcal{S},T\in\mathcal{T}} \bar{l}(S,T)$.

Output*:* $\hat{S}$ *and* $\hat{T}$.

An example is shown in Fig. 14.6, where Algorithm 2 is applied to the simulated data set in Fig. 14.1 generated from a DINA model. The candidate sets $\mathcal{S}$ and $\mathcal{T}$ are both chosen as {2, 3, 4, 5, 6}, and the number of replications M is chosen as 20. The left and right panels of Fig. 14.6 display the functions

$$l_1(S) = \max_{T\in\mathcal{T}} \bar{l}(S,T) \text{ and } l_2(T) = \max_{S\in\mathcal{S}} \bar{l}(S,T), \tag{14.19}$$

and the dashed lines in both panels show the value of $\bar{l}(\tilde{S},\tilde{T}) - SE(\tilde{S},\tilde{T})$. In this example, both S and T are correctly selected. In addition, the one standard error

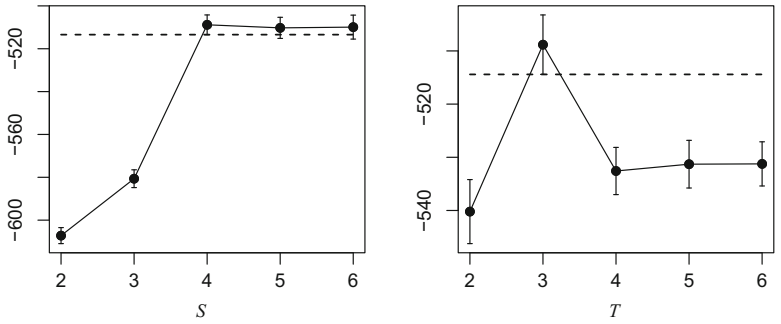


Fig. 14.6 Application of Algorithm 2 to a simulated data set generated from a DINA model. The left and right panels display $l_1(S)$ and $l_2(T)$ as defined in (14.19), respectively. The dashed line indicates the value $\bar{l}(\tilde{S}, \tilde{T}) - SE(\tilde{S}, \tilde{T})$. According to the left panel, $\hat{S} = 4$, because $S = 4$ is the smallest value of S for which $l_1(S) \geq \bar{l}(\tilde{S}, \tilde{T}) - SE(\tilde{S}, \tilde{T})$. Similarly, according to the right panel, $\hat{T} = 3$

rule, which is commonly used for variable selection in regression analysis (chapter 3, Hastie et al., 2009), avoids the over-selection of the number of student latent classes. That is, the values of $l_1(S)$ at $S = 4, 5$, and 6 are approximately the same. The one standard error rule tends to avoid the possible selection of $S = 5$ or 6.

Computational efficiency We remark on the computational efficiency of the proposed algorithms. The computation time of Algorithm 1 is dominated by step (3) of the algorithm, the singular value decomposition of an $N \times K$ matrix. As an extension of eigenvalue decomposition, singular value decomposition is a well developed algorithm for decomposing any $N \times K$ matrix (Golub & van Loan, 2012) into the form $\tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top$, where $\tilde{\mathbf{U}}$ is an $N \times N$ unitary matrix, $\tilde{\mathbf{\Sigma}}$ is a diagonal $N \times K$ matrix with non-negative real numbers on the diagonal, and $\tilde{\mathbf{V}}$ is a $K \times K$ unitary matrix. The singular value decomposition can be efficiently computed even for large values of N and K, with the computational complexity of the order $O(\min\{N^2K, K^3\})$ for the most efficient algorithms (Golub & van Loan, 2012). For Algorithm 2, the computation is mainly due to the application of singular value decomposition C times. The above analysis, including the application of both Algorithms 1 and 2 to the simulated data set from the DINA model, takes less than 1 s, where the Algorithms are implemented in statistical software R (R Core Team, 2013) and run on a personal computer.[1]

[1]The processor of the computer is: Intel (R) Core(TM) i5-5300 CPU @ 2.29 GHz.

14.4 Simulation Studies

14.4.1 Study I: Recovery of Latent Classes

We first evaluate the performance of Algorithm 1 based on simulated data from two diagnostic classification models, the DINA model and the RRUM. Specifically, we assume four attributes, which results in $S = 16$ student classes. The population proportion satisfies $P(\mathbf{a}_n = \mathbf{a}) = 1/16$, for all $\mathbf{a} \in \{0, 1\}^4$. In addition, we consider $T = 10$ item classes that are of equal size, corresponding to all item types that measure one or two attributes. Sample sizes $N = 500$ and 1000 and the number of items $K = 200$ and 400 are considered. Under the DINA model, the slipping and guessing parameters are generated from the uniform distribution $U[0.1, 0.3]$. Under the RRUM, β_{k0}s are generated from the uniform distribution $U[0.85, 0.95]$ and β_{kd}s are generated from the uniform distribution $U[0.2, 0.3]$, for $d = 1, \ldots, D$. For each combination of model, sample size, and number of items, 100 independent data sets are generated to evaluate Algorithm 1. Note that in this study, S and T are assumed to be known.

The evaluation of the clustering results is not straightforward, due to "label swapping". We measure the inconsistency between the clustering results of students $\hat{A}_n$s and the true attribute profiles $\mathbf{a}_n$s by

$$\mathrm{IC}_1 = \frac{2}{N(N-1)} \sum_{n \neq n'} \left(1_{\{\hat{A}_n = \hat{A}_{n'}, \mathbf{a}_n \neq \mathbf{a}_{n'}\}} + 1_{\{\hat{A}_n \neq \hat{A}_{n'}, \mathbf{a}_n = \mathbf{a}_{n'}\}} \right). \tag{14.20}$$

Here, $\sum_{i \neq j} 1_{\{\hat{A}_i = \hat{A}_j, \mathbf{a}_i \neq \mathbf{a}_j\}}$ is the number of student pairs who are classified into the same latent class while having different attribute profiles and $\sum_{i \neq j} 1_{\{\hat{A}_i \neq \hat{A}_j, \mathbf{a}_i = \mathbf{a}_j\}}$ is the number of pairs who are classified into different latent classes while sharing the same attribute profile. Thus, (14.20) is the proportion of inconsistent student pairs. Note that this index is invariant under "label swapping". Similarly, the inconsistency index for items is defined as

$$\mathrm{IC}_2 = \frac{2}{K(K-1)} \sum_{k \neq k'} \left(1_{\{\hat{\Lambda}_k = \hat{\Lambda}_{k'}, \mathbf{q}_k \neq \mathbf{q}_{k'}\}} + 1_{\{\hat{\Lambda}_k \neq \hat{\Lambda}_{k'}, \mathbf{q}_k = \mathbf{q}_{k'}\}} \right). \tag{14.21}$$

Results are shown in Table 14.1. The proportions of inconsistently classified pairs of students and pairs of items are low under all settings, which indicates the good performance of Algorithm 1. In addition, IC_1 and IC_2 decrease when the sample size N or the number of items K increases.

Table 14.1 Results from Study I: Mean values of IC_1 and IC_2 over 100 independent replications, under different combinations of models, sample sizes, and numbers of items

		$N = 500$		$N = 1000$	
		$K = 200$	$K = 400$	$K = 200$	$K = 400$
DINA	IC_1	1.1×10^{-2}	1.3×10^{-3}	6.6×10^{-3}	7.6×10^{-4}
	IC_2	1.2×10^{-3}	4.2×10^{-4}	0	0
RRUM	IC_1	1.6×10^{-2}	3.8×10^{-3}	8.9×10^{-3}	1.8×10^{-3}
	IC_2	4.9×10^{-4}	6.6×10^{-5}	2.6×10^{-5}	0

Table 14.2 Results from Study II: number of times that each candidate in $\mathcal{S}$ and $\mathcal{T}$ is chosen

		DINA				RRUM			
		$N = 500$		$N = 1000$		$N = 500$		$N = 1000$	
		$K = 200$	$K = 400$	$K = 200$	$K = 400$	$K = 200$	$K = 400$	$K = 200$	$K = 400$
$\hat{S}$	12	0	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0
	16	91	100	100	100	50	93	83	100
	18	9	0	0	0	50	7	17	0
	20	0	0	0	0	0	0	0	0
$\hat{T}$	6	0	0	0	0	0	0	0	0
	8	0	0	0	0	0	0	0	0
	10	100	100	100	100	100	100	100	100
	12	0	0	0	0	0	0	0	0
	14	0	0	0	0	0	0	0	0

14.4.2 Study II: Selection of Numbers of Latent Classes

This study investigates the performance of Algorithm 2 under the same simulation settings as in Study I. We consider candidate sets $\mathcal{S} = \{12, 14, 16, 18, 20\}$ and $\mathcal{T} = \{6, 8, 10, 12, 14\}$. The number of cross-validation replications M is set to be 50. Results are shown in Table 14.2, which shows the number of times that each candidate in $\mathcal{S}$ and $\mathcal{T}$ is chosen. In particular, the number of item classes is perfectly recovered under all settings, which may be due to the large amount of information available for the item classes. Except under the RRUM when $K = 200$ and $N = 500$, $\hat{S}$ recovers the number of students classes with reasonable accuracy. This accuracy improves when either the sample size N or the number of items K increases. Finally, the recovery of the number of student classes is less accurate under the RRUM than that under the DINA model, which may be due to the more complex model structure of the RRUM.

14.5 Real Data Example

In this section, we illustrate the use of the proposed method via a real data example based on the Examination for the Certificate of Proficiency in English (ECPE), a test developed and scored by the English Language Institute of the University of Michigan. The test measures advanced English skills in examinees whose primary language is not English and is administered internationally once a year. The ECPE data considered in this study contain 2922 examinees' responses to 28 items from a single year's administration. Initially, the ECPE was scored with unidimensional item response theory models and retrofitted with cognitive diagnostic models in Templin and Hoffman (2013) and Templin and Bradshaw (2014).

We apply the proposed method to this data set to investigate its latent structure. We first apply Algorithm 2, with inputs $M = 50$, and candidate sets $\mathcal{S} = \mathcal{T} = \{2, 3, 4, 5\}$. This analysis suggests that $\hat{S} = \hat{T} = 2$. We then apply Algorithm 1 to the data set with $S = T = 2$, which leads to estimated parameters $\hat{\mathbf{p}} = (0.5, 0.5)$, $\hat{\boldsymbol{\pi}} = (0.36, 0.64)$, and

$$\hat{\mathbf{B}} = \begin{pmatrix} 0.43 & 0.73 \\ 0.79 & 0.82 \end{pmatrix}. \tag{14.22}$$

According to $\hat{\mathbf{p}}$, the two student latent classes are of equal sizes. Suggested by $\hat{\mathbf{B}}$, students in the second class tend to have better English proficiency than those in the first class. This result suggests the underlying unidimensionality of the data. Based on $\hat{\boldsymbol{\pi}}$, the first item latent class is substantially smaller than the second class. In addition, according to $\hat{\mathbf{B}}$, items in the first latent class tend to be more difficult than those in the second class. The second type of items may not distinguish the two student groups well, with the corresponding success probabilities 0.73 versus 0.82.

We further explore the structure of data under different combinations of S and T. The results are listed below. When $T = 4$, the smallest item class has only one item and thus the corresponding results are not presented.

(1) $S = 2, T = 3$:

$$\hat{\mathbf{p}} = (0.5, 0.5), \quad \hat{\boldsymbol{\pi}} = (0.14, 0.46, 0.39), \quad \text{and } \hat{\mathbf{B}} = \begin{pmatrix} 0.25 & 0.57 & 0.81 \\ 0.74 & 0.78 & 0.86 \end{pmatrix}.$$

(2) $S = 3, T = 2$:

$$\hat{\mathbf{p}} = (0.30, 0.38, 0.32), \quad \hat{\boldsymbol{\pi}} = (0.36, 0.64), \quad \text{and } \hat{\mathbf{B}} = \begin{pmatrix} 0.36 & 0.72 \\ 0.59 & 0.76 \\ 0.86 & 0.84 \end{pmatrix}.$$

(3) $S = 3, T = 3$:

$$\hat{\mathbf{p}} = (0.33, 0.31, 0.36), \ \hat{\boldsymbol{\pi}} = (0.14, 0.11, 0.75), \ \text{and} \ \hat{\mathbf{B}} = \begin{pmatrix} 0.22 & 0.66 & 0.68 \\ 0.43 & 0.35 & 0.77 \\ 0.79 & 0.82 & 0.85 \end{pmatrix}.$$

(4) $S = 4, T = 2$:

$$\hat{\mathbf{p}} = (0.19, 0.29, 0.28, 0.24), \ \hat{\boldsymbol{\pi}} = (0.36, 0.64), \ \text{and} \ \hat{\mathbf{B}} = \begin{pmatrix} 0.32 & 0.72 \\ 0.49 & 0.73 \\ 0.69 & 0.80 \\ 0.88 & 0.84 \end{pmatrix}.$$

(5) $S = 4, T = 3$:

$$\hat{\mathbf{p}} = (0.25, 0.19, 0.23, 0.33), \ \hat{\boldsymbol{\pi}} = (0.14, 0.11, 0.75), \ \text{and} \ \hat{\mathbf{B}} = \begin{pmatrix} 0.20 & 0.44 & 0.71 \\ 0.28 & 0.80 & 0.68 \\ 0.55 & 0.36 & 0.80 \\ 0.80 & 0.83 & 0.85 \end{pmatrix}.$$

For the cases (1) $S = 2, T = 3$, (2) $S = 3, T = 2$, and (4) $S = 4, T = 2$, the corresponding student latent classes can be ordered from the least to the most proficient, according to the estimated relation matrix $\hat{\mathbf{B}}$. These results echo the finding under $S = T = 2$, suggesting that the data may be essentially unidimensional.

Some weak evidence on the multidimensionality of data is found from the results from the settings (3) $S = 3, T = 3$ and (5) $S = 4, T = 3$. Under setting (3) $S = 3$ and $T = 3$, the first two student classes cannot be ordered. Specifically, the first student class has better performance on the second type of items, while the second class is better at both the first and third types of items. The third student class may contain the most proficient students with dominantly better performance on all three types of items. The results under setting (5) $S = 4$ and $T = 3$ are similar to those from setting (3). According to the estimated relation matrix $\hat{\mathbf{B}}$, students in the first class perform poorly on all types of items. The second student class substantially outperforms the third class on the second type of items, but underperforms the third class on the first and third types of items. Finally, the four student class may contain the most proficient students, according to the large success probabilities in the last row of $\hat{\mathbf{B}}$. These results suggest that the first and second types of items may measure distinct latent dimensions. It is worth pointing out that, however, the evidence of multidimensionality is quite weak, since there are only 4 and 3 items in the first and the second item latent classes, respectively. To better investigate the latent dimensions measured by the ECPE, it is worth analyzing a larger item pool of ECPE.

In summary, our analysis suggests that the data may be essentially unidimensional and thus retrofitting DCMs to the data set may extract little additional information. This finding is consistent with the conclusion in von Davier (2014)

that a model with multiple skills and interactions between skills is not supported by this data set, as well as the comment in von Davier and Haberman (2014) that the skill distribution presented for this data set in Templin and Bradshaw (2014) provides support for a single ordered latent trait or a continuous latent trait being sufficient. Finally, it is worth pointing out that our results are exploratory that need to be validated by carefully examining the student profiles and item contents within each class and by further statistical inference.

14.6 Discussion

In this chapter, we propose a stochastic co-blockmodel as an exploratory model for cognitive diagnosis. This model reduces the dimensionality of data by co-clustering students and items into latent classes, where student classes may be due to students' attribute profiles and item classes may be explained by the attributes the items measure. A computationally efficient spectral co-clustering algorithm is proposed for fitting the model. We conjecture that this algorithm leads to statistical consistency in clustering under suitable conditions, which may be proved using the techniques in Rohe, Chatterjee, and Yu (2011) that are developed to prove the statistical consistency of a similar spectral clustering for fitting Holland et al. (1983)'s stochastic blockmodel. A cross-validation based approach is also developed for choosing the numbers of student and item classes. By making the connection between the proposed model and diagnostic classification models, and through simulated and real examples, we demonstrate that the proposed approach may be a good EDA tool for cognitive diagnosis modeling.

The proposed spectral co-clustering algorithm performs well when both the numbers of students and items are large. If either N or K is small, a likelihood based approach may be statistically more efficient. Although the classic EM algorithm is computationally infeasible for maximizing the marginal likelihood function of the proposed model, full information estimation may still be available using stochastic optimization algorithms, such as the stochastic EM algorithm (Celeux & Diebolt, 1985; Nielsen et al., 2000) and the Metropolis-Hastings Robbins-Monro algorithm (Cai, 2010). The performance of these algorithms and the comparison between different model fitting approaches are left for future investigations.

Finally, we point out that the proposed method has limitations. It does not directly suggest what diagnostic classification model should be used or suggest the Q-matrix for the items. To arrive at a confirmatory DCM for cognitive diagnosis, much subsequent analysis is need, including the development, estimation, and comparison of DCMs and the validation of the attributes, requiring inputs from both domain experts and psychometricians.

References

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics, 37*, 3099–3132. https://doi.org/10.1214/09-AOS689

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics, 41*, 2097–2122. https://doi.org/10.1214/13-AOS1138

Banerjee, S., & Roy, A. (2014). *Linear algebra and matrix analysis for statistics*. New York, NY: CRC Press.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75*, 33–57. https://doi.org/10.1007/s11336-009-9136-x

Celeux, G., & Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly, 2*, 73–82.

Chen, Y., Li, X., Liu, J., Xu, G., & Ying, Z. (2017). Exploratory item classification via spectral graph clustering. *Applied Psychological Measurement, 41*, 579–599. https://doi.org/10.1177/0146621617692977

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika, 82*, 660–692. https://doi.org/10.1007/s11336-016-9545-6

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*, 850–866. https://doi.org/10.1080/01621459.2014.934827

Choi, D., & Wolfe, P. J. (2014). Co-clustering separately exchangeable network data. *The Annals of Statistics, 42*, 29–63. https://doi.org/10.1214/13-AOS1173

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B, 39*, 1–38.

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco (pp. 269–274).

Golub, G. H., & van Loan, C. F. (2012). *Matrix computations*. Baltimore, MD: JHU Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association, 67*, 123–129. https://doi.org/10.1080/01621459.1972.10481214

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210. https://doi.org/10.1007/s11336-008-9089-5

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks, 5*, 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Joseph, A., & Yu, B. (2016). Impact of regularization on spectral clustering. *The Annals of Statistics, 44*, 1765–1791. https://doi.org/10.1214/16-AOS1447

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272. https://doi.org/10.1177/01466210122032064

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548–564. https://doi.org/10.1177/0146621612456591

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli, 19*, 1790–1817. https://doi.org/10.3150/12-BEJ430

Nielsen, S. F., et al. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli, 6*, 457–489. https://doi.org/10.2307/3318671

Qin, T., & Rohe, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3120–3128). Red Hook: NY Curran.

R Core Team. (2013). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Retrieved from http://www.R-project.org/

Rohe, K., Chatterjee, S., & Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics, 39*, 1878–1915. https://doi.org/10.1214/11-AOS887

Rohe, K., Qin, T., & Yu, B. (2016). Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences, 113*, 12679–12684. https://doi.org/10.1073/pnas.1525793113

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317–339. https://doi.org/10.1007/s11336-013-9362-0

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*, 37–50. https://doi.org/10.1111/emip.12010

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series, 2014*, 1–13. https://doi.org/10.1002/ets2.12043

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional diagnostic classification models: A commentary. *Psychometrika, 79*, 340–346. https://doi.org/10.1007/s11336-013-9363-z

Chapter 15
Recent Developments in Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT): A Comprehensive Review

Xiaofeng Yu, Ying Cheng, and Hua-Hua Chang

Abstract In this chapter, we provide a comprehensive and up-to-date review of cognitive diagnosis computer adaptive testing (CD-CAT). Similar to Cheng and Keng (Computerized adaptive testing in criterion-referenced testing. In Smith E, Stone G (eds) Applications of Rasch measurement in criterion-reference testing: practive analysis to score reporting. JAM Press, Maple Grove, 2009), which provided a flowchart for a typical CAT, we provide a typical CD-CAT flowchart. Compared to regular CAT, a key distinction is that in CD-CAT the goal is to obtain the latent mastery profile for each respondent in an efficient fashion, or alternatively to obtain both the latent mastery profile (formative) and the latent ability (summative) simultaneously. The former is referred to as single-purpose CD-CAT, and the latter dual-purpose CD-CAT. We discuss the main components of CD-CAT in this chapter. These components will be covered in the following order: starting rule, item selection strategies, stopping rule, scoring rule, and item bank development and more specifically online calibration.

This work is partially supported by NSF CAREER grant DRL-1350787 awarded to the corresponding author.

X. Yu
Department of Psychology, University of Notre Dame, Notre Dame, IN, USA

Jiangxi Normal University, Nanchang, China

Y. Cheng (✉)
Department of Psychology, University of Notre Dame, Notre Dame, IN, USA
e-mail: ycheng4@nd.edu

H.-H. Chang
Department of Educational Studies, Purdue University, IN, USA
e-mail: hhchang@illinois.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_15

15.1 Introduction

15.1.1 Basics of Computerized Adaptive Testing

An adaptive test is a computer-based test that dynamically adjusts itself to the estimated latent trait(s) level of each respondent during the test process to efficiently measure a test taker's latent trait(s) (e.g., math ability or English proficiency). A typical computerized adaptive testing (CAT) system includes five major components: a calibrated item bank, a starting rule, an item selection strategy, a scoring method, and a stopping rule. The adaptive feature, the key element of a CAT system, lies in the adaptive item selection, meaning different items are chosen for different examinees, or adaptive test termination, meaning that the test length varies among test takers, or both.

In general, an item bank with well-calibrated items is required for any CAT system. Given the bank, CAT begins with applying a certain rule, which is the starting rule, to start the test by selecting a first item (or first set of items) for an examinee. The scoring method refers to the estimation method used to obtain interim estimate of the latent trait(s), such as maximum likelihood (ML). Based on the most recent estimate, the item selection strategy (ISS) then determines which item or items to pick next. If a test consists of several stages where ISS selects a set of items for each stage, it is often referred to as multistage testing (MST; Yan, Lewis, & von Davier, 2014). MST adopts a group-level sequential design, whereas a typical CAT adopts item-level sequential design (Wang, Lin, Chang, & Douglas, 2016). In this regard, MST can be considered a variation of CAT. Depending on the stopping rule, CAT is often categorized into two types: fixed-length CAT and variable-length CAT.

Abundant literature exists on each of these important components of CAT. Item bank development requires an adequate number of items to be written and calibrated given a chosen item response theory (IRT) model. Most of the current operational CAT systems are built on unidimensional IRT models (e.g., the three-parameter logistic (3PL) model). There is an important line of CAT research on how to develop and maintain an item bank for CAT (He & Reckase, 2014; Reckase, 2010; Veldkamp & van der Linden, 2000), and online calibration of item parameters (Chen, Xin, Wang, & Chang, 2012; Stocking, 1988; Wainer & Mislevy, 2000).

Researchers have also proposed different starting rules. Note that initial items need to be selected in a non-deterministic manner; otherwise those items that are used often in the beginning of the test may quickly become known to test takers, which jeopardizes test security. Random selection is therefore a popular choice. Eggen and Straetmans (2000) suggested random selection from relatively easy items in the bank to start the test. Or a short pretest can be administered first so the testing program can collect additional information. Given the preliminary information, the formal testing will start with different items for different examinees (Riley, Conrad, Bezruczko, & Dennis, 2007).

There has also been a long-standing history of research on the ISS. In 1980, the maximum Fisher information (MFI) method was proposed by Lord (Lord, 1980; Thissen & Mislevy, 2000), which was the most popular ISS in the early days of CAT. The Fisher information measures the amount of information for the unknown ability θ produced by a response pattern. It can be determined by:

$$I\left(\theta\right) = -\mathrm{E}\left\{\left[\frac{\partial^2 \ln f\left(\boldsymbol{x};\theta\right)}{\partial\theta^2}\right]\right\}, \tag{15.1}$$

where $f(\boldsymbol{x};\theta)$ represents the likelihood function, θ is the underlying latent trait, and $\boldsymbol{x}$ refers to the observed response pattern. The item k's Fisher information is given by

$$I_k\left(\theta\right) = \frac{\left[P_k'\left(\theta\right)\right]^2}{P_k\left(\theta\right)Q_k\left(\theta\right)}, \tag{15.2}$$

where $P_k(\theta)$ is the item response function of item k specified by the chosen IRT model, and $Q_k(\theta) = 1 - P_k(\theta)$, and $P_k'(\theta)$ refers to the first derivative of the item response function with respect to θ. Assuming local independence the test information $I(\theta)$ is additive in item information, that is, $I(\theta) = \sum I_k(\theta)$.

For the three-parameter logistic (3PL) model, $P_j(\theta)$ is given by

$$P_k\left(\theta\right) = c_k + \left(1 - c_k\right)\frac{e^{a_k(\theta - b_k)}}{1 + e^{a_k(\theta - b_k)}}, \tag{15.3}$$

where a_k, b_k and c_k refer to the discrimination, difficulty, and guessing parameter for the k^{th} item, respectively. If the MFI method is used for item selection, an eligible item in the bank with the largest Fisher information given the current estimate of θ will be picked as the next item for administration. As the asymptotic variance of $\hat{\theta}_{ML}$, the maximum likelihood estimate of θ, is inversely proportional to the test information, the MFI method is widely regarded as a method to minimize the asymptotic variance of the θ estimate, or in other words, to asymptotically maximize measurement precision.

However, the ability estimate may not be accurate yet in the early stage of CAT. Maximizing information based on an unstable and inaccurate θ estimate can be characterized as "capitalization on chance". Thus, using the MFI early in a CAT program may not be ideal. Researchers also found that the MFI tends to select items with large discrimination parameters, but rarely uses items with smaller discrimination parameters. This means that a portion of items in the item bank can be grossly underutilized. Meanwhile, overexposure of a small number of highly discriminating items may pose a serious threat to test security. Additionally, to ensure face and content validity of CAT, the number of items from different content areas or subdomains oftentimes need to be balanced. Motivated by these concerns, many researchers have proposed alternative ISSs than the MFI method to address (a)

capitalization on chance, particularly early in CAT (Patton, Cheng, Yuan, & Diao, 2013; van der Linden & Glas, 2000); (b) the balance in item bank usage and control of item overexposure (Chang, 2015; Chang & Ying, 1999); and (c) the balance in test content (Cheng, Chang, & Yi, 2007; Yi & Chang, 2003).

Any ISS would rely on a suitable scoring rule that provides an efficient update of the ability estimate. Three mostly widely used methods for scoring in regular CAT are the maximum likelihood estimation (MLE), the maximum a posterior (MAP), and the expected a posterior (EAP) (Baker & Kim, 2004). The advantage of MLE lies in its asymptotic consistency and efficiency. A limitation of MLE, however, is that it leads to undefined estimates for those respondents with all incorrect or all incorrect answers, which can be common in the early stage of a test. The two Bayesian methods in contrast assume a prior of θ and therefore could avoid an undefined estimate. Once the posterior distribution $P(\theta|\boldsymbol{x})$ is obtained, the MAP and EAP takes the mode or the mean of the posterior distribution as the estimate of θ.

Another important component of CAT is the stopping rule and there are two main types. In a fixed-length CAT system, the test would be terminated when a pre-specified number of items are administered to a test taker. In a variable-length CAT system, the test would be ended when a pre-specified level of precision of the ability estimate is reached, or when a classification decision is ready to be made with pre-specified level of confidence.

15.1.2 Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT)

Cognitive Diagnostic Assessment (CDA) has both features of model-based measurement and formative assessment (Embretson, 2001). Instead of focusing on providing a summative score for each test taker, CDA tries to pinpoint the strengths and weaknesses on fine-grained skills (often referred to as attributes) for test takers. As described in Gierl and Zhou (2008), CDA oftentimes employs a cognitive diagnostic model (CDM; Rupp, Templin, & Henson, 2010), which assumes that one's responses to a test are governed by one's latent profile of mastery. Similar to an IRT model, a CDM specifies the probability of answering an item correctly given the item characteristics and the latent mastery profile of a test taker. Over the past decade or so, researchers have proposed dozens of CDMs with different parameterizations. For most of them, item characteristics include structural parameters that resemble the item parameters in IRT models, and the **Q** matrix is a K by D matrix that specifies the item-attribute relationship, where K is the total number of items on the test and D is the number of attributes. Elements in the **Q** matrix indicate whether an attribute is required to answer the item correctly or not. The entry q_{kd} equals 1 if the item k requires the mastery of attribute d, and $q_{kd} = 0$ otherwise, $d = 1, 2, \ldots, D$, and $k = 1, 2, \ldots, K$. The latent mastery profile for the n^{th} examinee is denoted by $\boldsymbol{\alpha}_n$,

where $\boldsymbol{\alpha}_n = \left(\alpha_{n1},\ \alpha_{n2},\ \cdots, \alpha_{nd},\ \cdots, \alpha_{nD}\right)'$, with $\alpha_{nd} = 1$ if the n^{th} examinee has mastered the d^{th} attribute, and $\alpha_{nd} = 0$ otherwise, where $n = 1, 2, \ldots, N$, N being the total sample size.

CD-CAT, as suggested by its name coined in Cheng (2009), is computerized adaptive testing built on a CDM. As the CDMs evolve, CD-CAT has also received increasing amount of attention. Xu, Chang, and Douglas (2003) conducted one of the earliest studies on developing ISSs for CD-CAT. Cheng (2008, 2009) further proposed and examined the performance of new ISSs for CD-CAT and examined their performances through simulation studies. An overview of CD-CAT was provided by Huebner (2010), which covered the major components of CD-CAT, such as item selection for the initial stage, item selection strategy, the methods for updating the attribute mastery patterns, and the stopping rule. These mirrored the components in a typical CAT system.

However, Huebner (2010) did not cover the development and maintenance of an item bank for CD-CAT. In addition, fueled by the demand of formative assessment, there has been numerous developments of CD-CAT since 2010. In this chapter, we will provide a more comprehensive and up-to-date review of CD-CAT. Similar to Cheng and Keng (2009), which provided a flowchart for a typical CAT, we created a typical CD-CAT flowchart in Fig. 15.1. A key distinction here is that in CD-CAT, the goal is to obtain the latent mastery profile for each respondent in an efficient fashion, or alternatively to obtain both the latent mastery profile (formative) and the latent ability (summative) simultaneously. The former is referred to as single-purpose CD-CAT, and the latter dual-purpose CD-CAT. Depending on the goal, some of the components of CD-CAT such as ISSs can be very different.

Next, we will discuss the corresponding components of CD-CAT. For the sake of convenience, these components will be covered in the following order: starting rule, ISS, stopping rule, scoring rule, and item bank development and more specifically online calibration.

15.2 Starting Rules

Theoretically, a CAT system can start at any level of difficulty. The simplest rule is to start the test for each respondent with the same item, or assuming the same ability level, such as 0, and then use some ISS to find the best item. In a CD-CAT, one can start the test by assuming the same attribute profile such as the most common profile in the population for every test taker. Such a simplistic rule may cause problem because the frequently used items in the beginning of the test may quickly become known to test takers. Therefore, some randomization mechanism is often introduced to the starting rule. For example, one rule is to start the test assuming a random θ chosen from an interval (e.g., $[-0.5, 0.5]$). Or one can randomly select the first item(s) among a predetermined set of items, often items in the middle range of difficulty. Riley et al. (2007) suggested administering a short pretest to gain initial information on each respondent before the test starts. For a CD-CAT system, some

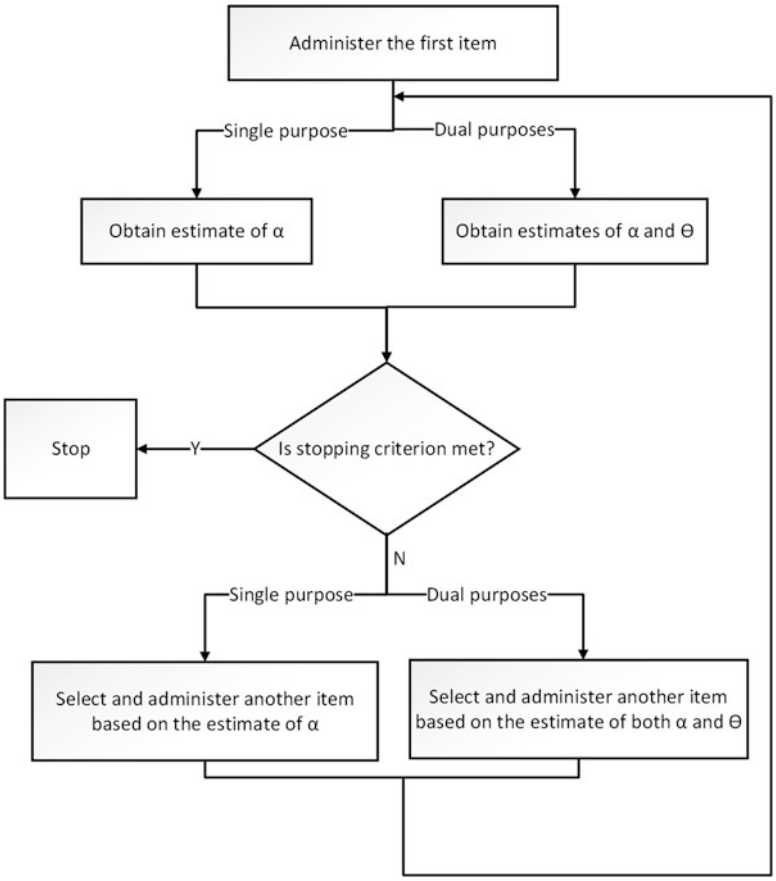


Fig. 15.1 The flowchart of a typical CD-CAT system

researches proposed to start the test with the same randomly chosen items for each respondent (e.g., Xu et al., 2003), some considered assigning a randomly generated mastery profile to each respondent and then pick items based on some ISS to suit that randomly generated profile (e.g., Chen et al., 2012). Von Davier and Cheng (2014) discussed the approach taken by the Programme for the International Assessment of Adult Competencies (PIAAC), which was to use auxiliary information collected from a background questionnaire to determine an initial set of items.

15.3 Item Selection Strategies

Most ISSs under a regular CAT are not directly applicable in CD-CAT, because the latter is based on constrained latent classes instead of continuous latent trait(s). To address this, researchers began to look for new algorithms that can be adopted in

CD-CAT. On one hand, the main purpose for ISSs under CD-CAT is to estimate each respondent's latent mastery profile $\boldsymbol{\alpha}$, or θ and $\boldsymbol{\alpha}$ simultaneously. On the other hand, the concerns in a typical CAT such as content and exposure control apply in CD-CAT as well. ISSs that have been developed for CD-CAT can be categorized into three groups: single-purpose for $\boldsymbol{\alpha}$ estimation, dual-purpose for estimation of both $\boldsymbol{\alpha}$ and θ, and ISSs that account for non-statistical constraints.

15.3.1 Single-Purpose ISS for Estimation of $\boldsymbol{\alpha}$

Two fundamental ISSs were proposed in the early 2000's. One is based on Kullback-Leibler information (Xu et al., 2003), and the other based on Shannon entropy (Tatsuoka, 2002).

Kullback-Leibler Information (KL) Method The Kullback–Leibler (K-L) information measures the divergence between two probability distributions (Cover & Thomas, 1991). Let $K(f\|g)$ represent K-L information between probability density functions $f(x)$ and $g(x)$. $K(f\|g)$ is defined as:

$$K\left(f\|g\right) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x)\mathrm{d}x. \tag{15.4}$$

If x is a discrete random variable, the integral in Eq. (15.4) can be replaced by summation.

$$K\left(f\|g\right) = \sum f(x) \log\left(\frac{f(x)}{g(x)}\right). \tag{15.5}$$

Suppose t items have been selected and administered in a CD-CAT program. Denote the set of items in the bank that are currently eligible for selection as $R^{(t)}$ and consider item h in $R^{(t)}$. Our focus is the conditional distribution of examinee n's response based on his or her cognitive profile $\boldsymbol{\alpha}_n$ (Cheng, 2009). The K-L distance between the distribution of x_{nh}, the response of person n to item h, given the most recent latent mastery profile estimate $\widehat{\boldsymbol{\alpha}}_n^{(t)}$ and the distribution of x_{nh} given any latent state $\boldsymbol{\alpha}_c$ can be computed as follows:

$$K_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\|\boldsymbol{\alpha}_c\right) = \sum_{x=0}^{1} \log\left(\frac{P\left(x_{nh}=x|\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)}{P\left(x_{nh}=x|\boldsymbol{\alpha}_c\right)}\right) P\left(x_{nh}=x|\widehat{\boldsymbol{\alpha}}_n^{(t)}\right). \tag{15.6}$$

A large $K_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\|\boldsymbol{\alpha}_c\right)$ suggests that item h provides large information to differentiate $\widehat{\boldsymbol{\alpha}}_n^{(t)}$and $\boldsymbol{\alpha}_c$. If $\boldsymbol{\alpha}_c$ is the true latent state, selecting such an item would help telling $\widehat{\boldsymbol{\alpha}}_n^{(t)}$ apart from the truth. In reality the true state is unknown. When there are

D attributes, there are 2^D possible latent states. Thus, Xu et al. (2003) suggested the following index for item selection:

$$KL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)} \| \boldsymbol{\alpha}\right) = \sum_{c=1}^{2^D} K_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)} \| \boldsymbol{\alpha}_c\right). \tag{15.7}$$

A large $KL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)} \| \boldsymbol{\alpha}\right)$ indicates that the item h contains a large amount of information to differentiate between $\widehat{\boldsymbol{\alpha}}_n^{(t)}$ from any possible $\boldsymbol{\alpha}_c$. Hence Xu et al. (2003) suggested to select items with the largest $KL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)} \| \boldsymbol{\alpha}\right), h \in R^{(t)}$ as the $(t+1)$-th item. This is referred to as the KL method in Cheng (2009).

Shannon Entropy (SHE) Method The Shannon entropy is a function of a random variable's probability distribution, which measures the uncertainty associated with the distribution (Shannon, 1948). Let $\boldsymbol{\Omega} = (\mathrm{F}, p)$ be a discrete probability space with M elements. In other words, each F_m has its probability p_m, $m = 1, 2, \ldots M$. The Shannon entropy of $\boldsymbol{\Omega}$ can be determined as

$$SH\left(\boldsymbol{\Omega}\right) = \sum_{m=1}^{M} p_m \log_b \frac{1}{p_m}, \tag{15.8}$$

where b without a subscript is the base of the logarithm used.

Denote the prior of the latent class as

$$p\left(\boldsymbol{\alpha}_c\right) = \pi_{0c}, c = 1, 2, \cdots, 2^D, \tag{15.9}$$

subject to $\sum_{c=1}^{2^D} \pi_{0c} = 1$ and $\pi_{0c} \geq 0$. The corresponding posterior distribution of the latent state after t items have been administered is

$$\pi_t\left(\boldsymbol{\alpha}_c | \boldsymbol{x}^{(t)}\right) \propto \pi_{0c} \prod_{k=1}^{t} p_{kc}^{x_k} (1 - p_{kc})^{(1-x_k)}, \tag{15.10}$$

where p_{kc} is the response probability to item k given latent state $\boldsymbol{\alpha}_c$ specified by a CDM, and $\boldsymbol{x}^{(t)}$ is the response vector to the t items that have been administered, which contains t elements, from x_1 to x_t. The SHE of the posterior distribution π_t can be determined as

$$SH\left(\pi_t | \mathbf{x}^{(t)}\right) = -\sum_{c=1}^{2^D} \pi_t\left(\boldsymbol{\alpha}_c\right) \log_b\left(\pi_t\left(\boldsymbol{\alpha}_c\right)\right). \tag{15.11}$$

When item h is administered, the expected SHE can be determined as

$$SH\left(\pi_h\left(\boldsymbol{\alpha}_c\right)\right) = -\sum_{x=0}^{1} SH\left(\pi_t|\boldsymbol{x}^{(t)}, x_h = x\right).p\left(x_h = x|\boldsymbol{x}^{(t)}\right). \tag{15.12}$$

Naturally, Tatsuoka (2002) proposed to select items that minimize the expected SHE of the posterior distribution of the latent state as quantified in Eq. (15.12). This is referred to as the SHE method in Cheng (2009).

Since the early 2000's many variations have been proposed based on the KL method and the SHE method. For example, Cheng (2009) noted that the KL algorithm assumed that all the latent states are equally likely, which is not necessary and might cause inefficiency. Cheng (2009) therefore proposed two alternative methods: the posterior-weighted KL method and the hybrid method.

Posterior-Weighted KL (PWKL) method The PWKL method modifies the original KL index in Eq. (15.7) by weighting each KL divergence with its respective probability of each possible latent state and created the PWKL index:

$$PWKL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right) = \sum_{c=1}^{2^D} KL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\|\boldsymbol{\alpha}_c\right)\pi_t\left(\boldsymbol{\alpha}_c\right), \tag{15.13}$$

where $\pi_t(\boldsymbol{\alpha}_c)$ is the probability of the latent state $\boldsymbol{\alpha}_c$ after t items have been administered. Cheng (2009) proposed to select the item in $R^{(t)}$ that maximizes $PWKL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)$. Compared to the original KL method, the $PWKL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)$ gives higher weight to the latent states that are more likely given previous responses.

Hybrid KL (HKL) Method Based on the PWKL index, HKL can be further weighted by the inverse of the distance between the current $\widehat{\boldsymbol{\alpha}}_n^{(t)}$ and other latent states:

$$HKL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right) = \sum_{c=1}^{2^D} KL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\|\boldsymbol{\alpha}_c\right)\pi_t\left(\boldsymbol{\alpha}_c\right)\frac{1}{dis\left(\boldsymbol{\alpha}_c, \widehat{\boldsymbol{\alpha}}_n^{(t)}\right)}, \tag{15.14}$$

where $dis\left(\boldsymbol{\alpha}_c, \widehat{\boldsymbol{\alpha}}_n^{(t)}\right)$ represents the Euclidean distance between $\boldsymbol{\alpha}_c$ and $\widehat{\boldsymbol{\alpha}}_n^{(t)}$:

$$dis\left(\boldsymbol{\alpha}_c, \widehat{\boldsymbol{\alpha}}_n^{(t)}\right) = \sqrt{\sum_{d=1}^{D}\left(\alpha_{cd} - \widehat{\alpha}_{nd}^{(t)}\right)^2}. \tag{15.15}$$

Such weighting favors items that can tell apart latent states that are close, or in other words, that are difficult to be distinguished from each other. Cheng (2009) proposed to select items that maximize the HKL index and reported that the PWKL and HKL methods outperform the KL and SHE methods in terms of efficiency.

Furthermore, Cheng (2009) showed the relationship between the KL and SHE methods, that is, "minimizing the expected SHE of the predicted posterior is equivalent to maximizing the expected KL distance between the predicted posterior and the discrete uniform distribution". It also discussed the connection between the KL method for CD-CAT and the global information method proposed by Chang and Ying (1996) for regular CAT, as well as the relationship between the SHE method for CD-CAT and the minimum expected posterior variance (MEPV) method for the regular CAT (van der Linden, 1998).

The methods described above were built with asymptotic statistical efficiency in mind. Similar as under regular CAT, there exist concerns over capitalization on chance in the beginning of a test, or for a short test. Some methods were developed to specifically address such concerns, for example, the mutual information method (MI; Wang, 2013), the modified PWKL and GDI method (Kaplan, de la Torre, & Barrada, 2015), and methods based on the CDM discrimination index or the CDI (Henson & Douglas, 2005; Zheng & Chang, 2016).

Mutual Information (MI) Method Mutual information is a measure of mutual dependence for two random variables, or information that can be obtained on one random variable from the other. Consider two discrete random variables Y and Z with joint distribution $p(y,z)$, the corresponding MI between Y and Z is given by

$$MI\left(Y\mid\mid Z\right)=\sum_{y,z}p\left(y,z\right)\log\frac{p\left(y,z\right)}{p(y)p(z)}. \tag{15.16}$$

Suppose $\pi(\boldsymbol{\alpha}|\boldsymbol{x}^{(t)})$and $p(x_h|\boldsymbol{x}^{(t)})$ are the posterior distribution of the latent state and the Bernoulli distribution of the next response given the responses on the first t items, respectively. Then the mutual information between them indicates the information obtained about the unknown $\boldsymbol{\alpha}$ when the item h is added to the test, which is:

$$\begin{aligned}&MI\left(\pi\left(\boldsymbol{\alpha}|\boldsymbol{x}^{(t)}\right)\|\mathrm{p}\left(x_h|\boldsymbol{x}^{(t)}\right)\right)\\&=\sum_{x=0}^{1}\sum_{c=1}^{2^D}p\left(\boldsymbol{\alpha}_c,x_h=x|\boldsymbol{x}^{(t)}\right)\log\left[\frac{p\left(\boldsymbol{\alpha}_c,x_h=x|\boldsymbol{x}^{(t)}\right)}{\pi\left(\boldsymbol{\alpha}_c|\mathbf{x}^{(t)}\right)p\left(x_h=x|\boldsymbol{x}^{(t)}\right)}\right].\end{aligned} \tag{15.17}$$

The elements $p(x_h=x|\boldsymbol{x}^{(t)})$ and $p(\boldsymbol{\alpha}_c,\ x_h=x|\boldsymbol{x}^{(t)})$ both involve summations over 2^D possible latent classes. Wang (2013) noted that MI-based algorithm requires a triple summation over 2^D possible cognitive profiles and is therefore computationally intensive. She provided the computational simplification for the MI-based algorithm in the paper. Based on the simulation results of Wang (2013),

MI is more efficient than the other competing indices for short tests. This is because it utilizes the full posterior of $\pi(\boldsymbol{\alpha}|\boldsymbol{x}^{(t)})$ instead of a point estimate from $\pi(\boldsymbol{\alpha}|\boldsymbol{x}^{(t)})$.

Modified Posterior-Weighted Kullback-Leibler (MPWKL) Method The motivation of the PWKL method (MPKWL; Kaplan et al., 2015) is to consider the entire posterior distribution of 2^D possible latent states when the test is relatively short, instead of focusing on the latest latent state estimate $\widehat{\boldsymbol{\alpha}}_n^{(t)}$. Again, this is because $\widehat{\boldsymbol{\alpha}}_n^{(t)}$ may not be stable or reliable yet at the beginning of CAT. The MPWKL index can be computed as

$$MPWKL_h^{(t)} = \sum_{c1=1}^{2^D}\left[\sum_{c2=1}^{2^D}\left[\sum_{x=0}^{1}\log\left(\frac{p\left(X_h=x|\boldsymbol{\alpha}_{c1}\right)}{p\left(X_h=x|\boldsymbol{\alpha}_{c2}\right)}\right)p\left(X_h=x|\boldsymbol{\alpha}_{c1}\right)\pi_t\left(\boldsymbol{\alpha}_{c2}\right)\right]\pi_t\left(\boldsymbol{\alpha}_{c1}\right)\right]. \tag{15.18}$$

Generalized Model Discrimination Index (GDI) Method Kaplan et al. (2015) proposed the GDI index which takes the (weighted) variance of the probabilities of answering an item correctly given a particular attribute distribution into consideration. Let A_h^* denote the set of attributes measured by item h, and $\boldsymbol{\alpha}_h^*$ the reduced mastery vector for the corresponding A_h^* attributes (de la Torre, 2011). $\overline{P}_h$ refers to the mean probability of answering item h correctly: $\overline{P}_h = \sum_{c=1}^{2^{A_h^*}}\pi^{(t)}\left(\boldsymbol{\alpha}_{ch}^*\right)P\left(X_h=1|\boldsymbol{\alpha}_{ch}^*\right)$, where $\pi^{(t)}\left(\boldsymbol{\alpha}_{ch}^*\right)$ is the posterior probability of the reduced attribute vector $\boldsymbol{\alpha}_{ch}^*$ after t items have been administered. The GDI is defined as:

$$GDI_h^{(t)} = \sum_{c=1}^{2^{D_h^*}}\pi^{(t)}\left(\boldsymbol{\alpha}_{ch}^*\right)\left[P\left(X_h=1|\boldsymbol{\alpha}_{ch}^*\right)-\overline{P}_h\right]^2. \tag{15.19}$$

It is the weighted variance of the probability of answering item h correctly. Items with larger GDI are better at differentiating among the reduced attribute vectors. Kaplan et al. (2015) showed that maximizing the GDI works well for short tests, and by using a reduced attribute vector the GDI is computationally more efficient.

Posterior-Weighted/Posterior-Weighted Attribute-Level CDM Discrimination Index (PWCDI/PWACDI) Method Based on the CDI index (CDM discrimination index; Henson & Douglas, 2005), Zheng and Chang (2016) proposed two algorithms that can be used for short-length CD-CAT, namely PWCDI and PWACDI. The original purpose for developing the CDI index is to facilitate test construction based on CDM. For a specific item h, the CDI DIS_h is a $2^D \times 2^D$matrix whose entry DIS_{huv} is the expected KL distance between the response distributions given latent classes $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$:

$$DIS_{huv} = E_{\alpha_u}\left[\log\left(\frac{P\left(X_h|\boldsymbol{\alpha}_u\right)}{P\left(X_h|\boldsymbol{\alpha}_v\right)}\right)\right], \tag{15.20}$$

The posterior-weighted DIS_{huv} is referred to as the PWD_{huv}:

$$PWD_{huv} = E_{\alpha_u}\left[\pi\left(\boldsymbol{\alpha}_u\right)\pi\left(\boldsymbol{\alpha}_v\right)\log\left(\frac{P\left(X_h|\alpha_u\right)}{P\left(X_h|\alpha_v\right)}\right)\right], \tag{15.21}$$

where $\boldsymbol{\alpha}_u$and $\boldsymbol{\alpha}_v$ are two candidate latent classes (u, $v = 1, 2, \cdots, 2^D$), and $\pi(\boldsymbol{\alpha}_u)$ and $\pi(\boldsymbol{\alpha}_v)$ are the posterior probabilities of these two states, respectively.

The PWCDI and PWACDI indices can be determined as follows:

$$PWCDI_h = \frac{1}{\sum_{u\neq v} H(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}}\sum\nolimits_{u\neq v} H(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} PWD_{huv}, \tag{15.22}$$

$$PWACDI_h = \sum_{k=1}^{K}\frac{1}{2^K}\sum_{all\ relevant\ cells} PWD_{huv}, \tag{15.23}$$

where $H\left(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v\right) = \sum_{d=1}^{D}\left|\alpha_{u1} - \alpha_{v1}\right|$ refers to the Hamming distance between $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$. All relevant cells denote the entries in the DIS_h matrix where only the d^{th} attribute is different for attribute patterns $\boldsymbol{\alpha}_u$and $\boldsymbol{\alpha}_v$. Because the calculation of CDI does not rely on the provisional attribute profile, Zheng and Chang (2016) showed that maximizing the PWCDI and PWACDI leads to better performance than maximizing the KL index in the early stage of CD-CAT.

15.3.2 Dual-Purpose ISS for Both α and θ Estimation

The different ISS discussed in the previous sections are designed for tests that focus on the estimation of each test taker's latent profile. However, in applications, providing the latent trait information (i.e., a summative score) can also be desirable. Thus, it is of practical importance to integrate the ability estimation into the ISSs under a CD-CAT system.

Dual-Information Method (DIM) To efficiently estimate both θ and $\boldsymbol{\alpha}$ simultaneously, Cheng and Chang (2007a) proposed the dual-information method, which incorporated the information from both θ and $\boldsymbol{\alpha}$ by using a weighted sum of them.

$$KL_h\left(\widehat{\theta}_n^{(t)}, \widehat{\boldsymbol{\alpha}}_n^{(t)}\right) = \omega \cdot KL\left(\widehat{\theta}_n^{(t)}\right) + (1-\omega)\, KL\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right), \tag{15.24}$$

where ω is a weight with $(0 \leq \omega \leq 1)$, and $KL\left(\widehat{\theta}_n^{(t)}\right)$ is the global information (Chang & Ying, 1996) for item selection under the regular CAT, and $KL\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)$ can be the KL index defined in Eq. (15.7), or the PWKL index defined in Eq. (15.13) or the HKL index in Eq. (15.14). The item with the largest $KL_h\left(\widehat{\theta}_n^{(t)}, \widehat{\boldsymbol{\alpha}}_n^{(t)}\right)$ will be chosen as the next item.

Along this line of research, numerous methods that serve dual purposes have been developed, such as McGlohen and Chang (2008), Wang, Chang, and Douglas (2012), Wang, Zheng, and Chang (2014), Dai, Zhang, and Li (2016), and Kang, Zhang, and Chang (2017).

Two-Stage Method McGlohen and Chang (2008) proposed a two-stage process. In the first stage, a shadow test (van der Linden, 2000) is assembled at the provisional ability estimate $\widehat{\theta}_n^{(t)}$. A shadow test refers to a test that is assembled at each step of item selection that maximizes test information at the current ability estimate, while meeting all non-statistical constraints (see the next section for the discussion of non-statistical constraints). Then in the second stage the SHE or KL method is used to pick the next item for administration from the shadow test. Thus, this method selects items that are suitable for both θ and $\boldsymbol{\alpha}$ estimation.

Aggregated Ranked Information Method (ARI) One notable problem with the DIM is that the information from θ and $\boldsymbol{\alpha}$ can be on different scales. Wang et al. (2014) modified the DIM by combining the percentile ranks of information coming from θ and α:

$$ARI = \omega \cdot pe\left(KL\left(\widehat{\theta}_n^{(t)}\right)\right) + (1-\omega)\, pe\left(PWKL\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)\right), \tag{15.25}$$

where $pe(\cdot)$ refers to percentile rank, ω refers to the weight $(0 \leq \omega \leq 1)$, and $KL\left(\widehat{\theta}_n^{(t)}\right)$ is the global information (Chang & Ying, 1996) evaluated at $\widehat{\theta}_n^{(t)}$. Wang et al. (2014) chose three weighting schemes for the weight ω: Theory-based weights, empirical weights, and attribute-level weights.

When both KL and PWKL index are standardized, the ARI turns into Aggregated standardized information (ASI) method:

$$Z_{KL^*\left(\widehat{\theta}_n^{(t)}\right)} = \frac{\left(KL\left(\widehat{\theta}_n^{(t)}\right) - mean\left(KL\left(\widehat{\theta}_n^{(t)}\right)\right)\right)}{SD\left(KL\left(\widehat{\theta}_n^{(t)}\right)\right)}, \tag{15.26}$$

$$Z_{PWKL^*\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)} = \frac{\left(PWKL\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right) - mean\left(PWKL\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)\right)\right)}{SD\left(PWKL\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)\right)}, \tag{15.27}$$

$$ASI = \omega \cdot Z_{KL\left(\widehat{\theta}_n^{(t)}\right)} + (1-\omega)\left(Z_{PWKL^*\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)}\right). \tag{15.28}$$

Dapperness with Information (DWI) Dai et al. (2016) combined SHE for $\boldsymbol{\alpha}$ estimation and Fisher information for θ estimation into one single index as a ratio between the two:

$$DWI_h = \frac{I_h\left(\widehat{\theta}_n^{(t)}\right)}{SHE_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)}, \tag{15.29}$$

where $I_h\left(\widehat{\theta}_n^{(t)}\right)$ and $SHE_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right)$ are the Fisher information based on the interim ability estimate and SHE index based on the interim posterior density given item h for the n^{th} respondent. By maximizing the ratio, the DWI method favors items that lead to large Fisher information for θ and small SHE for or $\boldsymbol{\alpha}$.

Jensen-Shannon Based Index (JS) To place information for θ and $\boldsymbol{\alpha}$ on the same scale, Kang et al. (2017) proposed a new index $JS(f_{\boldsymbol{\alpha}} \| f_\theta)$ based on Jensen-Shannon divergence,

$$JS\left(f_{\boldsymbol{\alpha}} | \mid f_\theta\right) = \omega KL\left(f_{\boldsymbol{\alpha}} \| g\right) + (1-\omega)\, KL,\left(f_\theta \mid | g\right), \tag{15.30}$$

where $g = \omega f_{\boldsymbol{\alpha}} + (1-\omega) f_\theta$, ω refers to the weight which was set as $\omega(k) = k/(L+1)$ in Kang et al. (2017), and L denotes the test length. $f_{\boldsymbol{\alpha}}$ and f_θ are the two mass functions corresponding to $P(x_k = 1|\,\boldsymbol{\alpha})$ and $P(x_k = 1|\,\theta)$, respectively. They also showed the relationship between $JS(f_{\boldsymbol{\alpha}} \| f_\theta)$ and other common indices, such as SHE and MI.

15.3.3 Constrained ISSs

Different from the above ISSs, constrained ISSs refer to those methods that take non-statistical constraints (van der Linden & Chang, 2003) or exposure control into consideration.

ISSs for Attribute Balancing As mentioned in Cheng (2010), attribute balancing, which is analogous to content balancing in a traditional CAT program, is very important for a CD-CAT program. Content balancing typically means balancing the proportion of items from different content areas or subdomains. A content balancing constraint is often imposed by test developers to ensure face validity and content validity of a test. It may come in the form of an upper bound of the number of items that can appear on the test from a certain content area, or in the form of both upper and lower bounds (Cheng & Chang, 2007b; Cheng, Chang, & Yi, 2007). The goal is to ensure that a content area is neither under- nor –over represented on a regular

CAT. In a CD-CAT program, attribute balancing means that the representation of each attribute on the test needs to be balanced.

MGDI Method Based on the KL index, Cheng (2010) proposed selecting items with the maximum MGDI, or the modified global discrimination index (MGDI), for the purpose of attribute balancing in CD-CAT. The MGDI index is formulated by weighting the KL index with an attribute-balancing index. The attribute-balancing index takes the following form,

$$\prod_{d=1}^{D}\left(\frac{B_d - b_d}{B_d}\right)^{q_{hd}}, \tag{15.31}$$

and the MGDI can be determined as follows:

$$MGDI_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)}\right) = \prod_{d=1}^{D}\left(\frac{B_d - b_d}{B_d}\right)^{q_{hd}} KL_h\left(\widehat{\boldsymbol{\alpha}}_n^{(t)} \| \boldsymbol{\alpha}_c\right), \tag{15.32}$$

where B_d is the minimum number of items required to measure the d^{th} attribute, and b_d represents the number of items measuring the d^{th} attribute that have already been selected. As we can see, if $q_{hd} = 0$, $\left(\frac{B_d - b_d}{B_d}\right)^{q_{hd}}$ is 1. This indicates that q_{hd} does not affect the corresponding MGDI. If B_d equals b_d, it means the d^{th} attribute has been measured by the minimum required number of items, and then MGDI would equal 0. Under these circumstances, items measuring the d^{th} attribute will have a MGDI of 0, and hence will not be favored by the MGDI method. On the other hand, items that tap into an under-represented attribute will be favored by the MGDI. The MGDI method can be viewed as analogous to the flexible content balancing method proposed in Cheng, Chang and Yi (2007) for regular CAT, when only lower bounds are imposed on attributes. It could also be viewed as an extension to CD-CAT of the maximum-priority-index (MPI) method proposed in Cheng and Chang (2009) for regular CAT when only simple content balancing constraints are present.

Q-control Method Wang et al. (2012) proposed the **Q**-control method to handle both upper and lower bounds in attribute balancing. Let q_{hd} denote the d^{th} entry in the **Q** matrix for the h^{th} item, u_d and l_d denote the upper and lower bounds on how many times each attribute should be measured, respectively. Let v_d represent the number of items that have been selected that measure attribute d. The control index P_h can be determined as

$$P_h = \prod_{d=1}^{D}\left[\frac{u_d - v_d - q_{hd}}{u_d}\right]\left[\frac{(L - l_d) - (t - v_d - q_{hd})}{L - l_d}\right]. \tag{15.33}$$

The control index P_h is then multiplied to the appropriate information index, for example, the KL information in Eq. (15.7) to formulate an item selection index

with built-in attribute balancing. For instance, Wang et al. (2012) proposed the MIQ index as follows:

$$\text{MIQ}_h = I_h\left(\widehat{\theta}_n^{(t)}\right) P_h, \tag{15.34}$$

where $I_h\left(\widehat{\theta}_n^{(t)}\right)$ denotes the item Fisher information at the current ability estimate. The MIQ index can serve as a dual-purpose item selection method.

ISSs for Exposure Control The ISSs may lead to some items to be overexposed, and some other items underexposed. Overexposed items may pose a threat to test security, and many underexposed items imply wasted resources on item writing and item bank development. It is therefore desirable to control item overexposure and balance item exposure in the bank.

RP_PWKL Method Wang, Chang, and Huebner (2011) proposed two methods for expose control based on the PWKL index, namely the P_PWKL method and the RP_PWKL method. The P_PWKL index is defined as follows:

$$P\_PWKL_h = \left(1 - \frac{t}{L}\right) R_h + \beta \frac{t}{L} PWKL_h, \tag{15.35}$$

where L is the test length, t refers to the number of items that have been administered, $\beta > 0$ is the weight which can be adjusted to exert more or less stringent control of item exposure, and R_h is a random number generated from $U(0, \max\{PWKL_h, h \in R^{(t)}\})$. The P_PWKL index is essentially a weighted sum of a random number and the PWKL. By adding that random number component, the P_PWKL method introduces more uncertainty to item selection and therefore results in more balanced item usage.

To further suppress item overexposure, the RP_PWKL index was proposed:

$$RP\_PWKL_h = \left(1 - \frac{\exp_h}{r}\right) P\_PWKL_h, \tag{15.36}$$

where r refers to the maximum exposure rate allowed, and exp_h is the exposure rate for item h. Wang et al. (2011) suggested selecting items with the largest RP_PWKL index at every step. Items that have already been administered very frequently would have high exp_h and consequently less favored by the RP_PWKL method. The P_PWKL method mimics the progressive approach for exposure control in a regular CAT proposed in Revuelta and Ponsoda (1998), where P_ means that the method is "progressively" relying more on the information component as test progresses. The RP_PWKL method is the restricted-progressive approach which added a restrictive cap on the maximum exposure rate (Georgiadou, Triantafillou, & Economides, 2007; Revuelta & Ponsoda, 1998).

RT_PWKL Method To avoid deterministic item selection using other ISSs, Wang et al. (2011) suggested selecting an item among those lead to the largest PWKLs,

instead of selecting the one item with the largest PWKL. In other words, an item will be randomly picked from the item set that contains all items that fall in the information interval defined as follows

$$[\max(PWKL_h) - \delta, \max(PWKL_h)], \quad h \in R^{(t)}, \tag{15.37}$$

where δ is a constant that regulates the size of the interval, or the size of the item set. Wang et al. (2011) suggested that δ should be larger in the beginning and smaller towards the end of the test. This makes the RT_PWKL method progressive, similar to the P_PWKL method. By adding the restrictive limit of maximum exposure rate, the RT_PWKL is similar to the RP_PWKL method.

Apparently some ISSs such as the P_PWKL and RP_PWKL methods were "developed from information indices in information science and attempted to achieve a balance among several objectives by assigning different weights" (Zheng & Wang, 2017). It is tricky how the weights should be assigned and often researchers must use "trial and error" to find appropriate weights. Zheng and Wang (2017) therefore proposed to adapt the classic binary searching algorithm to CD-CAT. The key idea of dynamic binary searching (DBS) is to select those items whose attribute vectors can split all possible attribute patterns into two mastery/non-mastery groups of equal size (i.e., the same number of attribute patterns in each group), or two groups of size that are as close as possible.

Q discrimination-control method. An innovative approach to balance item bank usage in regular CAT was the *a*-stratified method (Chang & Ying, 1999). It should be noted that under the MFI method items that are highly discriminating are favored and used most frequently. On the other hand, items with low discrimination parameters are seldom selected for administration. Chang and Ying (1999) therefore proposed to pre-stratify the item bank according to the *a*-parameter values into high-, medium- and low-discrimination stratum. Initially items are restricted to be selected from the low-discrimination stratum. Gradually as test progresses items can be chosen from higher strata.

The **Q** discrimination-control method (Wang et al., 2012) assumes that items with high "noise" parameters such as the guessing (g_h) and slipping (s_h) parameters in the deterministic input, noisy "and" gate (DINA; Junker & Sijtsma, 2001) model are low-quality items. They therefore proposed the MIQD index in contrast to the MIQ method by including the component$(1 - s_h)(1 - g_h)$ as follows:

$$P_h = (1 - s_h)(1 - g_h) \prod_{d=1}^{D} \left[\frac{u_d - x_d - q_{hd}}{u_d} \right] \left[\frac{(L - l_d) - (t - x_d - q_{hd})}{L - l_d} \right], \tag{15.38}$$

$$MIQD_h = I_h\left(\widehat{\theta}_n^{(t)}\right) P_h. \tag{15.39}$$

Items with larger $(1 - s_h)(1 - g_h)$ would be favored under the MIQD method. This again is a dual-purpose ISS as information from both the IRT model and CDM is considered. It can also be viewed as a variation of the MFI method by adding attribute balancing and item quality control. Therefore, it can still suffer from unbalanced item exposure. The StraQD method applies the MIQD to pre-stratified item bank, to address the potential item exposure issue by using the MIQD. The StraQD method serves dual-purpose and has built in attribute balancing and exposure control.

15.4 Stopping Rules

Depending on the stopping rule, a CD-CAT program can be of fixed length or variable-length. If the fixed-length rule is applied, a test will end when a pre-specified number of items are administered to each test taker. There are different stopping rules for variable-length CD-CAT programs. For example, Tatsuoka (2002) utilized a stopping rule when the maximum posterior probability of the respondent belonging to an attribute master pattern reaches 0.8. Hsu, Wang, and Chen (2013) further addressed the stopping rule issue and proposed two stopping rules for variable-length CD-CAT. The first rule is that a test would be stopped when the largest posterior probability (over all possible latent states) is no smaller than a pre-specified value (e.g., 0.7). The second rule requires not only the largest posterior probability to be greater than or equal to a pre-specified value (such as 0.7), but also that the second largest posterior probability is less than or equal to a pre-specified value (such as 0.1).

15.5 Respondent Classification Methods (Scoring Rules)

In a CD-CAT program, ISS or the termination rule or both may rely on sequential update of $\widehat{\boldsymbol{\alpha}}_n^{(t)}$. Similar to estimation of θ, MLE, MAP and EAP are currently the prevailing methods for estimation of $\boldsymbol{\alpha}$. Under the DINA model, the MLE, MAP and EAP estimates for $\boldsymbol{\alpha}$ can be obtained as follows:

$$\widehat{\boldsymbol{\alpha}}_{ML}^{(t)} = \underset{c}{\text{argmax}} \left\{ L\left(\boldsymbol{x}_n^{(t)}; \boldsymbol{\alpha}_c\right) \right\}, \tag{15.40}$$

$$\widehat{\boldsymbol{\alpha}}_{MAP}^{(t)} = \underset{c}{\text{argmax}} \left\{ P\left(\boldsymbol{\alpha}_c | \boldsymbol{x}_n^{(t)}\right) \right\}, \tag{15.41}$$

where $c = 1, 2, \ldots 2^D$, $L\left(\boldsymbol{x}_n^{(t)} | \boldsymbol{\alpha}_c\right)$ is the likelihood of responses on the first t items given an latent state $\boldsymbol{\alpha}_c$, and $P\left(\boldsymbol{\alpha}_c | \boldsymbol{x}_n^{(t)}\right)$ is the posterior probability of $\boldsymbol{\alpha}_c$ given $\boldsymbol{x}_n^{(t)}$.

The EAP method estimates the latent class by estimating the mastery status on each attribute:

$$\widehat{\alpha}_{EAP,d}^{(t)} = \sum_{c=1}^{2^D} P\left(\boldsymbol{\alpha}_c | \boldsymbol{x}_n^{(t)}\right) \mathbf{1}\left(\alpha_{cd} = 1\right), \tag{15.42}$$

where $\mathbf{1}(\cdot)$ is an indicator function, $\widehat{\alpha}_{EAP,d}^{(t)}$ is the mastery probability for the d^{th} attribute after administered t items. The final latent class estimate $\widehat{\boldsymbol{\alpha}}_{EAP}^{(t)}$ can be obtained by rounding each $\widehat{\alpha}_{EAP,d}^{(t)}$ to 0 or 1.

Huebner and Wang (2011) compared these three methods and examined the agreement among them on the classification of a given respondent. Based on their results, the performance of MLE and MAP is very similar. For total individual attribute classifications, EAP is superior to MLE and MAP. For tests of low diagnosticity (e.g., with large slipping and guessing parameters), there can be large discrepancies between EAP and MLE/MAP.

15.6 Online Calibration Under CAT

To maintain a large item bank, online calibration of new items is almost a necessity for large adaptive testing programs. Online calibration refers to estimating the item parameters of new items when they are administered to examinees, typically while they are administered in conjunction with previously calibrated, operational items during testing (Wainer & Mislevy, 2000). Then the calibrated new items can be added to the item bank to replenish the pool. Meanwhile, overexposed or obsolete items can be retired from the bank.

As discussed in Zheng (2014), it is possible to select an optimal sample of examinees to more efficiently calibrate item parameters during CAT. This is a cost-effective approach to maintain and replenish the item bank. It is also useful for the purposes of recalibrating existing items in the bank whose parameters may have drifted. Under a regular CAT, Stocking (1988) introduced two methods for online calibration: The Stocking-A method and the Stocking-B method. The Stocking-A method estimates examinee ability θs based on all the administered operational items first, and then estimate the parameters of new items by means of conditional maximum likelihood estimate (CMLE; Baker & Kim, 2004) assuming the estimated θs are fixed. The Stocking-B method extended the Stocking-A method by plugging in an equating step to correct the scale drift. Since then, numerous methods have been proposed and below we provide a brief review of several popular methods.

The OEM Method To calibrate the new items, Wainer and Mislevy (2000) used all the administered operational items to estimate the posterior θ distribution, and then marginalized the likelihood function with respect to the posterior distribution of θ based on the MMLE/EM algorithm with one EM cycle. The E-step is used to find

the posterior expectation of the log-likelihood of those new items, and the M-step is used to find the item parameter vector that maximizes the posterior expectation of the log-likelihood.

The MEM Method Different from the OEM, Ban, Hanson, Wang, and Harris (2001) increased the number of EM cycles until the predefined convergence condition was met and named this method Multiple EM cycles method (MEM). The first cycle of MEM is the same as the OEM. Starting from the second cycle, the posterior θ distribution will be updated based on both operational and pretest items, and then the pretest item parameters are updated. The E-step and the M-step then iterate until the predefined convergence condition is met.

The Bayesian Version of Stocking-A, OEM and MEM Methods To alleviate the non-convergence problems caused by small sample sizes and "bad" starting values for the Newton-Raphson algorithm, Zheng (2014) proposed Bayesian versions of the algorithms, that is, Bayesian Stocking-A, Bayesian OEM, and Bayesian MEM, by adding priors of item parameters to the likelihood function.

In addition to the calibration of item parameters (for example, slipping parameter and guessing parameter under the DINA model), the item attribute vector (which defines which attributes are measured by an item) also need to be specified. The **Q** matrix can be considered a compilation of item attribute vectors. Because of this, additional challenges may exist in online calibration under CD-CAT. So far there are two types of online calibration methods under CD-CAT. The first type, akin to its counterpart in regular CAT, estimates only the item parameters. The second type, on the other hand, estimates both the item parameters and the item attribute vector.

The Chen et al. (2012) method belongs to the first type. The paper extended the Stocking-A, OEM and MEM methods to CD-CAT, and denoted them as CD-Method A, CD-OEM and CD-MEM, respectively. The CD-Method A first estimates examinees' latent state based on the operational items (whose item parameters are known), and then calibrates the new items. It only employs a single EM cycle for calibration, the same as in the CD-OEM. In contrast to the CD-Method A, CD-OEM method does not fix estimated latent state; instead, it employs the full posterior distribution of the latent state. The difference between the CD-MEM and the CD-OEM is that the former includes multiple EM cycles until the predefined convergence criterion is met, whereas the latter includes only one single EM cycle.

Chen, Liu, and Ying (2015) proposed two methods to estimate both the item parameters and the corresponding attribute vectors for new items under the DINA model. They assume that there exists a set of operational items that have been well calibrated. Then they treat the attribute vector of the new items as additional item-specific parameters and estimated them in conjunction with item parameters. The two proposed methods are: the single-item estimation (SIE) method and the simultaneous item estimation (SimIE) method. The SIE method calibrates new items individually, while SimIE calibrates multiple new items simultaneously. Among the $2^{D}-1$ possible attribute vectors, the SIE method picks the one resulting in the largest likelihood function, and then obtains the corresponding item parameters as well as

latent class estimates. The SimIE adopts SIE to calibrate each new item and treats the new item as an additional operational item. The latent states of test takers were then updated via maximum likelihood or Bayesian estimation. Repeat this procedure to calibrate the next new item until all the new items are calibrated.

Granted, online calibration is one important piece to item bank development and maintenance, but not the only piece. Right now there hasn't been much research in CD-CAT on other aspects of item bank development, for example, the required size of the bank in relation to test length, optimal design of the item bank and so on. Further research is needed in these areas.

In summary, we have covered the main components of CD-CAT: Starting rule, ISS, stopping rule, scoring rule, and item bank development and more specifically online calibration. As a variation of CAT, multi-stage testing (MST) has lately become increasingly popular. Next we will discuss the CD-MST in relation to CD-CAT.

15.7 CD-MST

In both CAT and MST, items are selected sequentially based on respondent's provisional ability or proficiency estimates, be it latent trait(s) or latent class(es). Different from CAT, which is a fully sequential testing model, MST is a group-sequential testing model. The advantages of MST manifest in the use of stages, modules or panels, which allows test developers to preassemble a set of items for selection. Constraints such as content balancing can be met in the preassembly process. This means constraints do not need to be met on the fly as in CAT. It allows test developers to play a more important role in the process rather than only relying on the adaptive algorithm. See Yan, von Davier, and Lewis (2014) for more details on the comparison between CAT and MST. For a CD-MST program, there are several stages, and in each stage, choices need to be made to determine which block of items is administered next. Item blocks are pre-assembled. The reader is referred to von Davier and Cheng (2014) for a detailed discussion of CD-MST.

15.8 Large Scale Implementation

Numerous studies have been found on the various aspects of CD-CAT, but very few studies address its implementation. Liu, You, Wang, Ding, and Chang (2013) first reported on a large-scale development and implementation of a web-based CD-CAT in China. They developed an on-line assessment system to combine CAT with CDM and provided cognitive diagnostic feedback to the respondents on the Level 2 English Achievement. Based on the test blueprint, researchers and content experts constructed the **Q** matrix, which has 8 attributes covered by 400 items. They used 3PLM as the IRT model, and the DINA model as the CDM. For the ISS, they used

the SHE method to select items from the item bank and used Maximum a Posterior (MAP) method to estimate respondents' latent state. Liu et al. then evaluated the consistency between the results from the CD-CAT system and those obtained from an academic achievement test to obtain evidence of convergent validity. Von Davier and Cheng (2014) discussed the implementation of CD-MST in the Programme for the International Assessment of Adult Competencies (PIAAC).

15.9 Discussion

In this chapter we provided a comprehensive overview of the main components of CD-CAT, as well as up-to-date summary of existing research on these components. Apparently, there exists a large amount of research on ISS but relatively less attention on other aspects of CD-CAT. It is important to recognize the gaps in existing research and hopefully future research will fill in the gaps.

Besides these main components, other research topics may also be highly relevant to CD-CAT, for example, research on the validation and estimation of the **Q** matrix. Item selection and scoring under CD-CAT certainly relies heavily on the accuracy of the **Q** matrix. Researchers have proposed many approaches to deal with **Q** matrix estimation and validation, such as exploratory methods (Xiang, 2013; Chung, 2014), validation methods (de la Torre, 2008; de la Torre & Chiu, 2016), and hybrid methods (Liu, Xu, & Ying, 2012). These methods have their own strengths and weaknesses. Exploratory methods do not need an "initial **Q** matrix" to start but have a lower probability to obtain the correct **Q** matrix. On the other hand, validation methods require an "initial **Q** matrix" which may contain a small number of misspecifications but are more efficient. In contrast, hybrid methods only require part of the "initial **Q** matrix".

Another closely related topic is test assembly, that is, how to assemble an optimal test based on CDM. As alluded to in the section of CD-MST, test assembly may play an important role in the assembly of testing modules or panels of MST. Another important line of research in formative assessment is based on multidimensional item response theory (MIRT; Reckase, 1985, 1997) models. MIRT assumes a multi-dimensional vector for latent traits, in contrast to latent classes as in CDM. Discussion of CAT based on MIRT, or M-CAT, is beyond the scope of this chapter, but they clearly have important and close connections to CD-CAT.

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Ban, J. C., Hanson, B. A., Wang, T. Y., & Harris, D. J. (2001). A comparative study of on-line pretest item–calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*(3), 191–212.

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika, 80*(1), 1–20.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213–229.

Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(*3*), 211–222.

Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika, 77*(2), 201–222.

Chen, Y. X., Liu, J. C., & Ying, Z. L. (2015). Online item calibration for Q-Matrix in CD-CAT. *Applied Psychological Measurement, 39*(1), 5–15.

Cheng, Y. (2008). *Computerized adaptive testing: New development and applications.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*(4), 619–632.

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverge: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*(6), 902–913.

Cheng, Y., & Chang, H. H. (2007a). *Dual information method in cognitive diagnostic computerized adaptive testing*. Paper presented at the the meeting of the National Council on Measurement in Education, Chicago, IL.

Cheng, Y., & Chang, H. H. (2007b). *The modified maximum global discrimination index method for cognitive diagnostic computerized adaptive testing.* In the Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*(2), 369–383.

Cheng, Y., Chang, H. H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*(6), 467–482.

Cheng, Y., & Keng, L. (2009). Computerized adaptive testing in criterion-referenced testing. In E. Smith & G. Stone (Eds.), *Applications of Rasch measurement in criterion-reference testing: Practive analysis to score reporting*. Maple Grove, MN: JAM Press.

Chung, M. T. (2014). *Estimating the Q-matrix for cognitive diangnosis models in a Bayesian framework.* Unpublished doctoral dissertation, Columbia University.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Dai, B. Y., Zhang, M. Q., & Li, G. M. (2016). Exploration of item selection in dual-purpose cognitive diagnostic computerized adaptive testing: Based on the RRUM. *Applied Psychological Measurement, 40*(8), 625–640.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-Matrix validation. *Psychometrika, 81*(2), 253–273.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713–734.

Embretson, S. E. (2001). *The second century of ability testing: Some predictions and speculations*. Retrieved from http://www.ets.org/Media/Research/pdf/PICANG7.pdf

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing. *Journal of Technology, Learning, and Assessment, 5*(8), 4–38.

Gierl, M. J., & Zhou, J. W. (2008). Computer adaptive-attribute testing. *Journal of Psychology, 216*(1), 29–39.

He, W., & Reckase, M. D. (2014). Item pool design for an operational variable-length computerized adaptive test. *Educational and Psychological Measurement, 74*(3), 473–494.

Henson, R. A., & Douglas, J. (2005). Test construction for cognitive diagnostics. *Applied Psychological Measurement, 29*(4), 262–277.

Hsu, C. L., Wang, W. C., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement, 37*(7), 563–582.

Huebner, A. (2010). An overview of recent development in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation, 15*(3), 1–7.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*(2), 407–419.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kang, H. A., Zhang, S. S., & Chang, H. H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 54*(2), 165–183.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement, 39*(3), 167–188.

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for english achievement test in China. *Journal of Classification, 30*, 152–172.

Liu, J. C., Xu, G. J., & Ying, Z. L. (2012). Data driven learning of Q matrix. *Applied Psychological Measurement, 36*(7), 548–564.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

McGlohen, M., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40*(3), 808–821.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401–412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25–36.

Reckase, M. D. (2010). Designing item pools to optimized the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling, 52*(2), 127–141.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*(4), 311–327.

Riley, B. B., Conrad, K. J., Bezruczko, N., & Dennis, M. L. (2007). Relative precision, efficiency and construct validity of different starting and stopping rules for a computerized adaptive test: The GAIN substance problem scale. *Journal of Applied Measurement, 8*(1), 48–64.

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods and application*. New York: Guilford.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.

Stocking, M. L. (1988). *Scale drift in on-line calibration (RR-88-28-ONR)*. Princeton, NJ: Educational Testing Service.

Tatsuoka, C. (2002). Data analytic methods for latent parially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 51*(3), 337–350.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithm. In H. Wainer & N. J. Dorans (Eds.), *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbarm.

van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika, 63*(2), 201–216.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer.

van der Linden, W. J., & Chang, H. H. (2003). Implementing content constraints in a stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27(2)*, 107–120.

van der Linden, W. J., & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13*(1), 35–53.

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 149–162). Dordrecht, The Netherlands: Kluwer Academic Publishers.

von Davier, M., & Cheng, Y. (2014). Multistage testing using diagnostic models. In D. L. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 219–227). New York, NY: CRC Press.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration, and proficiency estimation. In H. Wainer & N. J. Dorans (Eds.), *Computerized adaptive testing: A primer* (2nd ed., pp. 61–100). Hillsdale, NJ: Lawrence Erlbaum.

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035.

Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods, 44*(1), 95–109.

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic CAT. *Journal of Educational Measurement, 48*(3), 255–273.

Wang, C., Zheng, C. J., & Chang, H. H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement, 51*(4), 358–380.

Wang, S. Y., Lin, H. Y., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement, 53*(1), 45–62.

Xiang, R. (2013). *Nonlinear penalized estimation of true Q-Matrix in cognitive diagnostic models.* Unpublished doctoral dissertation, Columbia University.

Xu, X. L., Chang, H. H., & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the the Annual Meeting of American Educational Research Association, Chicago, IL.

Yan, D. L., Lewis, C., & von Davier, A. A. (2014). Overview of computerized multistage tests. In D. L. Yan, A. A. von Davier, & C. Lewis (Eds.), Computerized multistage testing: Theory and applicatioins (pp. 3–20): CRC Press Boca Raton, FL.

Yan, D. L., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. CRC Press: Boca Raton, FL.

Yi, Q., & Chang, H. H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology, 56*(2), 359–378.

Zheng, C. J., & Chang, H. H. (2016). High-efficiency response distribution-based item selection algorithms for short-length cognitive diagnostc computerized adaptive testing. *Applied Psychological Measurement, 40*(6), 608–624.

Zheng, C. J., & Wang, C. (2017). Application of binary searching for item exposure control in cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement, 41*(7), 561–576.

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment*. Unpublished doctoral dissertation. University of Illinois at Urbana-Champaign.

Chapter 16
Identifiability and Cognitive Diagnosis Models

Gongjun Xu

Abstract Cognitive Diagnosis Models (CDMs) are popular statistical tools in cognitive diagnosis assessment. CDMs can be viewed as restricted latent class models with constraints introduced by the Q-matrix and assumptions of how skill variables that are assigned to items via the Q-Matrix interact in the item function. As many other latent variable models do, the CDMs often suffer from nonidentifiability. This chapter focuses on the identifiability issue of the CDMs and present conditions to ensure identifiability, which can be directly applied to most of the CDMs.

16.1 Introduction

Cognitive diagnosis is a type of assessment or measurement aiming to achieve a fine-grained description of an individual's latent traits, such as skills, knowledge, personality traits, or psychological disorders, based on his or her observed responses to certain diagnostic items. Compared with traditional tests for measuring proficiency that is usually characterized as a unidimensional latent trait, cognitive diagnosis focuses on detecting the presence or absence of multiple fine-grained latent traits, which are usually called *attributes*. Therefore, cognitive diagnosis assessment would provide more informative diagnostic profiles on each individual's attribute profile, such as school students mastery of the necessary component skills of mathematics. This feedback information allows for the design of more effective intervention strategies for remedy, such as to improve those latent attributes that a student has not sufficiently mastered yet.

Cognitive diagnosis models (CDMs), also called the diagnostic classification models (DCMs) in the literature, are statistical and psychometric tools in cognitive diagnosis assessment with the aim to estimate individuals' diagnostic attribute profiles from the response data of the assessment. Specifically, CDMs model

G. Xu (✉)
Department of Statistics, University of Michigan, Ann Arbor, MI, USA
e-mail: gongjun@umich.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_16

the complex relationship among items, multivariate binary latent trait vector, and categorical item responses for a set of items and a sample of respondents. Even though the earliest cognitively diagnostic models were proposed in 1980s, the topic of cognitive diagnosis modeling has gained great popularity in recent years due to the advancement of computation power to handle complex models and also due to the models' desirable diagnostic nature of providing informative cognitive profiles for every respondent. Various CDMs have been developed with different cognitive diagnosis assumptions, including the deterministic input noisy "and" gate (DINA) model and the noisy input deterministic "and" gate (NIDA) model (Junker & Sijtsma, 2001), the deterministic input noisy "or" gate (DINO) and noisy input deterministic "or" gate (NIDO) models (Templin & Henson, 2006), the reparameterized unified models (RUM; DiBello, Stout, & Roussos, 1995), the higher-order DINA model (de la Torre & Douglas, 2004), the general diagnostic model (GDM; von Davier, 2008), the loglinear CDM (LCDM; Henson, Templin, & Willse, 2009), and the generalized DINA model (GDINA; de la Torre, 2011), among others.

To achieve reliable and valid diagnostic assessment, a fundamental issue is to ensure that the CDMs applied in cognitive diagnosis are statistically identifiable, which is a necessity for statistically consistent estimation of the model parameters of interest and correct statistical inferences. The study of model identifiability has been an important topic in statistics and psychometrics, which dates back to Koopmans (1950) and Koopmans and Reiersøl (1950); see also McHugh (1956), Rothenberg (1971), Goodman (1974) and Gabrielsen (1978) for further developments. Identifiability issues of the CDMs have long been a concern, as noted in the literature (DiBello et al., 1995; Maris & Bechger, 2009; Tatsuoka, 2009a; DeCarlo, 2011; von Davier, 2014a). In practice, due to a lack of theoretical development on easy checkable identifiability conditions, there is often a tendency to overlook the issue, in part because available software tools tend not to provide checks of identifiability for applied research. Recently there have been several studies in the literature on the identifiability of the CDMs, including the DINA and DINO models (e.g., Liu, Xu, & Ying, 2013; Chen, Liu, Xu, & Ying, 2015; Xu & Zhang, 2016; Gu & Xu, 2018a,b) and general CDMs (e.g., Xu, 2017; Xu and Shang, 2018; Gu and Xu, 2018a).

This chapter presents practically checkable identifiability conditions with a selected review of recent developments and provides various examples for further illustration. It also aims to clarify some related concepts on the identifiability and estimability of the CDM parameters. For most CDMs, identifiability conditions can be characterized by the structure of the Q-matrix, a key component in cognitive diagnosis that describes the relationships between the items and the target latent attributes. A direct application of such results is that it would provide a guideline for designing statistically valid diagnostic tests. For instance, to ensure the model identifiability and consistent estimation, practitioners only need to construct Q-matrices that satisfy the identifiability structures when designing the diagnostic tests.

The rest of the chapter is organized as follows. In the following section, we give a review of several popularly used CDMs under the general framework of

the restricted latent class models. In what follows we introduce the identifiability definition and clarify some important concepts, such as the relationship between identifiability and consistent estimation; we also present an equivalent definition to check the identifiability of the model parameters, which is used to established the identifiability results. First, we consider two basic CDMs – the DINA and DINO models; due to their duality, we focus on the DINA model and present the sufficient and necessary conditions for identifying the slipping, guessing and population proportion parameters. Subsequently, we discuss the identifiability of general CDMs under the restricted latent class framework. Various examples, including a real data set, are given to illustrate the importance of the identifiability issue and how to check the proposed conditions. Some other interesting problems are further discussed in the closing section.

16.2 A Review of CDMs as Restricted Latent Class Models

We consider the setting of a cognitive diagnosis test with binary responses. The test contains J diagnosis items and a subject (such as an examinee or a patient) provides a J-dimensional binary response vector $\mathbf{X} = (X_1, \ldots, X_J)^\top$ to the items, where the superscript "$\top$" denotes the transpose operator. These responses are assumed to be dependent in a certain way on K unobserved latent attributes. Moreover, conditional on the K latent attributes, the responses are assumed to be independent, which is called the local independence assumption and is commonly used in the literature of CDMs and item response theory (e.g., Reckase, 2009; Rupp, Templin, & Henson, 2010; van der Linden & Hambleton, 2013; Embretson & Reise, 2013).

A complete set of the K latent attributes is known as an attribute profile, which is denoted by a column vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$, where $\alpha_k \in \{0, 1\}$ is assumed to be binary to indicate the absence or presence, respectively, of the kth attribute. Both $\boldsymbol{\alpha}$ and $\mathbf{X}$ are subject-specific; a particular subject i's attribute and response vectors are denoted by $\boldsymbol{\alpha}_i$ and $\mathbf{X}_i$, respectively, for $i = 1, \ldots, N$, where N denotes the number of subjects, i.e., the sample size.

Most CDMs assume the following two-step data generating process. The first step models the attribute profile $\boldsymbol{\alpha}$ from a population distribution. A common assumption is that the subjects' attribute profiles are a random sample of size N from a designated population so that their attribute profiles $\boldsymbol{\alpha}_i$, $i = 1, \ldots, N$ are random variables following a categorical distribution with probabilities

$$p_{\boldsymbol{\alpha}} := P(\boldsymbol{\alpha}_i = \boldsymbol{\alpha}),$$

where $p_{\boldsymbol{\alpha}} \in (0, 1)$, for any $\boldsymbol{\alpha} \in \{0, 1\}^K$, and $\sum_{\boldsymbol{\alpha} \in \{0,1\}^K} p_{\boldsymbol{\alpha}} = 1$. The distribution of $\boldsymbol{\alpha}$ is thus characterized by the column vector $\mathbf{p} = (p_{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in \{0, 1\}^K)^\top$.

The second step of the CDMs follows a restricted latent class model setting with incorporated constraints according the cognitive processes. Given a subject's

attribute profile $\boldsymbol{\alpha}$, the response X_j to item j under the corresponding model follows a Bernoulli distribution: $X_j \mid \boldsymbol{\alpha} \sim \text{Bernoulli}(\theta_{j,\boldsymbol{\alpha}})$, where

$$\theta_{j,\boldsymbol{\alpha}} := P(X_j = 1 \mid \boldsymbol{\alpha})$$

is the probability of providing positive response to item j for subjects with $\boldsymbol{\alpha}$. CDMs can be viewed as restricted latent class models where the parameters $\Theta = (\theta_{j,\boldsymbol{\alpha}})_{J\times 2^K}$ are constrained by the relationship between the J items and the K latent traits. Such relationship is specified through a Q-matrix (Tatsuoka, 2009b), which is defined as a $J \times K$ binary matrix with entries $q_{jk} \in \{0, 1\}$ indicating the absence or presence, respectively, of a link between the jth item and the kth attribute. The jth row vector, denoted by $\mathbf{q}_j$ of the Q-matrix correspond to the full attribute requirements of each item. For instance, we consider the following 3×2 Q-matrix, which gives the corresponding item and attribute relationships.

$Q =$

	Attribute α_1 (addition)	Attribute α_2 (multiplication)
Item1 :$1 + 3$	1	0
Item2 :4×2	0	1
Item3 :$(1 + 3) \times 2$	1	1

To define the Q-introduced constraints, we need some notation for vector ordering: given an attribute profile $\boldsymbol{\alpha}$ and the jth item's Q-matrix vector $\mathbf{q}_j$, we write $\boldsymbol{\alpha} \succeq \mathbf{q}_j$ if $\alpha_k \geq q_{jk}$ for any $k \in \{1, \ldots, K\}$, and $\boldsymbol{\alpha} \not\succeq \mathbf{q}_j$ if there exists k such that $\alpha_k < q_{jk}$; similarly we define the operations $\preceq$ and $\not\preceq$.

For most CDMs, a common assumption is that mastering non-required attributes will not change the response probability; if $\boldsymbol{\alpha} \succeq \mathbf{q}_j$, then a subject with $\boldsymbol{\alpha}$ has all the attributes for item j specified by the Q-matrix and would be most "capable" to provide a positive answer. On the other hand, if $\boldsymbol{\alpha}' \not\succeq \mathbf{q}_j$, the subject with $\boldsymbol{\alpha}'$ lacks some required attribute and is not expected to have a higher positive response probability than $\boldsymbol{\alpha} \succeq \mathbf{q}_j$. In addition, subjects without mastery of any latent traits ($\boldsymbol{\alpha} = \mathbf{0}$) are expected to have the lowest positive response probability. Such constraints on Θ are proposed through the following monotonicity relations (Xu, 2017):

$$\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j} \theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j} \theta_{j,\boldsymbol{\alpha}} \geq \theta_{j,\boldsymbol{\alpha}'} \geq \theta_{j,\mathbf{0}}, \quad \text{for any } \boldsymbol{\alpha}'; \tag{16.1}$$

in addition, for any $k \in \{1, \cdots, K\}$ and any item j such that it only requires the kth attribute, i.e., $\mathbf{q}_j = \mathbf{e}_k$, we assume

$$\theta_{j,\mathbf{1}} > \max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\not\succeq\mathbf{e}_k} \theta_{j,\boldsymbol{\alpha}} \quad \text{for } \mathbf{q}_j = \mathbf{e}_k. \tag{16.2}$$

The key assumption in (16.1) and (16.2) is $\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j}\theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j}\theta_{j,\boldsymbol{\alpha}}$, which is equivalent to

$$\theta_{j,\boldsymbol{\alpha}} = \theta_{j,\boldsymbol{\alpha}'}, \quad \text{for any } \boldsymbol{\alpha}\succeq\mathbf{q}_j \text{ and } \boldsymbol{\alpha}'\succeq\mathbf{q}_j.$$

The other assumptions in (16.1) and (16.2) are monotonicity requirements to avoid label-switching issue that occurs in all unrestricted latent class models. The requirements in (16.1) and (16.2) are satisfied by most of the CDMs, including the DINA, DINO, Reduced-RUM, LCDM, GDINA, and GDM. The identifiability results in the following section are established under this general framework in the sense that for any CDM, as long as (16.1) and (16.2) are satisfied, the theoretical results can be applied. To help with further discussions, we use the following examples to give a brief review of some of the popularly used CDMs.

Example 1 (DINA model) One of the basic cognitive diagnosis model is the deterministic input noisy "and" gate (DINA) model (Junker & Sijtsma, 2001), which assumes a conjunctive relationship among attributes. That is, it is necessary to possess all the attributes indicated by the Q-matrix to be capable of providing a positive response. In addition, having additional unnecessary attributes does not compensate for the lack of necessary attributes. For item j and attribute vector $\boldsymbol{\alpha}$, we define the ideal response $\xi_{j,\boldsymbol{\alpha}}^{DINA} = I(\boldsymbol{\alpha}\succeq\mathbf{q}_j)$. The uncertainty is further incorporated at the item level, using the slipping and guessing parameters $\mathbf{s}$ and $\mathbf{g}$. For each item j, the slipping parameter

$$s_j = P(X_j = 0 \mid \xi_{j,\boldsymbol{\alpha}}^{DINA} = 1)$$

denotes the probability of an incorrect response despite mastering all necessary skills; similarly, the guessing parameter

$$g_j = P(X_j = 1 \mid \xi_{j,\boldsymbol{\alpha}}^{DINA} = 0)$$

denotes the probability of a positive response despite an incorrect ideal response. The response probability $\theta_{j,\boldsymbol{\alpha}}$ then takes the form

$$\theta_{j,\boldsymbol{\alpha}} = (1-s_j)^{\xi_{j,\boldsymbol{\alpha}}^{DINA}} g_j^{1-\xi_{j,\boldsymbol{\alpha}}^{DINA}}.$$

It is usually assumed that $1 - s_j > g_j$ for any item j, which implies (16.1) and (16.2).

Example 2 (DINO model) In contrast to the DINA model, the DINO model assumes a disjunctive relationship among attributes, that is, one only needs to have one of the required attributes to be capable of providing a positive response. The ideal response of the DINO model is given by $\xi_{j,\boldsymbol{\alpha}}^{DINO} = I(\alpha_k \geq q_{jk}$ for at least one $k)$. Similar to the DINA model, there are two parameters s and g for each item, and

$$\theta_{j,\boldsymbol{\alpha}} = (1-s_j)^{\xi_{j,\boldsymbol{\alpha}}^{DINO}} g_j^{1-\xi_{j,\boldsymbol{\alpha}}^{DINO}}.$$

Again, assumptions (16.1) and (16.2) are satisfied if $1 - s_j > g_j$ for any j.

For the DINA or DINO model, if some item j does not require any of the attributes, then the guessing parameter of this item is not needed in the model specification. Without loss of generality, in the following discussion we define the guessing parameter of any item with $\mathbf{q}_j = \mathbf{0}$ to be a known value $g_j \equiv 0$.

Example 3 (Reduced-RUM model) Under the reduced version of the Reparameterized Unified Model (R-RUM; see DiBello et al., 1995; Rupp et al., 2010), we have

$$\theta_{j,\boldsymbol{\alpha}} = \pi_j \prod_{k=1}^{K} r_{j,k}{}^{q_{jk}(1-\alpha_k)}, \tag{16.3}$$

where π_j is the positive response probability for subjects who possess all required attributes and $r_{j,k}$, $0 < r_{j,k} < 1$, is the penalty parameter for not possessing the kth attribute. Note that the model is equivalent to the log-link model

$$\log \theta_{j,\boldsymbol{\alpha}} = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk}(q_{jk}\alpha_k).$$

For the reduced RUM in (16.3), it is easy to see that assumptions (16.1) and (16.2) are satisfied by the definition.

Example 4 (GDINA model) de la Torre (2011) generalizes the DINA model to the GDINA model. The formulation of the GDINA model based on $\theta_{j,\boldsymbol{\alpha}}$ can be decomposed into the sum of the effects due the presence of specific attributes and their interactions. Specifically,

$$\theta_{j,\boldsymbol{\alpha}} = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk}(q_{jk}\alpha_k) + \sum_{k'=k+1}^{K} \sum_{k=1}^{K-1} \beta_{jkk'}(q_{jk}\alpha_k)(q_{jk'}\alpha_{k'}) + \cdots + \beta_{j12\cdots K} \prod_{k} q_{jk}\alpha_k.$$

Note that for any $1 \le h \le K$ and any $1 \le k_1 < \cdots < k_h \le K$, if $\prod_{l=1}^{h} q_{j,k_l} = 0$, then $\beta_{j,k_1\cdots k_h}$ is not needed in the model and can be set as 0. For instance, when $\mathbf{q}_j \neq \mathbf{1}^\top$, we do not need parameter $\beta_{j12\cdots K}$ since $\prod_k (q_{jk}\alpha_k) = 0$. To interpret the model parameters, β_{j0} represents the probability of a positive response when none of the required attributes is present; when $q_{jk} = 1$, β_{jk} is included in the model and it shows the change in the positive response probability as a result of mastering a single attribute α_k; when $q_{jk} = q_{jk'} = 1$, $\beta_{jkk'}$ is in the model and it shows the change in the positive response probability due to the interaction effect of mastery of both α_k and $\alpha_{k'}$; when $\mathbf{q}_j = \mathbf{1}^\top$, $\beta_{j12\cdots K}$ represents the change in the positive

response probability due to the interaction effect of mastery of all the required attributes. Note that the assumption in (16.1), $\max_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j}\theta_{j,\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}:\boldsymbol{\alpha}\succeq\mathbf{q}_j}\theta_{j,\boldsymbol{\alpha}}$, is satisfied from the model definition.

Under the local independence assumption, the likelihood function of a subject's observed responses ($\mathbf{X}$) is

$$f(\mathbf{p}, \Theta; \mathbf{X}) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \boldsymbol{\theta}_{j,\boldsymbol{\alpha}}^{X_j}(1-\boldsymbol{\theta}_{j,\boldsymbol{\alpha}})^{1-X_j}.$$

For subjects $1, \cdots, N$, the joint likelihood is

$$f(\mathbf{p}, \Theta; \mathbf{X}_1, \cdots, \mathbf{X}_N) = \prod_{i=1}^{N} \sum_{\boldsymbol{\alpha}_i=\boldsymbol{\alpha}\in\{0,1\}^K} p_{\boldsymbol{\alpha}} \prod_{j=1}^{J} \boldsymbol{\theta}_{j,\boldsymbol{\alpha}}^{X_{i,j}}(1-\boldsymbol{\theta}_{j,\boldsymbol{\alpha}})^{1-X_{i,j}}.$$

Note that we consider the attribute profiles ($\boldsymbol{\alpha}$'s) as random effects and the above likelihood integrated out $\boldsymbol{\alpha}$'s through their population distribution that is characterized by $\mathbf{p}$. The likelihood function plays a central role in the identifiability research as well as statistical inference problems.

Remark 1 (Random effects CDMs vs. Fixed effects CDMs) There are two types of models for CDMs in terms of the interpretation of the attribute profiles. This work considers the attribute profile $\boldsymbol{\alpha}$ as random effects and further models the population distribution of $\boldsymbol{\alpha}$. On the other hand, we may consider the attribute profile $\boldsymbol{\alpha}$ as fixed effects, in which case the $\boldsymbol{\alpha}'s$ are considered as model parameters and the population distribution of the attribute profiles as well as the $\mathbf{p}$ parameters are no longer needed under the fixed effects CDMs. The likelihood of the fixed effects model for a subject with attribute profile $\boldsymbol{\alpha}$ can be written as

$$f(\boldsymbol{\alpha}, \boldsymbol{\theta}; \mathbf{X}) = \prod_{j=1}^{J} \boldsymbol{\theta}_{j,\boldsymbol{\alpha}}^{X_j}(1-\boldsymbol{\theta}_{j,\boldsymbol{\alpha}})^{1-X_j}.$$

For subjects $1, \cdots, N$ with attribute profiles $\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_N$, the joint likelihood of the fixed effects model is

$$f(\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_N, \boldsymbol{\theta}; \mathbf{X}_1, \cdots, \mathbf{X}_N) = \prod_{i=1}^{N} \prod_{j=1}^{J} \boldsymbol{\theta}_{j,\boldsymbol{\alpha}_i}^{X_{i,j}}(1-\boldsymbol{\theta}_{j,\boldsymbol{\alpha}_i})^{1-X_{i,j}}.$$

For the fixed effects CDMs, however, it is known that the fixed effect model parameters, including the item parameters and the fixed effects $\boldsymbol{\alpha}_i$'s, may not be consistently estimated even the sample size N goes to infinity. See Remark 2 for more details.

In this work, we focus on the by far overwhelmingly used random effects CDMs and consider the identifiability of item parameters and population mixture proportion parameters $\mathbf{p}$. We give identifiability definitions, identifiability conditions, and various examples in the following sections.

16.3 Identifiability and Related Concepts

16.3.1 Identifiability Definition

Following the statistics literature (e.g., Casella & Berger, 2002), we say a set of parameters β in the parameter space B for a family of distribution functions $\{f(\cdot|\beta) : \beta \in B\}$ is identifiable if distinct values of β correspond to distinct probability density (mass) functions, i.e., for any β there is no $\tilde{\beta} \in B \backslash \{\beta\}$ for which

$$f(\cdot|\beta) \equiv f(\cdot|\tilde{\beta}). \tag{16.4}$$

In addition, we say that a set of parameters β is locally identifiable if there exists a neighborhood of β, $\mathcal{N}_\beta \in B$, such that there is no $\tilde{\beta} \in \mathcal{N}_\beta \backslash \{\beta\}$ such that $f(\cdot|\beta) \equiv f(\cdot|\tilde{\beta})$.

Both identifiability and local identifiability of latent class models are well studied concepts in latent class analysis (McHugh, 1956; Goodman, 1974). Developments in the item response theory models can be found in Bechger, Verstralen, and Verhelst (2002), Maris and Bechger (2004), San Martín, Rolin, and Castro (2013) and others. Identifiability is an important prerequisite for many types of statistical inference, such as parameter estimation and hypothesis testing. Local identifiability is a weaker form of identifiability, which ensures that the model parameters are identifiable in a neighborhood of the true parameter values.

Remark 2 (Identifiable vs. Consistently Estimable) Identifiability is a prerequisite and necessary condition for the statistical consistency of an estimator. However, identifiability conditions are not always sufficient for consistent estimation. Here we say the parameter is consistently estimable if we can construct a consistent estimator for the parameter. That is, for parameter β, there exists $\hat{\beta}_N$ such that $\hat{\beta}_N - \beta \to 0$ in probability as the sample size $N \to \infty$.

An example of identifiable but not consistently estimable is the fixed effects CDMs, where the attribute profiles ($\boldsymbol{\alpha}$'s) are taken as parameters. Consider a simple example of the DINA model with nonzero slipping and guessing parameters. Under the fixed effects setting, the model parameters include $\boldsymbol{\alpha}_i, i = 1, \cdots, N$. In this case, $\boldsymbol{\alpha}$'s are identifiable if the Q-matrix has an identity submatrix (Chiu, Douglas, & Li, 2009). But with fixed number of items, even when the sample size N goes to infinity, the parameters cannot be consistently estimated. In this case, to have the consistent estimation of each $\boldsymbol{\alpha}$, the number of items needs to go to infinity and the number of identity sub-Q-matrices also needs to go to infinity (Wang & Douglas, 2015).

For the random effects CDMs as considered in this chapter, the identifiability conditions ensure consistent estimation of the model parameters. In particular, when the identifiability conditions are satisfied, the maximum likelihood estimators of the corresponding cognitive diagnosis model parameters are consistent as the sample size N increases. The result is applicable for all the CDMs under the restricted latent class model framework as introduced above.

Remark 3 (Model identifiability vs. partial identifiability) We say a model is identifiable if all parameters in the model are identifiable, and a model is partially identifiable if some but not all parameters are identifiable. Below, we present partial identifiability results of the CDMs parameters and illustrative examples, in addition to the model identifiability results.

Consider the CDMs under the restricted latent class model framework introduced in above. The model parameters can be equivalently represented as the parameter matrix $\Theta = (\theta_{j,\boldsymbol{\alpha}})_{J\times 2^K}$ and proportion parameter $\mathbf{p} = (p_{\boldsymbol{\alpha}})_{2^K\times 1}$, as shown in the examples, and the identifiability of $(\Theta, \mathbf{p})$ is equivalent to that of the CDM parameters. Without loss of generality, we focus on $(\Theta, \mathbf{p})$ in the following. Note the joint distribution of $\mathbf{X}$, conditional on the latent class $\boldsymbol{\alpha}$, is given by a J-dimensional $2 \times \cdots \times 2$ table

$$\mathbb{T}(Q, \Theta, \boldsymbol{\alpha}) = \bigotimes_{j=1}^{J} \begin{bmatrix} 1-\theta_{j,\boldsymbol{\alpha}} \\ \theta_{j,\boldsymbol{\alpha}} \end{bmatrix},$$

where $\otimes$ denotes the tensor product and the $\mathbf{x} = (x_1, \cdots, x_J)$-entry of the table $\mathbb{T}(Q, \Theta, \boldsymbol{\alpha})$ is $P(\mathbf{X} = \mathbf{x} \mid Q, \Theta, \boldsymbol{\alpha})$, i.e., the probability of observing $\mathbf{x}$ given $(Q, \Theta, \boldsymbol{\alpha})$. Following the above notation, we can write

$$P(\mathbf{X} = \mathbf{x} \mid Q, \Theta, \mathbf{p}) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^K} P(\mathbf{X} = \mathbf{x} \mid Q, \Theta, \boldsymbol{\alpha}) p_{\boldsymbol{\alpha}}.$$

We introduce the following identifiability definition.

Definition 1 We say that $(\Theta, \mathbf{p})$ is identifiable if for any $(\bar{\Theta}, \bar{\mathbf{p}}) \neq (\Theta, \mathbf{p})$, there exists at least one response pattern $\mathbf{x} \in \{0, 1\}^J$ such that

$$P(\mathbf{X} = \mathbf{x} \mid Q, \Theta, \mathbf{p}) \neq P(\mathbf{X} = \mathbf{x} \mid Q, \bar{\Theta}, \bar{\mathbf{p}}). \tag{16.5}$$

Definition 1 follows from the definition in (16.4) that different parameter values result in different probability distributions when the model is identifiable. Note that the above definition does not involve label swapping of the latent classes due to the fact that the labels of attributes are pre-specified from the knowledge of the Q-matrix and the monotonicity assumptions in (16.1) and (16.2). On the other hand, for unrestricted latent class models, the latent classes can be freely relabeled

without changing the distribution of the data and the model parameters are therefore identifiable only up to label swapping.

16.3.2 An Equivalent Definition of Identifiability

This subsection gives a brief introduction of some techniques to establish identifiability, which is based on an equivalent definition of identifiability.[1]

To establish (16.5) for the restricted latent class models, directly working with the vectors $P(\mathbf{X} = \mathbf{x} \mid Q, \Theta, \mathbf{p})$ is technically challenging. To better incorporate the induced restrictions by the Q-matrix, we consider the marginal response probability matrix as introduced in the following. The marginal response probability matrix is called the T-matrix, denoted by $T(Q, \Theta)$, which is defined as a $2^J \times 2^K$ matrix, where the entries are indexed by row index $\mathbf{x} \in \{0, 1\}^J$ and column index $\boldsymbol{\alpha}$. The $\mathbf{x} = (x_1, \cdots, x_J)$th row and $\boldsymbol{\alpha}$th column element of $T(Q, \Theta)$, denoted by $t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \Theta)$, is the marginal probability that a subject with attribute profile $\boldsymbol{\alpha}$ answers all items in subset $\{j : x_j = 1\}$ positively. Thus $t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \Theta)$ is the marginal probability that, given $Q, \Theta, \boldsymbol{\alpha}$, the random response $\mathbf{X} \succeq \mathbf{x}$, i.e.,

$$t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \Theta) = P(\mathbf{X} \succeq \mathbf{x} \mid Q, \Theta, \boldsymbol{\alpha}).$$

When $\mathbf{x} = \mathbf{0}$, $t_{\mathbf{0},\boldsymbol{\alpha}}(Q, \Theta) = P(\mathbf{X} \succeq \mathbf{0}) = 1$ for any $\boldsymbol{\alpha}$. When $\mathbf{x} = \mathbf{e}_j$, for $1 \leq j \leq J$,

$$t_{\mathbf{e}_j,\boldsymbol{\alpha}}(Q, \Theta) = P(X_j = 1 \mid Q, \Theta, \boldsymbol{\alpha}) = \theta_{j,\boldsymbol{\alpha}}.$$

Let $T_{\mathbf{x},\cdot}(Q, \Theta)$ be the row vector corresponding to $\mathbf{x}$. Then we know that for $j = 1, \cdots, J$, $T_{\mathbf{e}_j,\cdot}(Q, \Theta) = \Theta_{j,\cdot}$. In addition, for any $\mathbf{x} \neq \mathbf{0}$, we can write $T_{\mathbf{x},\cdot}(Q, \Theta) = \bigodot_{j:x_j=1} T_{\mathbf{e}_j,\cdot}(Q, \Theta)$, where $\odot$ is the element-wise product of the row vectors.

By definition, multiplying the T-matrix by the distribution of attribute profiles $\mathbf{p}$ results in a vector, $T(Q, \Theta)\mathbf{p}$, containing the marginal probabilities of successfully responding each subset of items correctly. The $\mathbf{x}$th entry of this vector is

$$\begin{aligned} T_{\mathbf{x},\cdot}(Q, \Theta)\mathbf{p} &= \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \Theta) p_{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \{0,1\}^K} P(\mathbf{X} \succeq \mathbf{x} \mid Q, \Theta, \boldsymbol{\alpha}) p_{\boldsymbol{\alpha}} \\ &= P(\mathbf{X} \succeq \mathbf{x} \mid Q, \Theta, \mathbf{p}). \end{aligned}$$

[1] For readers who are more interested in how to use the identifiability results in practice, this section can be skipped, as well as the discussion of Eqs. (16.12) and (16.13) and Remark 5, which are based on this section.

We can see that there is a one-to-one mapping between the T-matrix and the vectors $P(\mathbf{X} = \mathbf{x} \mid Q, \Theta, \mathbf{p})$, $\mathbf{x} \in \{0, 1\}^J$. Therefore, (16.5) directly implies the following proposition.

Proposition 1 (An equivalent definition of identifiability) $(\Theta, \mathbf{p})$ *is identifiable if and only if for any* $(\bar{\Theta}, \bar{\mathbf{p}}) \neq (\Theta, \mathbf{p})$*, there exists* $\mathbf{x} \in \{0, 1\}^J$ *such that*

$$T_{\mathbf{x},\cdot}(Q, \Theta)\mathbf{p} \neq T_{\mathbf{x},\cdot}(Q, \bar{\Theta})\bar{\mathbf{p}}. \tag{16.6}$$

From Proposition 1, to show the identifiability of $(\Theta, \mathbf{p})$, we only need to focus on the T-matrix and prove that if

$$T(Q, \Theta)\mathbf{p} = T(Q, \bar{\Theta})\bar{\mathbf{p}}, \tag{16.7}$$

then $\Theta = \bar{\Theta}$ and $\mathbf{p} = \bar{\mathbf{p}}$. This argument is used in the proofs of the identifiability results in Xu and Zhang (2016), Xu (2017), Xu and Shang (2018), and Gu and Xu (2018b).

16.4 Identifiability of the DINA and DINO Models

In this section we focus on the DINA model, a basic and popularly used CDM. Thanks to the duality of the DINA and DINO models (Chen et al., 2015; Xu & Zhang, 2016), the results can be directly applied to the DINO model. As introduced in Example 1, the model parameters under the DINA model include slipping parameters $\mathbf{s} = (s_1, \cdots, s_J)$, guessing parameters $\mathbf{g} = (g_1, \cdots, g_J)$, and population mixture proportion parameters $\mathbf{p} = (p_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \{0, 1\}^K)$. We assume that for each item, students mastering the required skills always have higher correct response probability than the students lacking one or more of the required skills, that is, $1 - s_j > g_j$, for $j = 1, \cdots, J$. In addition, we assume that the Q-matrix is pre-specified and correct. Under various model assumptions, we present the conditions for the identifiability of the unknown model parameters, most of which are based on the work of Xu and Zhang (2016) and Gu and Xu (2018b).

16.4.1 Identifiability Conditions When Both the Slipping and the Guessing Parameters Are Known

We first consider the ideal case when the jth item's response $X_j = \xi_{j,\boldsymbol{\alpha}}$, where $\xi_{j,\boldsymbol{\alpha}}$ denotes $\xi_{j,\boldsymbol{\alpha}}^{DINA}$ as defined in Example 1. In this ideal case, $\mathbf{s} = \mathbf{g} = \mathbf{0}$ and the only unknown parameters are $\mathbf{p}$. Note that $P(\mathbf{X} = \mathbf{e}_j \mid \boldsymbol{\alpha}, Q, \Theta) = \xi_{j,\boldsymbol{\alpha}}$ and the identifiability condition is equivalent to

$$(\xi_{j,\boldsymbol{\alpha}}; j = 1, \cdots, J) \neq (\xi_{j,\boldsymbol{\alpha}'}; j = 1, \cdots, J) \tag{16.8}$$

for all $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$. Otherwise, if there exists $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}'$ such that $(\xi_{j,\boldsymbol{\alpha}}; j = 1, \cdots, J) = (\xi_{j,\boldsymbol{\alpha}'}; j = 1, \cdots, J)$, this implies the nonidentifiability of $\mathbf{p}$ from the definition. To guarantee (16.8), the requirement on the Q-matrix structure is specified in the following definition.

Definition 2 A Q-matrix is said to be *complete* under the DINA model if $\{\mathbf{e}_j^\top : j = 1, \ldots, K\} \subset \{\mathbf{q}_j : j = 1, \cdots, J\}$; otherwise, we say that Q is *incomplete*.

Remark 4 The completeness concept was first introduced in Chiu et al. (2009) when studying the identification of an individual's attribute profile ($\boldsymbol{\alpha}$) in the fixed effects CDMs. The differences between the fixed effects CDMs and the random effects CDMs are discussed in Remarks 1 and 2. This work considers the random effects CDMs and focuses on the identifiability of item parameters and population parameters $\mathbf{p}$.

The Q-matrix is complete under the DINA model if there exist K rows of Q that can be ordered to form the K-dimensional identity matrix $\mathcal{I}_K$, that is, for each attribute there must exist a pure item requiring only that attribute. Note that even if the guessing and slipping parameters are known (which they are never in practice), the DINA (and equivalently the DINO) are not identifiable with respect to the attribute distribution unless there are pure items measuring each attribute separately in the test. Therefore, in most of the practical applications of the DINA, we would suffer from non-identifiability of the attribute distribution due to the incompleteness of the Q-matrix (e.g., see the analysis of the fraction subtraction data in DeCarlo, 2011).

A simple (and minimal) example of a complete Q-matrix under the DINA model is the $K \times K$ identity matrix $\mathcal{I}_K$. Completeness ensures that there is enough information in the response data for each attribute profile to have its own distinct ideal response vector. When a Q-matrix is incomplete, we can easily construct a non-identifiable example. For instance, consider the incomplete Q-matrix

$$Q = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}. \tag{16.9}$$

The population parameter $\mathbf{p}$ is non-identifiable. Subjects with attribute profiles $\boldsymbol{\alpha}_1 = (1, 0)^\top$ and $\boldsymbol{\alpha}_2 = (0, 0)^\top$ have the same ideal responses, so (16.8) is not satisfied. It is easy to see that such an argument holds for general incomplete Q-matrices.

We further consider the case in which both the slipping and the guessing parameters are known but may be nonzero. This corresponds to the applications where the item parameters are pre-calibrated. We need the following completeness condition for the identifiability of $\mathbf{p}$.

(C1) Q is complete. When this holds, we assume without loss of generality that the Q-matrix takes the following form:

$$Q = \begin{pmatrix} \mathcal{I}_K \\ Q' \end{pmatrix}. \tag{16.10}$$

Theorem 1 *Population proportion parameters* $\mathbf{p}$ *are identifiable only if Condition C1 is satisfied. Moreover, Condition C1 is sufficient and necessary when both the slipping and the guessing parameters are known.*

Theorem 1 states that when $\mathbf{s}$ and $\mathbf{g}$ are known, the completeness of the Q-matrix is a sufficient and necessary condition for the identifiability of $\mathbf{p}$. Similarly to the ideal case with $\mathbf{s} = \mathbf{g} = \mathbf{0}$, for the incomplete Q-matrix in (16.9), the population parameter $\mathbf{p}$ is non-identifiable even if $\mathbf{s}$ and $\mathbf{g}$ are known. Subjects with attribute profiles $\boldsymbol{\alpha}_1 = (1, 0)^\top$ and $\boldsymbol{\alpha}_2 = (0, 0)^\top$ still have the same conditional response probabilities $P(\mathbf{X} = \mathbf{x} \mid Q, \mathbf{s}, \mathbf{g}, \boldsymbol{\alpha})$, so weight can be transferred between $p_{\boldsymbol{\alpha}_1}$ and $p_{\boldsymbol{\alpha}_2}$ with no effect on the distribution probabilities $P(\mathbf{X} = \mathbf{x} \mid Q, \mathbf{s}, \mathbf{g}, \mathbf{p})$.

16.4.2 Identifiability Conditions When the Guessing Parameters Are Known

We now weaken our assumptions by taking only the guessing parameters $\mathbf{g}$ as known. Applications in this case may involve confirmatory type diagnosis analysis with the guessing parameters pre-determined due to the low possibility of correctly answering an item by guessing or certain multiple choice problem settings. Stronger conditions than Theorem 1 are needed for the identifiability of the unknown slipping parameters $\mathbf{s}$ and population proportion parameters $\mathbf{p}$.

(C1) Q is complete. When this holds, we assume without loss of generality that the Q-matrix takes the form in (16.10).
(C2) Each attribute is required by at least two items.

Condition C1 means the Q-matrix is complete, which is necessary to distinguish different latent attribute profiles. Condition C2 requires each attribute is needed by more than one item, which is necessary to identify the slipping parameters. The necessary and sufficient conditions for the identifiability of $\mathbf{s}$ and $\mathbf{p}$ are given in Theorem 2 below, which was proved in Xu and Zhang (2016).

Theorem 2 (Sufficient and Necessary Identifiability Conditions) *Under the DINA model with known guessing parameters* $\mathbf{g}$*, the slipping parameters* $\mathbf{s}$ *and the population proportion parameters* $\mathbf{p}$ *are identifiable if and only if Conditions C1 and C2 hold.*

Conditions C1 and C2 are easy to check. We use an example to illustrate.

Example 5 Consider the Q-matrices

$$Q_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad Q_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}. \tag{16.11}$$

From Theorem 2, we can see that when the guessing parameters $\mathbf{g}$ are known Q_1 describes a non-identifiable model while Q_2 describes an identifiable one.

The necessary of Condition C1 is shown in the previous section. To see why Condition C2 is necessary, consider any set of parameters $(\mathbf{s}, \mathbf{p})$ such that $1 - s_j \in (g_j, 1]$ for all $j \in \{1, \ldots, J\}$ and $\mathbf{p} \in (0, 1)^{2^K}$, $\sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} = 1$. In the following we construct a set of parameters $(\bar{\mathbf{s}}, \bar{\mathbf{p}})$ such that $(\bar{\mathbf{s}}, \bar{\mathbf{p}}) \neq (\mathbf{s}, \mathbf{p})$ and they are not identifiable if Condition C2 is not satisfied. We first choose $\bar{s}_1$ such that it is close enough to s_1 so that $1 - \bar{s}_1 \in (g_1, 1]$. We further choose $\bar{p}_{\boldsymbol{\alpha}} \in (0, 1)$ in the following way such that for any $\boldsymbol{\alpha} \in \{0, 1\}^K$,

$$\bar{p}_{\boldsymbol{\alpha}} = \begin{cases} p_{\boldsymbol{\alpha}}(1 - s_1)/(1 - \bar{s}_1) & \text{if } \alpha_1 = 1 \\ p_{\boldsymbol{\alpha}} + p_{\boldsymbol{\alpha}+\mathbf{e}_1}[1 - (1 - s_1)/(1 - \bar{s}_1)] & \text{if } \alpha_1 = 0 \end{cases}.$$

Under the above construction, we can see that for $\boldsymbol{\alpha}$ with the first attribute $\alpha_1 = 0$,

$$\bar{p}_{\boldsymbol{\alpha}} + \bar{p}_{\boldsymbol{\alpha}+\mathbf{e}_1} = p_{\boldsymbol{\alpha}} + p_{\boldsymbol{\alpha}+\mathbf{e}_1}[1-(1-s_1)/(1-\bar{s}_1)] + p_{\boldsymbol{\alpha}+\mathbf{e}_1}(1-s_1)/(1-\bar{s}_1) = p_{\boldsymbol{\alpha}} + p_{\boldsymbol{\alpha}+\mathbf{e}_1}.$$

We further take $\bar{s}_j = s_j$ for $j > 1$. Without loss of generality, in the following we consider $\mathbf{g} = 0$; otherwise, we can perform a linear transformation of the response T-matrices under the two sets of parameters in Proposition 1 to obtain equivalency (Xu & Zhang, 2016). With the above constructed parameters $(\bar{\mathbf{s}}, \bar{\mathbf{p}})$, we have that for any $\mathbf{x} \in \{0, 1\}^J$ satisfying $x_1 = 0$, $P(\mathbf{X} = \mathbf{x} \mid Q, \mathbf{s}, \mathbf{g} = \mathbf{0}, \boldsymbol{\alpha}) = P(\mathbf{X} = \mathbf{x} \mid Q, \bar{\mathbf{s}}, \mathbf{g} = \mathbf{0}, \boldsymbol{\alpha})$ and equivalently $T_{\mathbf{x},\cdot}(Q, \mathbf{s}, \mathbf{0}) = T_{\mathbf{x},\cdot}(Q, \bar{\mathbf{s}}, \mathbf{0})$. Therefore, we have

$$\begin{aligned} T_{\mathbf{x},\cdot}(Q, \mathbf{s}, \mathbf{0})\mathbf{p} &= \sum_{\{\boldsymbol{\alpha}:\alpha_1=0\}} t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \mathbf{s}, \mathbf{0})(p_{\boldsymbol{\alpha}} + p_{\boldsymbol{\alpha}+\mathbf{e}_1}) \\ &= \sum_{\{\boldsymbol{\alpha}:\alpha_1=0\}} t_{\mathbf{x},\boldsymbol{\alpha}}(Q, \bar{\mathbf{s}}, \mathbf{0})(\bar{p}_{\boldsymbol{\alpha}} + \bar{p}_{\boldsymbol{\alpha}+\mathbf{e}_1}) = T_{\mathbf{x},\cdot}(Q, \bar{\mathbf{s}}, \mathbf{0})\bar{\mathbf{p}}. \end{aligned} \tag{16.12}$$

Similarly, for any $\mathbf{x} \in \{0, 1\}^J$ satisfying $x_1 = 1$, we have

$$\begin{aligned} T_{\mathbf{x},\cdot}(Q, \mathbf{s}, \mathbf{0})\mathbf{p} &= \sum_{\boldsymbol{\alpha}:\alpha_1=1} t_{\mathbf{x}-\mathbf{e}_1,\boldsymbol{\alpha}}(Q, \mathbf{s}, \mathbf{0}) s_1 p_{\boldsymbol{\alpha}} \\ &= \sum_{\boldsymbol{\alpha}:\alpha_1=1} t_{\mathbf{x}-\mathbf{e}_1,\boldsymbol{\alpha}}(Q, \bar{\mathbf{s}}, \mathbf{0}) \bar{s}_1 \bar{p}_{\boldsymbol{\alpha}} = T_{\mathbf{x},\cdot}(Q, \bar{\mathbf{s}}, \mathbf{0})\bar{\mathbf{p}}. \end{aligned} \tag{16.13}$$

Thus we have found distinct sets of parameters satisfying (16.7), and shown that Condition C2 is necessary for the identifiability of slipping parameters. In summary, the above discussion and the proof of Theorem 2 give the following corollary.

Corollary 1 (Partial identifiability) *Consider a Q-matrix satisfying Condition C1. When the guessing parameters are known, the slipping parameters s_j, $j > K$, are all identifiable.*

For an item j with $1 \leq j \leq K$, which is single attribute item under C1, the following holds: (1) if the item's attribute only appears in itself in the Q-matrix, i.e., Q' does not require the attribute, then the slipping parameter s_j is not identifiable; (2) otherwise s_j is identifiable.

To explain the result of the corollary, consider Q_1 in (16.11). When the guessing parameters are known, the first and third items have their attributes appear twice and therefore s_1 and s_3 identifiable; on the other hand, the second item requires α_2 which only appears once in the Q-matrix and therefore s_2 not identifiable.

16.4.3 Identifiability Conditions When the Slipping and Guessing Parameters Are Unknown

When the slipping and guessing parameters are unknown, we need the following regularity conditions to establish identifiability of all model parameters.

(C1) Q is complete. When this holds, we assume without loss of generality that the Q-matrix takes the form in (16.10).
(C3) Each attribute is required by at least three items.
(C4) Any two different columns of the sub-matrix Q' in (16.10) are distinct.

Condition C1 is the same as previous sections and requires the Q-matrix to be complete. Condition C3 extends C2 and requires each attribute to be needed by more than two item, which is necessary to identify both the slipping and guessing parameters. Condition C4 assumes that any two different columns of the sub-matrix Q' in (16.10) are different, which is easy to check in practice.

When neither the slipping nor the guessing parameters are known, we have the following necessary and sufficient identifiability result, which was proved in Gu and Xu (2018b).

Theorem 3 (Sufficient and Necessary Identifiability Conditions) *Under the DINA model, $\mathbf{s}$, $\mathbf{g}$ and $\mathbf{p}$ are identifiable if and only if Conditions C1, C3 and C4 hold.*

Corollary 2 (Partial identifiability) *Suppose Conditions C1 and C3 hold but C4 does not hold. Then $\mathbf{s} = (s_1, \cdots, s_J)$ and $(g_{K+1}, \cdots, g_J)$ are identifiable while there exists at least one item $k \in \{1, \cdots, K\}$ such that g_k is not identifiable.*

Remark 5 Condition C1 requires the Q-matrix to be complete and is necessary for the identifiability of all DINA parameters. When the Q-matrix is incomplete, Gu and Xu (2018a) studied the partial identifiability of the DINA model and proposed

easily checkable conditions to ensure the identifiability of all item parameters. For instance, the slipping and guessing parameters are identifiable under the 20×8 Q-matrix of the fraction subtraction data specified in de la Torre and Douglas (2004). In addition, the following Q-matrices are incomplete under the DINA model, but they give identifiable slipping and guessing parameters:

$$Q_1 = \begin{pmatrix} 1\ 1\ 0 \\ 1\ 0\ 1 \\ 0\ 1\ 1 \\ 1\ 1\ 0 \\ 1\ 0\ 1 \\ 0\ 1\ 1 \end{pmatrix}, \qquad Q_2 = \begin{pmatrix} 1\ 1\ 0 \\ 1\ 0\ 1 \\ 0\ 1\ 1 \\ 1\ 1\ 0 \\ 1\ 0\ 1 \\ 1\ 1\ 1 \end{pmatrix}.$$

Theorem 3 and Corollary 2 give relatively simple conditions to check the identifiability of the DINA model. Proofs of these results are found in Gu and Xu (2018b). According to Theorem 3, neither Q_1 nor Q_2 from (16.11) describe identifiable DINA models when **s**, **g**, and **p** are all unknown. Particularly, since each attribute appears in less than three items, any of the item parameters (**s**, **g**, **p**) is not identifiable due to the necessity of the Condition C3. We further use the following example to illustrate the necessity of C4.

Example 6 (Necessity of C4 when $K = 2$) To illustrate the importance of Condition C4, consider the case when $K = 2$. For easy discussion, we assume there is no item requiring none of the attributes; then if C1 is satisfied but C4 is not satisfied, the Q can only have the following form (up to row switching)

$$Q = \begin{pmatrix} 1\ 0 \\ 0\ 1 \\ 1\ 1 \\ \vdots\ \vdots \\ 1\ 1 \end{pmatrix}_{J \times 2}, \tag{16.14}$$

that is, the first two items give an identity Q-matrix while all other items require both attributes. Without loss of generality, we assume C3 is also satisfied (but not C4) and show the necessity of C4. Suppose the true model parameters are (**s**, **g**, **p**). Next we construct another set of parameters $(\bar{\mathbf{s}}, \bar{\mathbf{g}}, \bar{\mathbf{p}}) \neq (\mathbf{s}, \mathbf{g}, \mathbf{p})$ which give the same response probabilities and therefore not distinguishable from the true parameters.

From Corollary 2, $\{s_j, j = 1, \cdots, J\}$ and $\{g_j, j = 3, \cdots, J\}$ are identifiable and we set $\bar{s}_j = s_j, j = 1, \cdots, J$ and $\bar{g}_j = g_j, j = 3, \cdots, J$. We further let $\bar{p}_{(11)} = p_{(11)}$. We next show $(g_1, g_2, p_{(00)}, p_{(10)}, p_{(01)})$ are not identifiable. Following the identifiability condition in Definition 1 and equivalently

Proposition 1, a direct calculation gives that $(\bar{g}_1, \bar{g}_2, \bar{p}_{(00)}, \bar{p}_{(10)}, \bar{p}_{(01)})$ and $(g_1, g_2, p_{(00)}, p_{(10)}, p_{(01)})$ are not identifiable if the following equations can be satisfied:

$$\begin{aligned}
\bar{p}_{(00)} + \bar{p}_{(10)} + \bar{p}_{(01)} &= p_{(00)} + p_{(10)} + p_{(01)};\\
(\bar{g}_1 + s_1 - 1)(\bar{p}_{(00)} + \bar{p}_{(01)}) &= (g_1 + s_1 - 1)(p_{(00)} + p_{(01)});\\
(\bar{g}_2 + s_2 - 1)(\bar{p}_{(00)} + \bar{p}_{(10)}) &= (g_2 + s_2 - 1)(p_{(00)} + p_{(10)});\\
(\bar{g}_1 + s_1 - 1)(\bar{g}_2 + s_2 - 1)\bar{p}_{(00)} &= (g_1 + s_1 - 1)(g_2 + s_2 - 1)p_{(00)}.
\end{aligned} \tag{16.15}$$

Here for any given set of true parameters $(g_1, g_2, p_{(00)}, p_{(10)}, p_{(01)})$, there are four constraints in (16.15) but there are five parameters $(\bar{g}_1, \bar{g}_2, \bar{p}_{(00)}, \bar{p}_{(10)}, \bar{p}_{(01)})$ to solve. Since we have more free parameters than the equations, we have the non-identifiability of $(g_1, g_2, p_{(00)}, p_{(10)}, p_{(01)})$.

The following types of complete Q-matrices satisfy C4:

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ * & * \\ \vdots & \vdots \\ * & * \end{pmatrix}_{J\times 2} \quad \textit{or equivalently} \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ * & * \\ \vdots & \vdots \\ * & * \end{pmatrix}_{J\times 2}, \tag{16.16}$$

where "$*$" can be either "0" or "1". Note that the two Q-matrices in (16.16) are equivalent up to the first two rows and the columns switching. For the Q-matrices in (16.16), if C3 is also satisfied (each attribute appears in at least three items), then all model parameters are identifiable.

Example 7 We consider more examples. The following Q-matrices satisfy Condition C4:

$$Q_1 = \begin{pmatrix} \mathcal{I}_K \\ \mathcal{I}_K \\ Q'' \end{pmatrix}, \quad Q_2 = \begin{pmatrix} \mathcal{I}_K \\ 1 - \mathcal{I}_K \\ Q'' \end{pmatrix}.$$

Therefore, if Condition C3 is satisfied for them, then all model parameters are identifiable. Note that for the above Q_2, C3 is automatically satisfied when $K \geq 3$.

Example 8 We use this example to show that Theorem 3 extends the identifiability result in Xu and Zhang (2016). Xu and Zhang (2016) gave a set of sufficient identifiability conditions, which however is not necessary. For instance, the following Q-matrix, which is given on page 633 in Xu and Zhang (2016), does not satisfy their sufficient condition, but still gives an identifiable model since it satisfies C1, C2 and C4.

$$Q = \begin{pmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \\ 1\,1\,1\,0 \\ 1\,1\,0\,1 \\ 1\,0\,1\,1 \\ 0\,0\,0\,1 \end{pmatrix}.$$

Example 9 (Real data example) This example considers a real data set collected by the Examination for the Certificate of Proficiency in English (ECPE) which is designed and organized by University of Michigan English Language Institute. The examination is developed to test high-level English language skills to determine the language proficiency of non-native speakers. It contains questions to evaluate grammar, vocabulary and reading skills of examinees. We use the data from 2003 to 2004 ECPE grammar section with 2922 subjects. In previous studies (e.g., Chiu et al. (2009)), 30 out of 40 items are selected to fit CDMs after 10 trial items are removed. The proposed Q-matrix contains three attributes: Morphosyntactic Form, Cohesive Form and Lexical Form. The Q-matrix is provided in Table 16.1. We can see it satisfies Conditions C1, C3 and C4; therefore under the DINA model, the model parameters are all identifiable. The identifiability under other CDMs, such as the Reduced-RUM, LCDM, and GDINA are considered in the next section.

When the identifiability conditions are satisfied, the maximum likelihood estimators of $\mathbf{s}$, $\mathbf{g}$ and $\mathbf{p}$ are consistent as the sample size $N \to \infty$. This can be proved following the argument given in Remark 5. If Bayesian estimation methods are employed, such as MCMC methods, the proposed identifiability conditions ensure the convergence of the estimates. When the identifiability conditions are not satisfied, the Bayesian estimators, such as MAP estimators, would depend on the modes of the priors; see Chen, Culpepper, Chen, and Douglas (2018). In practice, when researchers find that the proposed Q-matrix does not satisfy the identifiability conditions, it is recommended to design new items such that the identifiability conditions are satisfied.

16.5 Identifiability of General CDMs

We present the identifiability results of general CDMs under the restricted latent class model framework introduced in the section above, which include most of the existing CDMs as special cases, such as the R-RUM, LCDM and GDINA that are discussed in Examples 3, 4, and 5.

For diagnostic models introduced above, we provide in the following a unified sufficient condition that ensures their identifiability. Since the DINA model is a

Table 16.1 Q-matrix for the English proficiency test

	Q		
Item	Morphosyntactic	Cohesive	Lexical
1	1	1	0
2	0	0	0
3	0	1	0
4	1	0	1
5	0	0	1
6	0	0	1
7	0	0	0
8	0	0	1
9	1	0	1
10	0	1	0
11	0	0	1
12	1	0	0
13	1	0	1
14	1	0	1
15	1	0	0
16	1	0	0
17	0	0	1
18	1	0	1
19	0	1	1
20	0	0	1
21	0	0	1
22	1	0	1
23	1	0	1
24	0	0	1
25	0	1	0
26	0	1	0
27	1	0	0
28	0	0	1
29	1	0	0
30	0	0	1

special case of the restricted latent class models including higher order interaction effects of the attributes, such as GDINA and LCDM, it is recommended that we need to use a complete Q-matrix for the diagnostic models and we need at least three items for each attribute. To establish identifiability for the general class of models, we list below the conditions that will be used.

(D1) We assume that the Q-matrix takes the following form (after row swapping):

$$Q = \begin{pmatrix} \mathcal{I}_K \\ \mathcal{I}_K \\ Q' \end{pmatrix}. \tag{16.17}$$

(D2) Suppose Q has the structure defined in (16.17). We assume that for any $k \in \{1, \cdots, K\}$, $(\theta_{j,\mathbf{e}_k}; j > 2K)^\top \neq (\theta_{j,\mathbf{0}}; j > 2K)^\top$. That is, there exist at least one item in Q' such that subjects with $\boldsymbol{\alpha} = \mathbf{e}_k$ have different positively response probability from that of subjects with $\boldsymbol{\alpha} = \mathbf{0}$.

Condition D1 is a little stronger than the necessity of the complete matrix by requiring two identity submatrices. D1 itself implies that each attribute is required by at least two items. We need this condition to ensure enough information to identify the model parameters for each attribute. Condition D2 is a mild condition. For a general CDM such as LCDM and GDINA, it is satisfied if the main effects parameters are nonzero and each attribute appears once in Q'. Another example is that Condition D2 is automatically satisfied for all CDMs if Q' can be written as (after row swapping):

$$Q' = \begin{pmatrix} \mathcal{I}_K \\ \cdots \end{pmatrix},$$

equivalently, if there are three identity matrices in the Q-matrix, both D1 and D2 are satisfied. A third case is that if a general CDM allows that for one item j ($j > 2K$), $\theta_{j,\mathbf{0}} < \min_{\boldsymbol{\alpha} \neq \mathbf{0}} \theta_{j,\boldsymbol{\alpha}}$; that is, for subjects without any latent traits, the positive response probability is the lowest among all latent classes, then D2 is satisfied.

Our main identifiability results for general CDMs are as follows, which were established in Xu (2017).

Theorem 4 (Identifiability Conditions) *For any CDM satisfying the model setup in the sections above and Conditions D1–D2, the model parameters* $(\Theta, \mathbf{p})$ *are identifiable.*

Condition D1 itself is not enough to establish the identifiability of all parameters $(\Theta, \mathbf{p})$. An example is that under the DINA model with Q taking the form of (16.17) but Q' has at least one column being all zeros; such a Q-matrix satisfies D1 but as discussed above, the model parameters are not all identifiable since Condition C3 is not satisfied. However, Condition D1 ensures the identifiability of item parameters for items j, $j > 2K$. These partial identifiability results are given as follows.

Corollary 3 (Partial identifiability) *Under the model setup above, there exist Q-matrices satisfying D1 but the model is non-identifiable.*

On the other hand, if D1 is satisfied, the item parameters of items j, $j > 2K$, *are identifiable.*

The above theorem specifies the sufficient condition under which the CDM parameters $(\Theta, \mathbf{p})$ are identifiable from the response data. We use the following example to illustrate.

Example 10 (Example 9 continued) For the English proficiency data considered in Example 9, we can see that the Q-matrix in Table 16.1 satisfies Conditions D1 and D2 since it has three identity submatrices. Therefore, for popularly used CDMs, such as the DINA, DINO, R-RUM, LCDM, and GDINA, the model parameters are all identifiable.

Note that even when the CDMs are identifiable, empirically the estimation of the model parameters may still not be identifiable if the data does not support it, such as limited sample size, highly unbalanced attribute distribution, or misspecification of the dimensionality of the attribute profiles; see more discussions and further analysis of the English proficiency data in Templin and Bradshaw (2014) and von Davier (2014b).

Remark 6 (Consistency of MLE) When the identifiability conditions are satisfied, the maximum likelihood estimators of Θ and $\mathbf{p}$ are consistent as the sample size $N \to \infty$. Specifically, we introduce a 2^J-dimensional response vector $\boldsymbol{\gamma} = \{1, N^{-1}\sum_{i=1}^{N} I(\mathbf{X}_i \succeq \mathbf{e}_1), \cdots, N^{-1}\sum_{i=1}^{N} I(\mathbf{X}_i \succeq \mathbf{e}_J), N^{-1}\sum_{i=1}^{N} I(\mathbf{X}_i \succeq \mathbf{e}_1 + \mathbf{e}_2), \cdots, N^{-1}\sum_{i=1}^{N} I(\mathbf{X}_i \succeq \mathbf{1})\}$. From the definition of the T-matrix and the law of large numbers, we know $\boldsymbol{\gamma} \to T(Q, \Theta)\mathbf{p}$ almost surely as $N \to \infty$. On the other hand, the maximum likelihood estimators $\hat{\Theta}$ and $\hat{\mathbf{p}}$ satisfy

$$\|\boldsymbol{\gamma} - T(Q, \hat{\Theta})\hat{\mathbf{p}}\| \to 0,$$

where $\|\cdot\|$ is the L_2 norm. Therefore,

$$\|T(Q, \Theta)\mathbf{p} - T(Q, \hat{\Theta})\hat{\mathbf{p}}\| \to 0$$

almost surely. Then from the proof of the identifiability results, we can obtain the consistency result that $(\hat{\Theta}, \hat{\mathbf{p}}) \to (\Theta, \mathbf{p})$ almost surely. Furthermore, following a standard argument of the asymptotic theory, we take Taylor's expansion of the loglikelihood function at $(\Theta, \mathbf{p})$ and the central limit theorem would give the asymptotic normality of the estimators $(\hat{\Theta}, \hat{\mathbf{p}})$.

Remark 7 The proof of Theorem 4 in Xu (2017) is not based on the trilinear decomposition result in Kruskal (1976), which is applied in Allman, Matias, and Rhodes (2009) to show the generic identifiability up to label swapping. The generic identifiability results in Allman et al. (2009) can not be directly applied in the current model setting. This is because under the same Q-matrix, there may be several CDMs of interest. For instance, the DINA model can be taken as a submodel of the LCDM under the same Q-matrix, while more generally, the LCDM is a submodel of the GDM by extending the skill space (von Davier, 2014b). In this case, the parameters under the DINA model lie in a subspace of the parameter space under the LCDM and GDM, and generic identifiability results for the more general CDMs may not ensure the identifiability of the DINA model.

In addition, we would like to point out that Conditions D1 and D2 are different from the rank condition required by Kruskal's result and may be weaker in some cases. To apply Kruskal's result, a key condition is that there is a row partition of the Q-matrix, Q_1, Q_2, Q_3, such that

$$\text{rank}_k(\Gamma(Q_1, \Theta)) + \text{rank}_k(\Gamma(Q_2, \Theta)) + \text{rank}_k(\Gamma(Q_3, \Theta)) \geq 2^{K+1} + 2, \quad (16.18)$$

where the Kruskal rank of a matrix A, $\text{rank}_k(A)$, is the maximum number r such that any collection of r columns of A are linearly independent. The Kruskal rank is not larger than the matrix rank and generally the Kruskal rank and the matrix rank do not equal. Under Conditions D1 and D2, consider the case when $K = 2$ and $Q' = \mathbf{e}_1^\top$, i.e.,

$$Q = \begin{pmatrix} \mathcal{I}_2 \\ \mathcal{I}_2 \\ \mathbf{e}_1^\top \end{pmatrix}.$$

We have $\text{rank}(\Gamma(Q', \Theta)) = 2$ but $\text{rank}_k(\Gamma(Q', \Theta)) < 2$ since the column vectors corresponding to $\mathbf{e}_1$ and $\mathbf{1}$ are the same. Then if we decompose the Q-matrix as $\mathcal{I}_2, \mathcal{I}_2$ and Q', we can see that the sum of the Kruskal's ranks of the three parts is less than $2^{K+1} + 2$, due to the fact that $\text{rank}_k(\Gamma(\mathcal{I}_K, \Theta)) \leq 2^K$ and $\text{rank}_k(\Gamma(Q', \Theta)) < 2$. Similarly, for the other decompositions of the Q-matrix, we can verify the rank condition is also not satisfied. Therefore the rank condition (16.18) may not be satisfied under our conditions.

Theorem 4 gives the strict identifiability of general CDMs. A key requirement is the identity submatrix in Q. When such a requirement is not satisfied, researchers have recently considered the generic identifiability of the restricted latent class models (Gu & Xu, 2018a). The generic identifiability is defined following algebraic geometry terminology. It implies that the set of parameters for which the identifiability does not hold has Lebesgue measure zero (Allman et al., 2009). As for the general CDMs, Gu and Xu (2018a) proposed mild conditions on the form of the Q-matrix that lead to *generic* identifiability, which ensures that the model parameters $(\Theta, \mathbf{p})$ are identifiable almost everywhere in the restricted parameter space except a Lebesgue measure zero set. In particular, Gu and Xu (2018a) established the following result.

Theorem 5 (Generic Identifiability Conditions) *a Q matrix takes the following form up to row permutations*

$$Q = \begin{pmatrix} Q_1 \\ Q_2 \\ Q^\star \end{pmatrix}; \quad Q_i = \begin{pmatrix} 1 * \ldots * \\ * 1 \ldots * \\ \vdots \vdots \ddots \vdots \\ * * \ldots 1 \end{pmatrix}_{K \times K}, \quad i = 1, 2, \tag{16.19}$$

where each "$*$" can either be 1 or 0 and each attribute is required by at least one item in $Q^\star$, then the main-effect CDMs, such as R-RUM and ACDM, and the all-interaction-effect CDMs, such as the LCDM, GDINA and GDM, are generically identifiable.

The result in Theorem 5 would be helpful for practitioners, especially when it becomes difficult or even impossible to design pure items with single attribute

specifications. The identifiability results in Theorems 4 and 5 would also provide a helpful guideline for designing diagnostic tests. For the diagnostic classification models introduced in above, the model parameters are identifiable if the Q-matrix satisfies the proposed conditions. The theoretical results would also help to improve existing diagnostic tests. For instance, when researchers find that the estimation results are problematic and the Q-matrix does not satisfy the identifiability conditions, it is then recommended to add new items to satisfy them.

16.6 Further Discussions

This chapter reviews some of the existing identifiability results for CDMs. The completeness of the Q-matrix plays an important role for identifiability. When the Q-matrix is incomplete, the model parameters $(\Theta, \mathbf{p})$ are not identifiable under Definition 1. A particular case is when each row of the Q-matrix is $\mathbf{1}^\top$, then the model becomes similar to the unrestricted latent class models with 2^K classes. As shown in the literature (Gyllenberg, Koski, Reilink, & Verlaan, 1994), the unrestricted general latent class model is not identified. In such case, generic identifiability results would be practically useful as shown in Theorem 5.

When the identifiability conditions are not satisfied, we may expect to obtain partial identification results as discussed above. It is also possible in practice that there exist certain hierarchical structures among the latent attributes. For instance, a certain attribute may be a prerequisite for other attributes. In this case, some $\mathbf{p}$'s are restricted to be 0. In this chapter the attribute profile is modeled using a saturated model with $2^K - 1$ attribute profile parameters. It would be also interesting to consider the identifiability conditions under the unsaturated models. For these case, weaker conditions are expected for the identifiability of the model parameters, as studied in Gu and Xu (2018a).

The Q-matrix in this chapter is assumed to be correctly specified. In practice, the Q-matrix is usually constructed by the users and may not be accurate. A misspecified Q-matrix could lead to substantial lack of fit and, consequently, erroneous classification of subjects (Rupp & Templin, 2008; de la Torre, 2008). Thus it is recommended to apply the proposed identifiability results after validating the constructed Q-matrix. Various methods for Q-matrix validation and estimation can be found in recent works (e.g., Liu, Xu, & Ying, 2012; Liu et al., 2013, DeCarlo, 2012; Chen et al., 2015, 2018; de la Torre & Chiu, 2016; Gu, Liu, Xu, & Ying, 2018; Xu & Shang, 2018).

References

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics, 37*, 3099–3132.

Bechger, T. M., Verstralen, H. H., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika, 67*(1), 123–136.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd Ed.). Pacific Grove, CA: Duxbury Pacific Grove.

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika, 83*(1), 89–108.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*, 850–866.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.

de la Torre, J., & Douglas, J. A. (2004). Higher order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Erlbaum Associates.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Mahwah, NJ: Psychology Press.

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics, 8*(2), 261–263.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215–231.

Gu, Y., Liu, J., Xu, G., & Ying, Z. (2018). Hypothesis testing of the Q-matrix. *Psychometrika, 83*(3), 515–537.

Gu, Y., & Xu, G. (2018a). Partial identifiability of restricted latent class models. arXiv preprint arXiv:1803.04353.

Gu, Y., & Xu, G. (2018b, in press). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika*, https://doi.org/10.1007/s11336-018-9619-8.

Gyllenberg, M., Koski, T., Reilink, E., & Verlaan, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability, 31*, 542–548.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Koopmans, T. C. (Ed.) (1950). *Statistical inference in dynamic economic models* (Vol. 10). New York: Wiley, Inc.

Koopmans, T. C., & Reiersøl, O. (1950). The identification of structural characteristics. *Annals of Mathematical Statistics, 21*, 165–181.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika, 41*(3), 281–293.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548–564.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli, 19*(5A), 1790–1817.

Maris, G., & Bechger, T. M. (2004). Equivalent MIRID models. *Psychometrika, 69*(4), 627–639.

Maris, G., & Bechger, T. M. (2009). Equivalent diagnostic classification models. *Measurement, 7*, 41–46.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika, 21*, 331–347.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York/London: Springer.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society, 39*, 577–591.

Rupp, A. A., & Templin, J. L. (2008). Effects of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*, 78–98.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York: Guilford Press.

San Martín, E., Rolin, J.-M., & Castro, L. M. (2013). Identification of the 1PL model with guessing parameter: Parametric and semi-parametric results. *Psychometrika, 78*(2), 341–379.

Tatsuoka, C. (2009a). Diagnostic models as partially ordered sets. *Measurement, 7*, 49–53.

Tatsuoka, K. K. (2009b). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317–339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. New York, NY: Springer Science & Business Media.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

von Davier, M. (2014a). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*(1), 49–71.

von Davier, M. (2014b). The log-linear cognitive diagnostic model (lcdm) as a special case of the general diagnostic model (GDM). *ETS Research Report Series, 2014*(2), 1–13.

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika, 80*(1), 85–100.

Xu, G. (2017). Identifiability of restricted latent class models with binary responses. *The Annals of Statistics, 45*, 675–707.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association, 113*(523), 1284–1295.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*, 625–649.

Chapter 17
Measures of Agreement: Reliability, Classification Accuracy, and Classification Consistency

Sandip Sinharay and Matthew S. Johnson

Abstract Gierl, Cui, and Zhou (J Educ Meas 46:293–313, 2009), Cui, Gierl, and Chang (J Educ Meas 49:19–38, 2012), Templin and Bradshaw (J Classif 30:251–275, 2013), Wang, Song, Chen, Meng, and Ding (J Educ Meas 52:457–476, 2015), Johnson and Sinharay (J Educ Meas, 55: 635–664, 2018), and Johnson and Sinharay (J Educ Behav Stat, in press) suggested reliability-like measures for the estimates obtained from a diagnostic classification model. These measures mostly express the agreement between the estimated skill and the true skill, or between estimated skills from parallel assessments. This paper provides a review of these measures and demonstrates some of them for a real data example.

17.1 Introduction

Diagnostic classification models (DCM; e.g., Rupp, Templin, & Henson, 2010) or cognitive diagnostic models (CDMs) have been suggested several decades ago; for example, DiBello, Stout, and Roussos (1995), Maris (1999), and Mislevy, Almond, Steinberg, and Yan (1999) suggested such models in the 1990s. Further, DCMs have been fitted to data from a wide variety of assessments including the National Assessment of Educational Progress (Xu & von Davier, 2006), international large-scale survey assessments (Lee, Park, & Taylan, 2011; Oliveri & von Davier, 2011), language testing (von Davier, 2008), and the SAT (Gierl et al., 2009).

DCMs are mostly used to estimate the *mastery status* or probability of mastery of examinees. Standards 1.14 and 2.3 of the Standards for Educational and Psycholog-

Note: Any opinions expressed in this publication are those of the authors and not necessarily of Columbia University or Educational Testing Service.

S. Sinharay (✉) · M. S. Johnson
Educational Testing Service, Princeton, NJ, USA
e-mail: ssinharay@ets.org; msjohnson@ets.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_17

ical Testing (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 2014) recommend reporting of the reliability of subscores.[1] Because the estimated attributes from a DCM can be considered to be types of subscores, the standards imply that the reliability of the estimated mastery or of the estimated probability of mastery should be reported. However, as Sinharay and Haberman (2009) commented, those applying DCMs rarely reported the reliability of the estimated attributes with a few exceptions such as von Davier (2008); furthermore, no reliability-like measures existed for the estimates obtained from DCMs until recently.[2] Such a lack of research on the reliability of the estimates from DCMs is surprising given the abundance of research on the quality of subscores (e.g., Haberman, 2005; Haladyna & Kramer, 2004; Harris & Hanson, 1991) and on classification consistency and accuracy in the context of item response theory and strong true score theory (e.g., Hanson & Brennan, 1990; Lee, Hanson, & Brennan, 2002).

More recently, researchers such as Cui et al. (2012), Gierl et al. (2009), Johnson and Sinharay (2018, in press), Templin and Bradshaw (2013), and Wang et al. (2015) have recognized the importance of reporting reliability-like measures for the estimates obtained from the application of a DCM. Because the estimated attributes can be viewed as discrete mastery classifications (that denote what skills or attributes an examinee possesses), Cui et al. (2012) and Wang et al. (2015) suggested the use of measures of *classification accuracy* or *classification consistency* as indicators of reliability. Classification accuracy is the probability that the estimated classification is equal to the true classification; classification consistency is the probability that two parallel forms of the assessment result in the same estimated classification. Whereas Cui et al. (2012) discussed measures of accuracy and consistency for the entire vector of attributes, Wang et al. and Johnson and Sinharay (2018) suggested several measures of classification accuracy and consistency of the individual attributes. Templin and Bradshaw (2013) and Johnson and Sinharay (in press) developed, for each attribute in an application of a DCM, measures of reliability for the posterior probability of mastery or skill attainment of a randomly selected examinee.

The measures of Cui et al. (2012), Gierl et al. (2009), Johnson and Sinharay (2018, in press), Templin and Bradshaw (2013), and Wang et al. (2015) can all be considered to be measures of agreement, either between the true attributes and estimated attributes or between attributes estimated from two parallel forms. We will focus on DCMs that involve binary classifications, but several concepts discussed in this paper apply in a straightfoward manner to DCMs involving polytomous classifications. As in Johnson and Sinharay (2018), we define parallel forms of a DCM as two tests with the same Q-matrix and identical item parameters. We only

[1] The 1999 version of the standards also recommended the reporting of the reliability of subscores.

[2] von Davier (2008) used Monte Carlo simulation to examine classification accuracy, but only considered simulation studies where the true mastery patterns of the examinees were known.

consider dichotomous items in this paper, but most of the concepts discussed in this paper apply to polytomous items as well.

17.2 Existing Methods

17.2.1 Notation

To formalize the problem of developing reliability-like measures in applications of DCMs, consider an assessment that includes K items and measures D binary attributes. Let A_d denote the binary latent variable indicating whether a randomly chosen examinee truly possesses the attribute d (where $A_d = 1$ if the examinee possesses attribute d and $A_d = 0$ otherwise). Let $\boldsymbol{A} = (A_1, A_2, \cdots, A_D)'$ denote the collection of the A_d's, or the true attribute mastery pattern, for the examinee. Let Ω denote the set of all possible values of $\boldsymbol{A}$; thus, Ω consists of 2^D attribute patterns. Let $\boldsymbol{a} = (a_1, a_2, \cdots, a_D)'$ denote a realization of $\boldsymbol{A}$, with a_d being a realization of A_d. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_k, \ldots, X_K)^T$ denote the random vector of the item scores of a randomly chosen examinee on a K-item test. Now let $\tilde{a}_d(\boldsymbol{x})$ denote a binary estimate of a_d based on the set of observed item responses $\boldsymbol{x} = (x_1, x_2, \ldots, x_K)^T$, and $\tilde{\boldsymbol{a}}(\boldsymbol{x}) = (\tilde{a}_1(\boldsymbol{x}), \tilde{a}_2(\boldsymbol{x}), \cdots, \tilde{a}_D(\boldsymbol{x}))'$ denote the vector of all attribute estimates for an examinee. One could set $\tilde{a}_d(\boldsymbol{x})$ equal to the maximum a posteriori (MAP) estimator as did Wang et al. (2015) or Johnson and Sinharay (2018) or to any other satisfactory dichotomous estimator of the attribute.

17.2.2 Reliability of Each Attribute for the Attribute Hierarchy Method

Leighton, Gierl, and Hunka (2004) suggested the attribute hierarchy model (AHM) for diagnostic assessments in which the attributes can be ordered according to some hierarchy, that is, when an individual must possess, for example, attribute a_1 in order to possess attribute a_2, etc. For such assessments, Gierl et al. (2009) developed a reliability measure based on Cronbach's α for modified observed scores. The measure is of the form

$$\alpha_{AHM_d} = \frac{K_d}{K_d - 1}\left(1 - \frac{\sum_{k \in S_d} W_{kd}^2 \sigma_{x_k}^2}{\sigma^2_{\sum_{k \in S_d} W_{kd} x_k}}\right),$$

where S_d is the set of items that measures attribute d, K_d is the size of S_d, W_{kd} is the weight for item k in the calculation of attribute d and is fixed by the investigator, $\sigma_{x_k}^2$ is the variance of the observed scores on item k and $\sigma^2_{\sum_{k \in S_d} W_{kd} x_k}$ is the variance

of the weighted observed number-correct scores. The measure of Gierl et al. (2009) is the first reliability-like measure for any DCM according to the knowledge of the authors of this chapter, but applies only to applications of the AHM and does not easily generalize to applications of other DCMs.

17.2.3 Classification Accuracy and Consistency of the Attribute Pattern

Cui et al. (2012) suggested examining the agreement at the attribute pattern level using the measures

$$\text{Classification Accuracy} = P_c = P(\tilde{\boldsymbol{a}}(\boldsymbol{X}) = \boldsymbol{A}), \text{ and}$$

$$\text{Classification Consistency} = P_a = P(\tilde{\boldsymbol{a}}(\boldsymbol{X}_1) = \tilde{\boldsymbol{a}}(\boldsymbol{X}_2)),$$

where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are the item-scores of the same randomly selected examinee on two parallel assessments. Cui et al. (2012) noted that a DCM classifies each observed response pattern into one of H mutually exclusive latent classes where ideally, each latent class is associated with a distinct attribute pattern and $H = 2^D$. Let C_h denote the h-th latent class and π_h denote the set of all possible item-reponse patterns that lead to an examinee being classified into C_h. Cui et al. (2012) suggested computing classification accuracy and classification consistency using the formulas

$$P_a = \sum_{\boldsymbol{A}\in\Omega} \sum_{\boldsymbol{x}\in\pi_{\boldsymbol{A}}} P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{A}) r_{\boldsymbol{A}},$$

and

$$P_c = \sum_{\boldsymbol{A}\in\Omega} \left[\sum_{h=1}^{H} \left(\sum_{\boldsymbol{x}\in\pi_h} P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{A}) \right)^2 \right] r_{\boldsymbol{A}},$$

respectively, where r_A is the relative frequency of the attribute pattern A and $\pi_{\boldsymbol{A}}$ is the set of all x's that produce an estimated classification equal to $\boldsymbol{A}$. Wang et al. (2015) noted that the computation of either of the two above measures requires a summation over all possible item-response patterns for the test—therefore these measures can be extremely computation-intensive for tests with a modestly large or a large number of items.[3]

[3] Cui et al. (2012) considered test length of up to 20 in their data examples. Most likely, computation for tests consisting of much more than 20 items would take very long.

17.2.4 Classification Accuracy for Each Attribute

Wang et al. (2015) suggested examining agreement measures at the attribute level, in the same manner as Gierl et al. (2009). They defined the following measure of attribute-level classification accuracy:

$$P_{\text{CA}d} = P(\tilde{a}_d(\boldsymbol{X}) = a_d)\cdot \tag{17.1}$$

The above measure and several other measures discussed below are simple summaries of agreement for a hypothetical 2×2 table that consists of the proportions of the examinee population cross-classified by their true attribute indicator A_d and their estimated attribute $\tilde{a}_d(\boldsymbol{X})$.

Table 17.1 shows such a cross-classification. If the true proportions, p_{ij}, $i = 0, 1$, $j = 0, 1$, where $p_{ij} = P(A_d = i, \tilde{a}_d(\boldsymbol{X}) = j)$, were known, then the attribute-level accuracy would simply be

$$P_{\text{CA}d} = p_{00} + p_{11}. \tag{17.2}$$

Given a specific DCM such as the Deterministic Inputs, Noisy And gate (DINA) model, Noisy Inputs, Deterministic And gate (NIDA) model, or Loglinear Cognitive Diagnosis Model (LCDM; e.g., Rupp et al., 2010), a Q-matrix specifying which item requires which attribute, and a set of estimated item parameters, the p_{ij}'s can be calculated exactly by noting that

$$\begin{aligned} p_{ij} &\equiv P(A_d = i, \tilde{a}_d(\boldsymbol{X}) = j) \\ &= \sum_{\{\boldsymbol{x}:\, \tilde{a}_d(\boldsymbol{x})=j\}} P(A_d = i, \boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{\{\boldsymbol{x}:\, \tilde{a}_d(\boldsymbol{x})=j\}} P(A_d = i|\boldsymbol{X} = \boldsymbol{x})P(\boldsymbol{X} = \boldsymbol{x}). \end{aligned}$$

Noting that $P(A_d = i|\boldsymbol{X} = \boldsymbol{x})$ is the posterior probability of A_d being equal to i, the cell proportion p_{ij} is the average of the mean posterior probability, where the average is computed over all possible item response patterns.

While the accuracy indices can be calculated exactly with knowledge of the model parameters, if the number of items is large, the computation, which requires a summation over 2^K item response patterns, would be computationally prohibitive.

Table 17.1 The contingency table classifying individuals by their true and estimated attributes

	Estimate $\tilde{a}_d$		
True a_d	0	1	Total
0	p_{00}	p_{01}	p_{0+}
1	p_{10}	p_{11}	p_{1+}
Total	p_{0+}	p_{1+}	1

However, the p_{ij}'s, and therefore, the classification accuracy and consistency indices, can be estimated using a summation only over the response patterns observed in a random sample of examinees. Wang et al. (2015) and Johnson and Sinharay (2018) showed that one can estimate the p_{ij}s by

$$\hat{p}_{ij} = \frac{1}{N}\sum_{n=1}^{N} \Pr\{A_d = i|\boldsymbol{X} = \boldsymbol{x}_n\} I\{\hat{a}_{nd} = j\} \tag{17.3}$$

for all $d = 1, \ldots, D$, $i = 0, 1$, and $j = 0, 1$, where a_{nd} is the observed value of a_d for individual n whose item-score vector is denoted as $\boldsymbol{x}_n$, and $\hat{a}_{nd} = \tilde{a_d}(\boldsymbol{x}_n)$ is the binary estimate of a_{nd}.

Using Eqs. 17.2 and 17.3, Wang et al. (2015) estimated the classification accuracy of skill d with

$$\hat{P}_{\text{CA}d}=\frac{1}{N}\sum_{n=1}^{N}\tilde{a}_d(\boldsymbol{x}_n)P(A_d = 1|\boldsymbol{X} = \boldsymbol{x}_n)+\frac{1}{N}\sum_{n=1}^{N}(1-\tilde{a}_d(\boldsymbol{x}_n))P(A_d = 0|\boldsymbol{X}=\boldsymbol{x}_n)\cdot \tag{17.4}$$

Note that Wang et al. (2015) used the notation $\hat{\tau}_d$ (see their Eq. 17.6) to denote the quantity that is denote here as $\hat{P}_{\text{CA}d}$. The estimator $\hat{P}_{\text{CA}d}$ allows for consistent estimation of the true classification accuracy given a random sample from the population.

Wang et al. (2015) offered an additional method for estimating the attribute-level classification accuracy index in equations 25 through 27 in their paper, but Johnson and Sinharay (2018) argued that this method does not lead to a satisfactory estimation of $P_{\text{CA}d}$.

17.2.5 Classification Consistency for Each Attribute

Classification consistency for an attribute is the probability that a randomly selected individual would receive the same score on the attribute on two parallel forms of a test. Wang et al. (2015) suggested computing classification consistency at the attribute level as

$$P_{\text{CC}d} = P(\tilde{a}_d(\boldsymbol{X}_1) = \tilde{a}_d(\boldsymbol{X}_2))\cdot \tag{17.5}$$

Wang et al. (2015) suggested two approaches for estimating $P_{\text{C}d}$. The first estimator, denoted as $\hat{\gamma}_d$, is defined (Wang et al., Equation 9) as

$$\hat{\gamma}_d = \frac{1}{N}\sum_{n=1}^{N}\left([P(A_d = 1|\boldsymbol{X} = \boldsymbol{x}_n)]^2 + [P(A_d = 0|\boldsymbol{X} = \boldsymbol{x}_n)]^2\right). \tag{17.6}$$

Table 17.2 The contingency table classifying individuals on two parallel forms

	Estimate from Parallel Form 2 ($\tilde{a}_d(\boldsymbol{X}_2)$)		
Estimate from Parallel Form 1 ($\tilde{a}_d(\boldsymbol{X}_1)$)	0	1	Total
0	r_{00}	r_{01}	r_{0+}
1	r_{10}	r_{11}	r_{1+}
Total	r_{0+}	r_{1+}	1

Wang et al. (2015) justified this estimator by claiming that the posterior probability $P(A_d = 1|\boldsymbol{X} = \boldsymbol{x}_n)$ is constant across parallel forms of the test (at least almost surely). In an application of a DCM, the equality of the posterior probability would require individuals to produce the exact same item response vectors on parallel forms of the assessment (that is, $\boldsymbol{X}_1 = \boldsymbol{X}_2$). However, Wang et al. did not provide any theoretical proof or empirical results to support this claim and Johnson and Sinharay (2018) used a simple example to prove that the claim does not hold in general and that the estimator $\hat{\gamma}_d$ was biased. Wang et al. suggested another estimator of P_{CC_d} in their equation 24, but Johnson and Sinharay (2018) found in their simulations that this estimate is biased as well.

Whereas the attribute-level classification accuracy index can be viewed as a summary of the 2×2 table formed by cross-classifying individuals by their true and estimated attributes, classification consistency examines agreement in a 2×2 table formed by classifying individuals by their estimates on two parallel forms of a test. The cell proportions of such a table are

$$r_{ij} = P(\tilde{a}_d(\boldsymbol{X}_1) = i,\ \tilde{a}_d(\boldsymbol{X}_2) = j)\cdot$$

Table 17.2 shows such a table. Then the (true) classification consistency index is

$$P_{\text{CC}_d} = r_{00} + r_{11}. \tag{17.7}$$

Given estimated item parameters, these proportions can be calculated by applying the result that

$$\begin{aligned} r_{ij} &= P(\tilde{a}_d(\boldsymbol{X}_1) = i, \tilde{a}_d(\boldsymbol{X}_2) = j) \\ &= \sum_{\boldsymbol{a}} P(\tilde{a}_d(\boldsymbol{X}_1) = i, \tilde{a}_d(\boldsymbol{X}_2) = j|\boldsymbol{A} = \boldsymbol{a})P(\boldsymbol{A} = \boldsymbol{a}) \\ &= \sum_{\boldsymbol{a}} P(\tilde{a}_d(\boldsymbol{X}_1) = i|\boldsymbol{A} = \boldsymbol{a})P(\tilde{a}_d(\boldsymbol{X}_2) = j|\boldsymbol{A} = \boldsymbol{a})P(\boldsymbol{A} = \boldsymbol{a}). \end{aligned}$$

The final equality holds because of the assumption of conditionally independent test responses given the true latent attribute vector $\boldsymbol{a}$.

Johnson and Sinharay (2018) suggested estimating the proportions r_{ij}'s with their consistent estimators

$$\hat{r}_{ij} = \sum_{\boldsymbol{a} \in \Omega} \frac{\left(\sum_{n=1}^{N} P(\boldsymbol{A}=\boldsymbol{a}|\boldsymbol{X}=\boldsymbol{x}_n) I\{\hat{a}_{nd}=i\}\right)\left(\sum_{n=1}^{N} P(\boldsymbol{A}=\boldsymbol{a}|\boldsymbol{X}=\boldsymbol{x}_n) I\{\hat{a}_{nd}=j\}\right)}{N^2 P(\boldsymbol{A}=\boldsymbol{a})}. \tag{17.8}$$

Johnson and Sinharay (2018) suggested estimating P_{CC_d} by

$$\hat{P}_{\text{CC}_d} = \hat{r}_{00} + \hat{r}_{11}. \tag{17.9}$$

The estimator $\hat{P}_{\text{CC}_d}$ is different from the two suggested by Wang et al. (2015) and can be computed using data from only one form. In addition, $\hat{P}_{\text{CC}_d}$ is a consistent estimator of the classification consistency index if the model parameters are known and the examinees are randomly sampled from the population of interest, since both $\hat{r}_{00}$ and $\hat{r}_{11}$ are consistent estimators of r_{00} and r_{11}.

17.2.6 Reliability of the Posterior Probability of Mastery

Templin and Bradshaw (2013) noted that in applications of DCMs, one often reports the marginal posterior probabilities of mastery given the observed item responses, that is, reports

$$\hat{a}_d(\boldsymbol{x}_n) = P(A_{nd} = 1|\boldsymbol{X} = \boldsymbol{x}_n) \equiv E[A_{nd}|\boldsymbol{x}_n], \tag{17.10}$$

where A_{nd} is the random variable indicating whether examinee n possesses attribute d. Templin and Bradshaw (2013) developed a reliability measure using this estimator by constructing a 2×2 contingency table that is shown in Table 17.3, where

$$\hat{m}_{ij} = \frac{1}{N} \sum_{n=1}^{N} P(A_d = i|\boldsymbol{X} = \boldsymbol{x}_n) P(A_d = j|\boldsymbol{X} = \boldsymbol{x}_n). \tag{17.11}$$

Table 17.3 Hypothetical contingency table used to calculate the reliability measure proposed by Templin and Bradshaw

	a_2	
a_1	0	1
0	$\hat{m}_{00}$	$\hat{m}_{01}$
1	$\hat{m}_{10}$	$\hat{m}_{11}$

The reliability measure proposed by Templin and Bradshaw (2013), denoted henceforth as $\hat{\rho}_{TB}$, is the tetrachoric correlation calculated from the contingency table (shown in Table 17.3) defined by the proportions $\hat{m}_{ij}$'s.

Templin and Bradshaw (2013) did not make it clear whether $\hat{\rho}_{TB}$ measures the reliability of the estimated attributes or of the posterior probability of mastery.[4] Based on the facts that

- the heading used by the authors is "Measuring the Reliability of Diagnostic Model Examinee Estimates",
- the heading used also by the Templin & Bradshaw (2013) is "Diagnostic Classification Model Examinee Estimates",
- they did not include a binary classification as an examinee estimate,
- in their article on examinee estimates, they stated that "Examinees are more often provided with marginal probabilities of attribute mastery ..." and listed the probability of attribute mastery as an examinee estimate,
- the probabilities of attribute mastery play a major role in the computation of their measure,

we decided to treat $\hat{\rho}_{TB}$ as a measure of the reliability of the posterior probability of mastery.

Templin and Bradshaw (2013) derived $\hat{\rho}_{TB}$ on the basis of their assumption that the posterior probability $P(A_d = 1|\boldsymbol{X})$ is constant for an examinee across parallel forms of the test. This assumption is virtually identical to the assumption that Wang et al. (2015) made in deriving $\hat{\gamma}_d$ and was proved incorrect by Johnson and Sinharay (in press, 2018) under the traditional definition of parallel forms; Johnson and Sinharay (in press) found $\hat{\rho}_{TB}$ to overestimate the reliability of the posterior probability of mastery; the overestimation can most likely be attributed to the incorrect assumption.

Johnson and Sinharay (in press) suggested three measures of reliability of the posterior probability of mastery for DCMs. The first one of them is the *squared point biserial correlation* between the binary attribute mastery status A_d and its posterior expectation $E[A_d|\boldsymbol{X}]$,

$$\rho_{bis}(E[A_d|\boldsymbol{X}]) = (\text{cor}(A_d, E[A_d|\boldsymbol{X}]))^2 .$$

Johnson and Sinharay (in press) provided alternative expressions of this measure in terms of agreement statistics suggested by Yule (1912) and Youden (1950). Johnson and Sinharay (in press) also expressed $\rho_{bis}(E[A_d|\boldsymbol{X}])$ as the ratio of the observed score variance to the total variance in the context of DCMs. This ratio of observed to true score variance has been called the generic form of reliability in classical test theory (see Lord & Novick, 1968, Section 9.7). If we have a random sample of size

[4] They mentioned "reliability for the categorical attribute" and "reliability of the attribute" in a few places, but these are ambiguous regarding what $\hat{\rho}_{TB}$ actually measures.

N from the population of interest and the estimated item parameters are available, $\rho_{bis}(E[A_d|\boldsymbol{X}])$ can be consistently estimated by

$$\hat{\rho}_{bis}(E[A_d|\boldsymbol{X}]) = \frac{\frac{1}{N}\sum_{n=1}^{N}(E[A_d|\boldsymbol{X}=\boldsymbol{x}_n])^2 - p_d^2}{p_d(1-p_d)}, \tag{17.12}$$

where $p_d = P(A_d = 1)$.

The second measure suggested by Johnson and Sinharay (in press) is the *parallel-forms reliability* defined as

$$\rho_{pf}(E[A_d|\boldsymbol{X}])) = \text{cor}(E[A_d|\boldsymbol{X}_1], E[A_d|\boldsymbol{X}_2])\cdot$$

This correlation can be expressed as

$$\rho_{pf}(E[A_d|\boldsymbol{X}]) = \frac{\sum_{\boldsymbol{a}}\left[\sum_{\boldsymbol{x}} E[A_d|\boldsymbol{x}]P(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{A}=\boldsymbol{a})\right]^2 P(\boldsymbol{A}=\boldsymbol{a}) - p_d^2}{\text{var}(E[A_d|\boldsymbol{X}])}. \tag{17.13}$$

When a single skill is assessed, $\rho_{bis}(E[A_d|\boldsymbol{X}])$ and $\rho_{pf}(E[A_d|\boldsymbol{X}])$ are equivalent. However, when more than one skill is assessed, they need not be equal. Given a DCM, Q-matrix, and parameters, $\rho_{pf}(E[A_d|\boldsymbol{X}])$ can be calculated exactly. However, because it requires a double summation, one over both the entire set of the 2^D possible attribute patterns and the other over all 2^K possible item score vectors, it could be computationally prohibitive. Given a random sample from the population of interest, one can estimate $\rho_{pf}(E[A_d|\boldsymbol{X}])$ consistently with the estimator

$$\hat{\rho}_{pf}(E[A_d|\boldsymbol{X}]) = \frac{\sum_{\boldsymbol{a}}\frac{1}{P(\boldsymbol{A}=\boldsymbol{a})}\left[\frac{1}{N}\sum_{n=1}^{N}E[A_d|\boldsymbol{X}=\boldsymbol{x}_n]P(\boldsymbol{A}=\boldsymbol{a}|\boldsymbol{X}=\boldsymbol{x}_n)\right]^2 - p_d^2}{\frac{1}{N}\sum_{n=1}^{N}(E[A_d|\boldsymbol{X}=\boldsymbol{x}_n])^2 - p_d^2}. \tag{17.14}$$

The third measure suggested by Johnson and Sinharay (in press), referred to as the *informational reliability*, is the squared informational correlation (e.g., Linfoot, 1957) between A_d and $\hat{a}_d(\boldsymbol{X})$, that is,

$$\rho_{\mathcal{J}} = 1 - \exp\left\{2(H(A_d) - H(A_d|\hat{a}_d(\boldsymbol{X})))\right\}, \tag{17.15}$$

where $\hat{a}_d(\boldsymbol{X})$ is defined in Eq. 17.10. The term $H(A_d)$ is the prior entropy and is defined as

$$H(A_d) = -p_d \ln p_d - (1-p_d)\ln(1-p_d).$$

The other term in Eq. 17.15, $H(A_d|\hat{a}_d(\boldsymbol{X}))$, is the conditional entropy and is computed as

$$H(A_d|\hat{a}_d(\boldsymbol{X})) = -\sum_{\boldsymbol{x}} P(\boldsymbol{X} = \boldsymbol{x}) \left[\hat{a}_d(\boldsymbol{x}) \ln \hat{a}_d(\boldsymbol{x}) + (1 - \hat{a}_d(\boldsymbol{x})) \ln(1 - \hat{a}_d(\boldsymbol{x}))\right].$$

Given a random sample from the population of interest, $H(A_d|\hat{a}_d(\boldsymbol{X}))$ can be estimated with

$$\hat{H}(A_d|\hat{a}_d) = -\frac{1}{N} \sum_{n=1}^{N} \left(\hat{a}_d(\boldsymbol{x}_n) \ln \hat{a}(\boldsymbol{x}_n) + (1 - \hat{a}_d(\boldsymbol{x}_n)) \ln(1 - \hat{a}_d(\boldsymbol{x}_n))\right),$$

and the informational reliability can be estimated with

$$\hat{\rho}_{\mathcal{J}} = 1 - \exp\left\{2(H(A_d) - \hat{H}(A_d|\hat{a}_d))\right\}. \tag{17.16}$$

17.2.7 Measures of Agreement Beyond Classification Accuracy and Consistency

While classification accuracy and consistency measures are intuitive and widely used, it has long been established that some of the properties of the probability of agreement in 2×2 tables render the measures difficult to interpret in many cases (e.g., Goodman & Kruskal, 1954; Youden, 1950). For example, if the proportion of individuals with the skill is 0.9 and the proportion of individuals estimated to have the skill is also 0.9, then the probability of agreement is equal to $0.9(0.9) + (0.1)(0.1) = 0.82$ even if the true and estimated attributes are independent of each other. When trying to quantify the quality of a CDA or the associated DCM, this issue can lead to very surprising results. For example, suppose that we know that 90% of the population possess an attribute. Now suppose that we fit a DCM that provides absolutely no information about the attribute. The MAP estimator will be 1 for *all* examinees, since the prior distribution and the posterior distribution are equal to one another.[5] Therefore, $p_{11} = 0.9$, $p_{01} = 0.1$, $p_{10} = p_{00} = 0$, $r_{11} = 1$, and the accuracy and consistency measures of the assessment will be $P_{\text{CA}_d} = 0.9$ and $P_{\text{CC}_d} = 1.0$ respectively. These values would make the CDA and DCM appear very informative on the attribute when they actually provide no information.

Johnson and Sinharay (2018) suggested the following measures that overcome these limitations of the probability of agreement:

- *Youden's statistic*. Youden's statistic $\mathcal{J}$ (Youden, 1950), which, in the case of DCMs, compares the probability that a DCM correctly classifies an individual

[5] That is because for each examinee, the posterior distribution for the attribute will be a discrete distribution with a probability of 0.9 on the value of 1 and a probability of 0.1 on the value 0, which would result in a MAP of 1.

that has the skill (true positive) to the probability that the DCM incorrectly classifies an examinee that does not have the skill (false positive); that is,

$$\begin{aligned}\mathcal{J} &= P(\tilde{a}_d(\boldsymbol{x}) = 1|A_d = 1) - P(\tilde{a}_d(\boldsymbol{x}) = 1|A_d = 0) \\ &= \frac{p_{11}}{p_{11} + p_{10}} - \frac{p_{01}}{p_{00} + p_{01}}.\end{aligned}$$

The statistic ranges from -1 to $+1$. It is $+1$ only when there are no classification errors, and is zero when the same proportions of individuals with and without the skill are estimated to have the attribute. For example, in the above non-informative example, the Youden's statistic for accuracy would be 0.

- *Goodman & Kruskal's Lambda*. Goodman and Kruskal (1954) introduced the statistic Λ which, in applications of DCMs, adjusts for a baseline case where the DCM would always choose the modal category. The statistic is computed as

$$\Lambda = \frac{p_{11} + p_{00} - \max\{p_{1+}, p_{0+}\}}{1 - \max\{p_{1+}, p_{0+}\}}.$$

The values of Λ will usually be non-negative. The statistic will be zero if the posterior mode never differs from the prior modal classification (as in the above non-informative test example) and will be one if there is perfect classification.

- *Cohen's kappa*. Cohen (1960) suggested computing the difference of the observed agreement (between the two classifications) and the agreement that is expected if they are independent, and normalizing the difference as described by the following equation

$$\kappa = \frac{p_{11} + p_{00} - p_{1+}p_{+1} - p_{0+}p_{+0}}{1 - p_{1+}p_{+1} - p_{0+}p_{+0}}.$$

The statistic will be zero when classifications are independent (as with a non-informative test) and one when there is perfect agreement.

- *Tetrachoric correlation*. The tetrachoric correlation describes the association between two binary variables as the correlation between two correlated normal random variables that would result in quadrant probabilities $(p_{11}, p_{01}, p_{10}, p_{00})'$, where, e.g., $p_{11} = P(Z_1 > 0, Z_2 > 0)$. When any of the four proportions is zero (as in the non-informative test example), the tetrachoric correlation is not defined.
- *Sensitivity and Specificity*. The true positive (TP) rate or sensitivity is the proportion of those with the skill that are correctly identified as having the skill. The true negative (TN) rate or specificity is the proportion of those lacking the skill that correctly identified as lacking the skill. They are computed as

$$\begin{aligned}\text{TP} &\equiv P(\tilde{a}_d(\boldsymbol{X}) = 1|A_d = 1) = \frac{p_{11}}{p_{1+}}, \text{ and} \\ \text{TN} &\equiv P(\tilde{a}_d(\boldsymbol{X}) = 0|A_d = 0) = \frac{p_{00}}{p_{0+}}.\end{aligned}$$

For the non-informative test example, $TP = 1$ and $TN = 0$. The classification index P_{CA_d} is simply a weighted average of the true positive and true negative rates, where the weights are equal to the prevalence of the attribute in the population, $P_{CA_d} = TP \times p_{1+} + TN \times p_{0+}$.

While the association measures described above were described as measures of accuracy, and thus computed with p_{ij}'s, they can also be described as measures of consistency and hence can be calculated using the estimates of r_{ij}'s.

17.2.8 Other Measures

The `CDM` package in `R` (Robitzsch, Kiefer, George, & Uenlue, 2014) includes a function `cdm.est.class.accuracy` that calculates accuracy and consistency measures at both the pattern level like Cui et al. (2012) and at the attribute level like Wang et al. (2015). However, there is no peer-reviewed publication supporting the method used in this function at the current time. In the simulation study in Johnson and Sinharay (2018), the results from the `R` function differs substantially from the accuracy and consistency indices of Wang et al. (2015) or Johnson and Sinharay (2018).

17.2.9 A List of the Existing Reliability-Like Measures

Table 17.4 provides a list of most of the statistics described above, states whether they are measures of consistency or accuracy or both, or of reliability, and provides the source of each statistic.

17.3 A Real Data Example

To demonstrate the measures discussed above, we calculated them for a data set from the grammar section of the Examination for the Certificate of Proficiency in English (ECPE). The data set was analyzed by Templin and Hoffman (2013) and von Davier (2014). The grammar section includes 28 multiple choice items in which a set of words is missing. Examinees are asked to select the appropriate word(s) for the missing part of the statement from four response options; for example, in an item, the examinees had to fill the blank in the statement "Mary had to lean _____ the counter to open the window." by choosing one word from the four following options: (a) above, (b) over, (c) after, and (d) around. The data set is included in the `CDM` package in `R` (Robitzsch et al., 2014) and includes the responses of 2922 examinees to the items. The Q-matrix for the data appears in Table 1 of Templin

Table 17.4 Summary of accuracy, consistency, and reliability (of posterior probability of mastery) measures

Statistic	Notation	Proposed by	Accuracy/Consistency
Accuracy index	$\hat{P}_{CAd}$	Wang et al. (2015)	Accuracy
Consistency est	$\hat{\gamma}_d$	Wang et al. (2015)	Consistency
Corrected consistency	$\hat{P}_{CCd}$	Johnson and Sinharay (2018)	Consistency
Youden's statistic	$\mathcal{J}$	Johnson and Sinharay (2018)	Both
Goodman & Kruskal	Λ	Johnson and Sinharay (2018)	Both
Cohen	κ	Johnson and Sinharay (2018)	Both
Tetrachoric correlation	ρ_T	Johnson and Sinharay (2018)	Both
Sensitivity & Specificity	TP, TN	Johnson and Sinharay (2018)	Both
Tetrachoric correlation	$\hat{\rho}_{TB}$	Templin and Bradshaw (2013)	Reliability
Biserial correlation	$\rho_{bis}(E[A_d\|\boldsymbol{X}])$	Johnson and Sinharay (in press)	Reliability
Parallel-forms reliability	$\rho_{pf}(E[A_d\|\boldsymbol{X}])$	Johnson and Sinharay (in press)	Reliability
Informational reliability	$\rho_{\mathcal{J}}$	Johnson and Sinharay (in press)	Reliability

Note: "Both" means "Both Accuracy and Consistency"

and Hoffman (2013). The items measure knowledge of one or more of the three following attributes: (1) morphosyntactic rules; (2) cohesive rules; and (3) lexical rules. The test includes respectively five, four, and ten items that measure only the first attribute, only the second attribute, or only the third attribute. Two items measure both the first two attributes, seven measure the first and third attribute, and one measures the second and third attribute.

As in Templin and Hoffman (2013), we fitted a saturated log-linear CDM (Henson, Templin, & Willse, 2009) to the ECPE data, where items measuring more than one skill contain both main effects and the interaction effect of the skills.

The estimated pattern-level accuracy and consistency measures (Cui et al., 2012) were 0.75 and 0.67, respectively, for the data set. We examined classification accuracy and consistency for the MAP estimators and the reliability of the posterior probability of mastery of the three individual attributes. In addition, we also produced Monte Carlo approximated measures of the accuracy and consistency, denoted ACCR and ATRCR. These Monte Carlo approximated indices were obtained by performing the following steps:

1. Generate 100,000 attribute patterns $\boldsymbol{a}_i$ from the categorical distribution with probabilities $\widehat{\Pr}\{\boldsymbol{A} = \boldsymbol{a}\}$.
2. For each generated attribute pattern $\boldsymbol{a}_i$, generate two parallel sets of item responses from the loglinear CDM.
3. For each of the 100,000 generated examinees, calculate two sets of MAP estimators from the parallel sets. For example, $\hat{\boldsymbol{a}}_{i1}$ and $\hat{\boldsymbol{a}}_{i2}$ are two pattern level estimators for individual i.

Table 17.5 Reliability-like measures for the ECPE data

Measure	Symbol	Attribute 1	Attribute 2	Attribute 3
Prevalence	π_d	0.39	0.55	0.67
Accuracy	ACCR	0.91	0.87	0.95
	$\hat{P}_{CAd}$	0.90	0.86	0.92
	Λ	0.74	0.68	0.75
	κ	0.72	0.70	0.65
	$\mathscr{J}$	0.79	0.71	0.81
	ρ_T	0.95	0.90	0.96
	TP	0.87	0.88	0.95
	TN	0.92	0.82	0.86
Consistency	ATRCR	0.85	0.82	0.88
	$\hat{P}_{CCd}$	0.83	0.81	0.86
	Λ	0.57	0.56	0.56
	κ	0.72	0.68	0.63
	$\mathscr{J}$	0.65	0.61	0.68
	ρ_T	0.85	0.82	0.88
	TP	0.78	0.83	0.90
	TN	0.86	0.78	0.78
Reliability	$\hat{\rho}_{TB}$	0.89	0.80	0.92
	$\rho_{bis}(E[A_d\|\boldsymbol{X}])$	0.70	0.59	0.74
	$\rho_{pf}(E[A_d\|\boldsymbol{X}])$	0.74	0.73	0.76
	$\rho_{\mathscr{J}}$	0.58	0.51	0.59

4. Calculate the Monte Carlo approximated accuracy and consistency indices. For example, for the attribute d,

$$\text{ACCR} = \frac{1}{2M}\sum_{i=1}^{M}\sum_{m=1}^{2} I\{\hat{a}_{imd} = a_{id}\},$$

$$\text{ATRCR} = \frac{1}{M}\sum_{i=1}^{M} I\{\hat{a}_{i1d} = \hat{a}_{i2d}\},$$

where $M = 100,000$ is the total number of generated examinees and $\hat{a}_{imd}$ is the estimate for individual i for attribute d calculated from parallel form m.

Note that ACCR and ATRCR, because of the way they are computed using simulated data, provide the estimated classification accuracy of consistency when the DCM perfectly fits the data whereas measures such as $\hat{P}_{CAd}$ provide corresponding estimates irrespective of how well the data set fits the DCM. In addition, the computation of ACCR and ATRCR involves simulations whereas the computation of measures such as $\hat{P}_{CAd}$ and $\hat{P}_{CCd}$ does not involve simulations.

Table 17.5 includes the values of some of the abovementioned reliability-like measures for the ECPE data. The table shows that the values of the classification accuracy measure $\hat{P}_{CAd}$ (Wang et al., 2015) are very similar to the results obtained through Monte Carlo simulation. The values of the classification consistency measure proposed by Johnson and Sinharay (2018), $\hat{P}_{CCd}$, are also very similar to those derived by Monte Carlo approximation. For any skill, the value of a consistency measure is smaller than that of the corresponding accuracy measure; this is expected and researchers such as Cui et al. (2012) also found the consistency measures to be smaller than accuracy measures. The values of $\hat{\rho}_{TB}$ (Templin & Bradshaw, 2013) are considerably larger than the corresponding values of the measures suggested by Johnson and Sinharay (in press). The values of $\hat{\rho}_{TB}$ are not close to the ATRCR's either; therefore, even if one considers $\hat{\rho}_{TB}$ as an estimate of classification consistency, the measure does not perform very well for these data. Overall, the accuracy and consistency measures in Table 17.5 seem to indicate that the assessment does a decent job of estimating the attributes of examinees.[6]

17.4 Discussion

In an application of a DCM, it is important to provide the end-user with some measure of the quality of the CDA and the associated DCM. It is also important to recognize the limitations of those reliability measures. This chapter reviews most of such existing measures, referred to as measures of agreement—they primarily consist of measures of classification consistency and accuracy of the individual attributes and measures of reliability of the posterior probability of mastery.

Researchers and practitioners often wonder about the interpretation of reliability and reliability-like measures, mainly because of the lack of unanimous guidelines on what values of reliability can be considered large enough for a given test. Johnson and Sinharay (2018) suggested some guidelines on measures of classification consistency and accuracy of the individual attributes, but more research on establishing further guidelines on these measures would be helpful.

An important and practically relevant question, given the prevalence of all the measures discussed in this paper, is "Which measure(s) should be reported in an application of a DCM to a real data set?" Our recommendation is to report a collection of different types of measures. For example, one may report a measure of classification accuracy and classification consistency,[7] and one of the measures of the reliability of the posterior probability of mastery. We also think that one should,

[6]However, von Davier and Haberman (2014) found that a located latent class analysis with four levels also fits the data well.

[7]In their simulation study, Johnson and Sinharay (2018) found the Goodman & Kruskal Λ for accuracy to have the best properties among such measures.

in addition, report a couple of measures from Rows 4–8 of Table 17.4 given that classification accuracy and classification consistency can be misleading for skills that have very high or low prevalence. One option would be to report the prevalence of each attribute, the TP and TN rates. These three numbers would allow the users to weight the different types of errors (false positive and false negative) differently, depending on the type of application. For example, for one test user, it may be more of a problem to inaccurately estimate the skill of someone who does not have it, i.e., to make a false positive; in this case, the user could give more importance to the true negative rate (that is equal to one minus the false positive rate) than the true positive rate.

While reliability-like measures provide information about the quality of DCMs, there exist other ways to obtain information about the quality of DCMs. The fit of a DCM to the data also provide information about the quality of DCMs. Researchers such as de la Torre and Lee (2013), Oliveri and von Davier (2011), and Park, Johnson, and Lee (2015) have investigated the issue of goodness of fit of DCMs.

There are several topics related to reliability of DCMs that can be further investigated. First, the simulations and the real data examples in existing research on reliability-like measures focused only a few DCMs such as the DINA model; it is possible to consider other DCMs such as NIDA, LCDM (e.g., Rupp et al., 2010), and general diagnostic model (von Davier, 2008) in future research. Second, most researchers who investigated the reliability of DCMs considered only binary items; therefore, it is possible to examine reliability-like measures for polytomous data. Third, it is possible to consider more simulated and real data sets in future research on such measures. Fourth, it will be interesting to examine how the reliability-like measures are affected by misfit of DCMs. Finally, all the measures of agreements are computed under the assumption that the item parameters are known and the DCM fits the data. While Johnson and Sinharay (2018) performed a brief investigation of how the agreement measures are affected by uncertainty in the item parameters and by misfit of the DCM, more research on these areas would be useful.

References

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cui, Y., Gierl, M., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*, 19–38.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Iowa City, IA: Lawrence Erlbaum.

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 293–313.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*, 732–764.

Haberman, S. J. (2005). *When can subscores have value?* (ETS Research report No. RR-05–08). Princeton, NJ: ETS.

Haladyna, S. J., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions, 24*, 349–368.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement, 27*, 345–359.

Harris, D. J., & Hanson, B. A. (1991). *Methods of examining the usefulness of subscores.* Paper Presented at the Annual meeting of the National Council of Measurement in Education, Chicago, IL.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Johnson, M. S., & Sinharay, S. (in press). The reliability of the posterior probability of skill attainment in diagnostic classification models. *Journal of Educational and Behavioral Statistics*.

Johnson, M. S., & Sinharay, S. (2018). Measures of agreement to assess attribute-level classification accuracy and consistency for cognitive diagnostic assessments. *Journal of Educational Measurement*, *55*, 635–664.

Lee, W.-C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*, 412–432s.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*, 144–177.

Leighton, J., Gierl, M., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237.

Linfoot, E. (1957). An informational measure of correlation. *Information and Control, 1*, 85–89.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Mislevy, R. J., Almond, R. G., Steinberg, L. S., & Yan, D. (1999). Bayes nets in educational assessment: Where do the numbers come from? In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, Stockholm, Sweden (pp. 437–446).

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*, 315–333.

Park, J. Y., Johnson, M. S., & Lee, Y.-S. (2015). Posterior predictive model checks for cognitive diagnostic models. *International Journal of Quantitative Research in Education, 2*(3/4), 244.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2014). CDM: Cognitive diagnosis modeling [Software-Handbuch]. http://CRAN.R-project.org/package=CDM (R package version 4.1).

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford.

Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives, 6*, 46–49.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*, 251–275.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using mplus. *Educational Measurement: Issues and Practice, 32* (2), 37–50.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (ETS research report No. RR-14–40). Princeton, NJ: ETS.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models—a commentary. *Psychometrika, 79*, 340–346.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for diagnostic assessment. *Journal of Educational Measurement, 52*, 457–476.

Xu, X., & von Davier, M. (2006). *Cognitive diagnosis for NAEP proficiency data* (ETS research report No. RR-06–08). Princeton, NJ: ETS.

Youden, W. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35.

Yule, G. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statisical Society, 75*, 579–652.

Chapter 18
Differential Item Functioning in Diagnostic Classification Models

Xue-Lan Qiu, Xiaomin Li, and Wen-Chung Wang

Abstract Assessment of differential item functioning (DIF) in diagnostic classification models (DCMs) has begun to attract research attention. In previous studies, authors found that DIF detection in DCMs appeared to be very powerful even when most or all the items on the studied test had DIF and no scale purification was necessary. This surprisingly good result was built on studies that made the unrealistic assumption of equality of the model and the Q-matrix across groups. The present study clarifies these weaknesses in previous studies, identifies various types of DIF, and proposes new DIF detection methods that are powerful in detecting DIF in DCMs. An illustrative simulation study was conducted to demonstrate the feasibility and advantages of the new methods. Finally, conclusions and suggestions for future studies are provided.

Given the wide use of educational and psychological tests all over the world, it is a logical and moral imperative for test developers and users to ensure test fairness. One serious threat to test fairness is differential item functioning (DIF). An item is deemed to expose DIF when two test-takers who have an identical level of the latent variable the test intends to measure but belong to different groups (e.g., gender or ethnicity) have different probabilities of success or endorsement on the item (Holland & Wainer, 1993). Expressed mathematically, an item exhibits DIF when the following equation does not hold:

$$f\left(Y|\theta, G\right) = f\left(Y|\theta\right), \tag{18.1}$$

where Y is the item response; θ is the latent variable that the test intends to measure, which can be continuous as in classical test theory (CTT) or item response theory (IRT) or an attribute profile as in diagnostic classification models (DCMs); and G

X.-L. Qiu (✉) · X. Li · W.-C. Wang (deceased)
The University of Hong Kong, Pokfulam, Hong Kong
e-mail: xlqiu@hku.hk; xmli@eduhk.hk

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_18

is the group membership, which can be categorical (such as gender), continuous (such as age), or latent (such as a problem-solving strategy). If a test contains DIF items, then the test measures different latent variables for different groups of test-takers (i.e., measurement invariance does not hold across groups); therefore, the raw total score in CTT, the person ability measure in IRT, or the attribute profile in DCM is no longer comparable across groups (Holland & Wainer, 1993, p. xi). If theoretically incomparable scores are compared across groups (e.g., boys and girls), the resulting hypothesis testing (e.g., whether boys have a higher mean ability than girls, whether John has a higher ability than Mary) and subsequent decision making will be erroneous. Acknowledging these consequences, practitioners usually remove DIF items from a test to ensure score comparability across groups.

Many DIF detection methods have been developed in which responses are assumed to follow IRT models, subsequently referred to as IRT-DIF in this chapter. When item responses are generated from DCMs, the corresponding DIF detection methods are referred to as DCM-DIF in this chapter. Given the great similarity between IRT-DIF and DCM-DIF, the following review of IRT-DIF will shed light on DCM-DIF.

18.1 Review of IRT-DIF

There are two major categories of DIF detection methods in IRT-DIF: One is IRT-based methods, and the other is non-IRT-based methods. In IRT-based DIF detection methods, an IRT model is fit to the data, and the item parameter estimates for different groups of test-takers are estimated, placed on the same scale, and compared. If the item parameter estimates are found to differ systematically across groups, then the studied item is deemed to have DIF. Lord's (1980) method and the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988) belong to this category. In non-IRT-based detection methods, raw scores are usually used to match test-takers from different groups so that the performances of the different groups on the studied item can be compared. If the response probabilities are statistically different given comparable test performance, then the studied item is deemed to have DIF. The Mantel-Haenszel (MH; Mantel & Haenszel, 1959) method (Holland & Thayer, 1988) and the logistic regression (LR) method (Swaminathan & Rogers, 1990) belong to this category.

Based on the definition of DIF, test-takers from different groups must be matched on the target latent variable θ that the test intends to measure to enable DIF assessment. When tests contain DIF items, the raw total score or the IRT person measure that is calibrated from the contaminated test no longer has the same meaning for different groups. Therefore, these scores do not represent the uncontaminated latent variable, and cannot be used to match test-takers. Take an English-version logical reasoning test as an example. For English speakers, the test measures logical reasoning as intended, whereas the test may measure a composite of English proficiency and logical reasoning for non-English speakers. It is not

possible to match English speakers and non-English speakers because the logical reasoning test measures different latent variables for these two groups of test-takers. Only when the matching variable consists exclusively of DIF-free items, so the latent variable has the same meaning across groups, can it be used to match test-takers from different groups for subsequent DIF detection of the other items.

This logic is not as widely recognized as one would expect. Consider the following scenario: The item responses for each group are generated from the Rasch (1980) model with a mean item difficulty of zero for each group, but every item has different difficulties across groups (i.e., all items have DIF). Then, simulated data are analyzed with the Rasch model separately (one group at a time) or jointly, with the constraint that the mean item difficulty is zero for each group, which is the default model identification for many computer programs, such as ConQuest and BILOG-MG. In other words, the data-generating Rasch model together with the correct constraint for model identification is fit to the simulated data. Under such an ideal situation, the item and person parameters can be recovered well. Then, one can test every item for evidence of DIF using the Wald test if the standard errors are estimated accurately (Wright & Stone, 1979) or the likelihood ratio test for two nested models (Thissen et al., 1988). If the sample size is sufficiently large, then the null hypothesis of no DIF will be rejected. As a result, all items will be deemed to have DIF, which is identical to the generating setup so the DIF detection is perfect. Thus, it is tempting to conclude that matching variables is not necessary and DIF detection is accurate even when all items have DIF. In addition, because the person parameters are recovered accurately, it is also tempting for practitioners to compare person measures across groups.

Assuming the mean item difficulty is equal across groups, referred to as the equal-mean-difficulty (EMD) method, is not uncommon in DIF studies (Wang, 2004, 2008). In this method, each group has its own scale so the latent variable measured by the test is different for different groups, just like the logical reasoning test example. There is no way to match test-takers from different groups. A major drawback of the EMD method is that the assumptions made are in most cases wrong when one item on the studied test has DIF because the mean item difficulty will not be equal across groups. Even when there are multiple DIF items, it is very unlikely that their DIF magnitudes will cancel out exactly across groups so that the mean item difficulty is equal across groups.

There are other methods that are more appropriate than the EMD method. The all-other-item (AOI) method is an example. All items except the studied item on the test are assumed to be DIF-free and thus serve as the basis for a matching variable, and the studied item is evaluated for DIF (Wang, 2004, 2008). This procedure repeats until all items on the test have been evaluated for DIF. Unfortunately, the AOI method is theoretically correct only when the test does not have a DIF item (the test is perfect) or when the studied item is the only DIF item on the test. To make the AOI method more applicable for real (imperfect) tests, scale purification procedures are advocated. Specifically, when all items have been evaluated for DIF with the AOI method, and a few items have been identified as having DIF while the other items are DIF-free, we then use these presumably DIF-free items as the

matching variable and examine all items on the test again. This procedure repeats until the same set of items is detected as having DIF at two successive iterations or a maximum number of iterations is reached. This purified AOI method, denoted as AOI-P, outperforms the AOI method when tests have multiple DIF items, but the AOI-P begins to lose control of Type I error rates (false positive rates) when tests consist of a high percentage (say, 20% or higher) of DIF items (French & Maller, 2007; Wang, Shih, & Yang, 2009; Wang & Su, 2004).

Because scale purification procedures such as the AOI-P method cannot completely purify the matching variable to yield good DIF detection when tests have many DIF items, other methods should be pursued. The DIF-free-then-DIF (DFTD) strategy (Wang, Shih, & Sun, 2012) is such an attempt, in which DIF detection involves two steps: (a) identify a set of items that are most likely to be DIF-free and (b) use these presumably DIF-free items as the matching variable and evaluate the other items for DIF. This strategy can be applied to any DIF detection method. For example, we can adopt the AOI-P method to identify a set of items that are most likely DIF-free, use them to match test-takers from different groups, and evaluate the other items for DIF. In a series of simulations, the DFTD strategy demonstrates its superiority in DIF detection over traditional methods, such as the AOI and the AOI-P, especially when the tests have a high percentage of DIF items (Chen, Chen, & Shih, 2014; Chen & Hwu, 2017; Wang et al., 2012).

18.2 Review of DCM-DIF

A DCM can provide a fine-grained profile of multiple latent binary attributes to support learning and teaching. Popular DCMs include the deterministic inputs, noisy and gate (DINA) model (Junker & Sijtsma, 2001), the deterministic input, noisy or gate (DINO) model (Templin & Henson, 2006), and the (reduced) reparameterized unified model (RUM; Hartz, 2002). General forms of DCMs have also been proposed, such as the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), the generalized DINA model (de la Torre, 2011), and the general diagnostic model (GDM; von Davier, 2008).

Recently, several DIF detection methods have been developed. Zhang (2006) fit the DINA model to simulated data generated from the DINA model with or without DIF items, obtained the attribute profile estimate for each test-taker, and compared the performances in DIF detection of two methods: (a) the standard MH method with the total score as matching variable and (b) the MH method with the estimated attribute profile as matching variable (denoted as the profile MH [PMH] method). Zhang (2006) found the PMH method outperformed the standard MH method. Similar results were found by Zhang (2006) when the simultaneous item bias test method was compared with the raw score and the estimated attribute profile as the matching variables.

Li (2008) modified the higher-order DINA model (de la Torre & Douglas, 2004) to simultaneously detect differential functioning at the item and attribute levels, by

adding a group indicator at the attribute level while allowing for differences in item parameters across groups at the item level. Li (2008) found the method performed well in controlling Type I error rates at the item level but poorly at the attribute level when the tests included 20% DIF items. Bozard (2010) adapted methods of testing measurement invariance in confirmatory factor analysis models into the LCDM by testing the invariance of the intercepts, loadings of the main effects and interaction effects as well as the residual variance, but the performance of the method were not evaluated with simulations. Hou, de la Torre, and Nandakumar (2014) utilized the Wald test to compare differences in the item parameters that were obtained by fitting the DINA model to the reference and focal groups separately. Results showed that the method yielded well-controlled Type I error rates and high statistical power, even when all items on the test had DIF, although the method's performance decreased when guessing and slip parameters increased.

Hou et al. (2014) focused on DIF detection between two groups (e.g., gender). More recently, Li and Wang (2015) developed a flexible approach to detecting DIF based on the LCDM to accommodate various DCMs (e.g., DINA or DINO), more than two groups of test-takers, and multiple grouping variables that are categorical, continuous, observed, or latent (e.g., strategy usage). Li and Wang (2015) replaced the Wald statistic used by Hou et al. (2014), which is often inappropriate because of inaccurate estimation of standard errors, with the Bayesian Markov chain Monte Carlo method (MCMC) and observed good parameter recovery and superiority in DIF detection to the Wald method in a series of simulation studies.

In these previous studies, all items were simulated from the same DCM (e.g., the DINA model), the reference and focal groups shared the same Q-matrix, and DIF was simulated by specifying different guessing and/or slip parameters for different groups. Then, the simulated data were analyzed jointly for all test-takers (assuming all items had DIF) or separately for each group, based on the data-generating DCM and the correctly specified Q-matrix. These settings ensure that the model parameters are identifiable (Chiu, Douglas, & Li, 2009; Xu & Zhang, 2016). Because different groups are assumed to share the same DCM and Q-matrix, this method is referred to as the equal model and Q-matrix (denoted as EMQ) method. It was observed that the item parameters and person attributes were recovered well and DIF detection was good even when most or all items had DIF (Hou et al., 2014; Li & Wang, 2015). Given these excellent results, it is tempting to conclude that no matching variable or scale purification is needed, and the person attributes are comparable across groups. This is exactly what happens in the EMD method in IRT-DIF.

18.3 New Methods in DCM-DIF

There are many conditions in which Eq. 18.1 will not hold. Allowing different groups to have different guessing and/or slip parameters is one possibility. Other possibilities include that different groups have different Q-matrices or follow

different measurement models. For demonstration, we adopt the LCDM because it not only includes many DCMs as special cases but also allows for further model generalization in the logistic framework. Nevertheless, the new DIF detection methods proposed in this study can apply easily to other general DCMs, such as the generalized DINA model (de la Torre, 2011) and the GDM (von Davier, 2008).

Let $\boldsymbol{\alpha}_n^{\mathrm{T}} = (\alpha_{n1}, \ldots \alpha_{nK})$ be the attribute profile for person n, and P_{ni1} and let P_{ni0} denote the probabilities of scoring 1 and 0 for person n on item i, respectively. Under the LCDM, the log-odds of scoring 1 over scoring 0 are defined as follows:

$$\text{logit}\left(P_{ni}\right) \equiv \log\left(P_{ni1}/P_{ni0}\right) = \lambda_{i,0} + \boldsymbol{\lambda}_i^{\mathrm{T}}\mathbf{h}\left(\boldsymbol{\alpha}_n, \mathbf{q}_i\right), \tag{18.2}$$

where $\lambda_{i,0}$ defines the probability of success for persons who have not mastered any of the attributes required by item i; $\boldsymbol{\lambda}_i^{\mathrm{T}}$ is a vector of weights for item i with a length of $2^K - 1$; K is the number of latent binary attributes; q_{ik} $(k = 1, \ldots, K)$ is the entry for item i in the Q-matrix, indicating whether attribute k is required to answer item i correctly; $\mathbf{q}_i$ is a collection of q_{ik}; $\mathbf{h}(\boldsymbol{\alpha}_n, \mathbf{q}_i)$ is a set of linear combinations of α_n and $\mathbf{q}_i$; and $\boldsymbol{\lambda}_i^{\mathrm{T}}\mathbf{h}\left(\boldsymbol{\alpha}_n, \mathbf{q}_i\right)$ can be expressed as

$$\boldsymbol{\lambda}_i^{\mathrm{T}}\mathbf{h}\left(\boldsymbol{\alpha}_n, \mathbf{q}_i\right) = \sum\nolimits_{k=1}^{K} \lambda_{ik}\left(\alpha_{nk}q_{ik}\right) + \sum\nolimits_{k=1}^{K}\sum\nolimits_{v>k} \lambda_{ikv}\left(\alpha_{nk}\alpha_{nv}q_{ik}q_{iv}\right) + \cdots, \tag{18.3}$$

which includes an intercept, all the main effects of the attributes, and all possible interaction effects between the attributes, presenting compensatory and noncompensatory combinations. By constraining some of these effects, many common DCMs can be formed.

Below, we identify several conditions in which Eq. 18.1 does not hold. First, as simulated in previous studies (Hou et al., 2014; Li & Wang, 2015; Zhang, 2006), DIF occurs when different groups have different item parameters (e.g., the guessing and/or slip parameters in DINA model). For simplicity, this type of DIF is denoted as GS-DIF in this study. To formulate GS-DIF in the LCDM framework, the log-odds can be defined as follows:

$$\text{logit}\left(P_{nig}\right) = \lambda_{i,0,g} + \boldsymbol{\lambda}_{ig}^{\mathrm{T}}\mathbf{h}\left(\boldsymbol{\alpha}_{ng}, \mathbf{q}_i\right), \tag{18.4}$$

where g stands for group membership; the other variables were defined previously. Note that $\mathbf{q}_i$ does not have a subscript g, suggesting the q elements of item i are identical across groups.

Second, Eq. 18.1 does not hold when different groups have different Q-matrices (more specifically different q elements on the studied item), which is referred to as Q-DIF. Consider the following analogical reasoning item:

strawberry: red ≡
(A) peach: ripe
(B) leather: brown

(C) grass: green
(D) orange: round
(E) lemon: yellow

The correct answer is (E). For test-takers in the majority group, this item may measure three attributes: (a) encoding, (b) inference, and (c) mapping. Because some test-takers in the minority group have never seen yellow lemons, option (E) is not an option to them. For the minority group, this item requires not only the three attributes but also the attribute "life experience."

Another example is an arithmetic item measuring fraction subtraction (de la Torre & Douglas, 2004):

$$4\frac{4}{12} - 2\frac{7}{12} =$$

Two strategies can be used to solve this item. For strategy 1, the following four attributes are required: separating a whole number ($4\frac{4}{12} - 2\frac{7}{12} = 2\frac{4}{12} - \frac{7}{12}$), borrowing one from a whole number ($2\frac{4}{12} = 1\frac{16}{12}$), basic operation ($1\frac{16}{12} - \frac{7}{12}$), and simplifying ($1\frac{9}{12} = 1\frac{3}{4}$). For strategy 2, the following three attributes are required: converting a mixed number ($4\frac{4}{12} - 2\frac{7}{12} = \frac{52}{12} - \frac{31}{12}$), basic operation ($\frac{52}{12} - \frac{31}{12} = \frac{21}{12}$), and simplifying ($\frac{21}{12} = 1\frac{9}{12} = 1\frac{3}{4}$).

In these two examples, different attributes are required for different strategy groups, and the items are said to have Q-DIF. To create Q-DIF in the LCDM framework, the log-odds can be defined as follows:

$$\text{logit}\left(P_{nig}\right) = \lambda_{i,0,g} + \boldsymbol{\lambda}_{ig}^{\mathrm{T}}\mathbf{h}\left(\boldsymbol{\alpha}_{ng}, \mathbf{q}_{ig}\right), \tag{18.5}$$

where $\mathbf{q}_{ig}$ has a subscript g, indicating different groups have different q elements on studied item i. In this study, we focus on the case in which the focal group requires more attributes than the reference group to answer the studied item correctly. The additional attributes can be internal (already included in the Q-matrix, denoted as IQ-DIF) or external (not included in the Q-matrix, denoted as EQ-DIF).

Third, DIF can occur when different groups follow different measurement models on the studied item, which is denoted as MM-DIF. For example, the reference group may follow the DINA model while the focal group follows the DINO model or the RUM. In the LCDM framework, MM-DIF can be defined as follows:

$$\text{logit}\left(P_{nig}\right) = \lambda_{i,0,g} + \boldsymbol{\lambda}_{ig}^{\mathrm{T}}\mathbf{h}_g\left(\boldsymbol{\alpha}_{ng}, \mathbf{q}_{ig}\right), \tag{18.6}$$

where $\mathbf{h}_g$ has a subscript g, indicating different groups have different functions to combine α_n and $\mathbf{q}_i$, resulting in different measurement models.

To detect GS-DIF, the EMQ method is ideal as shown in previous studies. Because misspecification of the Q-matrix leads to biased parameter estimates and poor classification of person attributes (Kunina-Habenicht, Rupp, & Wilhelm, 2012;

Madison & Bradshaw, 2015; Rupp & Templin, 2008), it is expected the EMQ method will be inappropriate for detecting IQ-DIF, EQ-DIF, and MM-DIF. In practice, we are seldom (if ever) sure which type of DIF the studied item may have; therefore, it is desirable to develop methods that are sensitive to any type of DIF.

In IRT-DIF, when a test has DIF items, it measures different latent variables for different groups so that the latent variables are not comparable across groups. Therefore, a clean matching variable is essential in DIF detection, scale purification is advocated, and the DFTD strategy is helpful. This study applies the same logic and adapts those scale purification procedures and the DFTD strategy to DCM-DIF.

We focus on the PMH method in this study. Because the latent variables in most DCMs are binary attributes and every test-taker belongs to one of the possible attribute profiles, DIF occurs when test-takers have the same attribute profiles but belong to different groups and have different probabilities of success of endorsement on the studied item. To statistically test the null hypothesis of no DIF on the studied item, the PMH method appears feasible. Specifically, a DCM is fit to the data to obtain an attribute profile for each test-taker. The MH statistic on the studied item is computed by using attribute profiles to match test-takers. If the statistic is statistically significant, then the studied item is deemed to exhibit DIF. The standard PMH method (denoted as PMH-S), in which no scale purification is implemented, should be appropriate when tests do not have a DIF item (Zhang, 2006) or when the studied item is the only item on the test which might have DIF. However, empirical results indicate that most real tests usually have multiple DIF items. Therefore, we propose to implement the following scale purification procedure in the PMH method:

1. When detecting the studied item (e.g., item 1), assume all the other items are DIF-free and fit a DCM (e.g., DINA model) to all the other items to obtain an attribute profile for each test-taker. Compute the MH statistic for the studied item and conduct hypothesis testing for evidence of DIF.
2. Repeat step 1 until all items have been checked for DIF. (Steps 1 and 2 are the PMH-S method.)
3. Fit the DCM to those presumably DIF-free items found previously and obtain an attribute profile for each test-taker. Compute the MH statistic using the obtained attribute profiles to match test-takers from different groups and evaluate all items for DIF.
4. Repeat Step 3 until the same set of items is deemed to have DIF at two successive iterations or a maximum number of iterations (say, 10) is reached.

This purified PMH method, denoted as PMH-P, is similar to the AOI-P method in IRT-DIF, except that the attribute profile serves as the matching variable in the PMH method.

As documented in the IRT-DIF literature, scale purification procedures are helpful but begin to lose control of Type I error rates when the tests consist of a high percentage of DIF items. The DFTD strategy was developed to solve the problem. In DCM-DIF, this strategy can be implemented as follows:

1. Use the PMH-P method to classify items into two categories: DIF items and DIF-free items.
2. Among those classified as DIF-free items, select a subset of items (e.g., a fixed number, say 10 items, or a fixed percentage, say 50%) which have the largest p values for the MH statistic (meaning that these items are less likely to have DIF). Fit a DCM to these selected DIF-free items to obtain an attribute profile for each test-taker. Compute the MH statistic using the obtained attribute profiles to match test-takers from different groups and detect the other items for DIF.

This PMH method together with the DFTD strategy, denoted as PMH-D, should outperform the PMH-S and the PMH-P when tests contain a high percentage of DIF items.

18.4 An Illustration Using Simulated Data

We conducted a small-scale simulation study to illustrate the four types of DIF (GS-, IQ-, EQ-, and MM-DIF) and compare the performances in DIF detection of various PMH-based methods and others.

18.4.1 Design

The test had 30 items measuring five attributes. To facilitate estimation, the popular and parsimonious DINA model, which is a special case of LCDM, was used to generate item responses. The specifications of the Q-matrix were similar to those in Li and Wang (2015). Both groups had a sample size of 500 test-takers. Previous studies (de la Torre & Lee, 2010; Hou et al., 2014; Li, 2008) showed the item parameter for the DINA model can be well recovered with this sample size. Two independent variables were manipulated: (a) number of DIF items: 0, 3, 6, and 12 DIF items, representing 0%, 10%, 20%, and 40% DIF items on the test, respectively; (b) DIF type: GS-, IQ-, EQ-, and MM-DIF. For GS-DIF, similar to the settings in Li and Wang (2015), the guessing and slip parameters for the focal group were set 0.1 or 0.2 larger than those for the reference group. For IQ-DIF, the focal group required one additional attribute compared to the reference group to answer the DIF items correctly, and the additional attribute was part of the Q-matrix. For EQ-DIF, the additional attribute was outside the Q-matrix. For MM-DIF, the DIF items for the focal group followed the RUM (Hartz, 2002), which can be expressed as follows:

$$P_{nic} = \left[\pi_i \prod_{k=1}^{K} \omega_{ik}^{(1-\alpha_{nk})q_{ik}}\right] \times \frac{\exp(\gamma_c + \delta_i)}{1 + \exp(\gamma_c + \delta_i)}, \tag{18.7}$$

where P_{nic} is the probability of success on item i for person n with attribute profile c, π_i is the baseline probability of success on item i when all required attributes have been mastered, ω_{ik} is the penalty for the probability of success on item i for not mastering attribute k, γ_c is the latent variable for attribute profile c, and δ_i is the easiness parameter of item i for the attributes that are not indicated in the Q-matrix; the other variables have already been defined. In this study, we set $\pi_i = 0.9$ for item i, $\omega_{ik} = 0.2$ for attribute k in item i, and γ_c and δ_i as generated from $N(0, 1)$.

Five PMH-based methods were used to detect DIF: (a) the PMH-S method, (b) the PMH-P method in which scale purification procedures were implemented, (c) the PMH-D method in which the DFTD strategy was incorporated and 50% of items that had the largest p values for the MH statistic in the first step were selected as anchor items, (d) the optimal PMH method (denoted as PMH-O) in which all DIF-free items by design were used as anchors, and (e) the EMQ method.

To implement the PMH-S, PMH-P, PMH-D, and PMH-O methods, the CDM package in R was used to analyze the item responses to obtain attribute profiles for the test-takers, and the difMH function in the difR package in R was modified to allow using the attribute profiles as matching variable to test whether the studied item had DIF at the .05 nominal level. The customized R code is available upon request. To implement the EMQ method, the item responses were estimated with the freeware Just Another Gibbs sampler (JAGS; Plummer, 2003), which implements Bayesian MCMC methods. DIF detection was conducted by checking whether the 95% credible intervals of the DIF magnitude in the item parameters contained zero. A total of 100 replications were conducted. The dependent variables were Type I error rates and power rates of DIF detection.

It was expected that all methods would perform well when the test did not have a DIF item. When there were DIF items, the PMH-O method would perform the best because it was the true model. When the test had a high percentage of DIF items, the PMH-D method would outperform the PMH-P and PMH-S methods because it used a presumably clean matching variable. The PMH-P would perform well when tests did not contain a high percentage of DIF items. The PMH-S would perform well when tests had only a few DIF items. The EMQ method would perform well in detecting GS-DIF but poorly in detecting other types of DIF.

18.4.2 Results

Due to space constraints, we are not able to show the Type I error rates and power rates for individual items. Instead, we present the mean Type I error rate across all DIF-free items and the mean power rate across all DIF-items in Table 18.1. In general, the results confirm our expectations. Specifically, all methods controlled the Type I error rates at the .05 nominal level when there was no DIF item. The PMH-O, PMH-D, and PMH-P methods yielded well-controlled Type I error rates and higher power rates under all conditions in detecting IQ-DIF, EQ-DIF and MM-DIF, and the PMH-O method had the highest power rates, followed by the PMH-D method

Table 18.1 Mean type I error rates (%) and mean power rates (%) in the simulation study

DIF		PMH-O		PMH-D		PMH-P		PMH-S		EMQ	
Percentage	DIF Type	Type I	Power	Type I	Power	Type I	Power	Type I	Power	Type I	Power
0		4.9		4.2		3.7		3.9		5.2	
10	GS-DIF-0.1	5.1	78.0	4.0	77.0	4.8	74.0	5.1	76.3	8.2	93.7
	GS-DIF-0.2	4.1	99.3	3.7	99.3	3.9	99.1	4.2	99.0	8.2	100.0
	IQ-DIF	4.4	100.0	5.9	100.0	4.6	100.0	6.5	100.0	9.7	100.0
	EQ-DIF	5.1	100.0	4.7	100.0	4.7	100.0	5.9	99.6	8.6	100.0
	MM-DIF	4.4	94.7	3.9	92.0	4.1	90.0	4.5	89.7	7.9	99.7
20	GS-DIF-0.1	5.4	57.6	4.5	56.8	5.4	53.7	5.4	53.5	8.3	93.2
	GS-DIF-0.2	4.5	83.8	5.3	85.2	3.8	84.0	4.2	83.5	8.9	100.0
	IQ-DIF	3.8	100.0	4.7	100.0	4.1	100.0	9.5	99.2	8.4	100.0
	EQ-DIF	4.3	100.0	5.1	100.0	4.1	100.0	6.3	99.8	9.4	100.0
	MM-DIF	4.9	83.3	4.2	82.3	4.5	84.0	7.3	83.5	8.4	91.5
40	GS-DIF-0.1	4.4	49.7	5.3	48.5	4.5	48.1	4.6	48.4	8.1	92.4
	GS-DIF-0.2	4.1	67.3	4.9	66.8	4.9	66.8	4.6	66.7	8.6	99.9
	IQ-DIF	4.6	98.7	5.6	94.3	5.6	92.3	16.7	45.3	18.5	99.2
	EQ-DIF	3.9	98.2	5.6	95.6	5.2	94.5	18.6	37.4	16.4	99.4
	MM-DIF	5.0	89.3	5.2	88.3	5.1	87.2	13.8	81.2	9.7	90.3

Note: *PMH-O* = PMH method using all DIF-free items by design as anchors, *PMH-D* = PMH method with the DIF-free-then-DIF strategy, *PMH-P* = PMH method with scale purification procedures, *PMH-S* = standard Mantel-Haenszel method using the attribute profile as the matching variable, *EMQ* = equal model and Q-matrix method. *GS-DIF-0.1/0.2*: DIF occurs in guessing and slip parameters in DINA model with DIF size of 0.1/0.2, *IQ-DIF*: DIF occurs in the Q-matrix where the additional attribute for focal group is already included in the Q-matrix, *EQ-DIF*: DIF occurs in the Q-matrix where the additional attribute for focal group is not included in the Q-matrix, *MM-DIF*: DIF occurs when different groups have different DCMs

and then the PMH-P method. Because the differences in the power rates among the three methods were not large, it appeared that they performed very similarly under all conditions. In contrast, the PMH-S began to yield inflated Type I error rates and deflated power rates when tests contained 20% or more DIF items. For example, in the detection of IQ-DIF, when tests had 10%, 20%, and 40% DIF items, the PMH-S yielded mean Type I error rates of 6.5%, 9.5%, and 16.7%, respectively, and mean power rates of 100.0%, 99.2%, and 45.3%, respectively. It appeared that the larger the inflation in the Type I error rates, the larger the deflation in the power rate.

In terms of GS-DIF which was the focus of previous studies (Hou et al., 2014; Li & Wang, 2015; Zhang, 2006), it was found the four new PMH-based methods (i.e., the PMH-O, the PMH-D, the PMH-P and the PMH-S) yielded well-controlled Type I errors in all conditions. The larger DIF effect, the higher power rates of detection. For example, when tests contained 10% DIF items, the PMH-D methods yielded mean power rates of 77.0% and 99.3%, respectively, for DIF size of 0.1 and 0.2. However, the power rates decreased as the DIF percentage increased. For instance, when the DIF percentage increased to 40%, the mean power rates of PMH-D methods decreased to about 48.5% and 66.8%, respectively, for DIF size of 0.1 and 0.2.

As anticipated, the EMQ method performed well in detecting GS-DIF, in terms of only a slight inflation on Type I error rates (between 8.1% and 8.9% for 10–40% DIF items) and high power rates (between 92.4% and 100%). The power rates were much higher than those yielded by the other methods, which was because the EMQ method directly modeled the difference in the guessing and slip parameters between groups, whereas the other methods obtained the attribute profiles and used them to assess DIF indirectly. However, the EMQ method performed poorly in detecting other types of DIF. For example, in detection of IQ-DIF, the EMQ yielded mean Type I error rates of 9.7%, 8.4%, and 18.5% when tests had 10%, 20%, and 40% DIF items, respectively. Because the Type I error rates were seriously inflated, the corresponding high power rates were meaningless.

18.5 Discussion and Conclusions

Many DIF detection methods have been developed and evaluated for use in IRT. Because real tests often contain multiple DIF items, the biggest challenge in IRT-DIF is to find a clean matching variable to match test-takers from different groups. Scale purification procedures and the DFTD strategy appear effective in overcoming this challenge. Although research on DCM has increased rapidly in recent years, few studies have been conducted on DCM-DIF. One reason for this phenomenon is that the EMQ method, without the necessity of matching variables, has been found to be very promising for detecting DIF, even when many or all items on a test have DIF. It appears that the troublesome challenge in IRT-DIF does not exist in DCM-DIF. Similar to the EMD method in IRT-DIF, the EMQ method in DCM-DIF performs well only when the true model and the correct model constraints are implemented, which can seldom (if ever) happen in practice. Thus, the troublesome challenge remains unresolved in DCM-DIF.

Many reasons can cause DIF. That different groups have different guessing and/or slip parameters but an identical DCM and Q-matrix (GS-DIF) is only one possibility. Other possibilities include different groups have different Q-matrices or follow different DCMs. It is desirable to develop DIF detection methods that are sensitive to all types of DIF. In this study, we identified four types of DIF (GS-, IQ-, EQ-, and MM-DIF) and proposed four new PMH-based methods, including the PMH-S, PMH-P, PMH-D, and PMH-O methods.

We conducted a small-scale simulation study to demonstrate the superiority of the PMH-P and PMH-D methods when tests have many DIF items and different types of DCM-DIF. For comparison, we also adopted the EMQ method and the PMH-O method. The simulation study showed that (a) all methods performed well when the tests were perfect, (b) the EMQ method performed well only for GS-DIF, (c) the PMH-P method outperformed the PMH-S method, indicating scale purification procedures were helpful, (d) the PMH-D method outperformed the PMH-P method when the tests had 40% DIF items, indicating the advantages of the DFTD strategy in high percentages of DIF items, and (e) the PMH-O performed

very well under all conditions, indicating the necessity of a long and pure matching variable. In sum, when detecting IQ-, EQ-, and MM-DIF, the PMH-P and PMH-D methods outperformed the EMQ method in controlling type I error rate and yielding high power. When detecting GS-DIF for which the EMQ method performed very well, it appeared that the larger the DIF size, the higher power of the new methods. When the test contains 10% DIF items and the DIF size is 0.2, the new methods yielded type I error rate close to the nominal level and as high power rate as the EMQ method. In this study, the EMQ method (implemented with JAGS) took approximately 1 hour on average to converge for 30 items and 500 persons in each group; whereas the proposed PMH-based methods (implemented with customized R) took only seconds. Thus, the new methods are more suitable for practical use.

This study is not without limitations. The simulation study did not cover a large number of different conditions. Future studies should evaluate the new methods under more conditions, such as different DCMs, Q-matrices, test lengths, sample sizes, and DIF magnitudes. This study opens up several research lines that can facilitate a deeper understanding of DCM-DIF. First, we focused on adapting the MH method to DCM-DIF. As a competitor of the MH method, the LR method that is commonly used in IRT-DIF can be adapted, in which raw total scores are replaced with attribute profiles to match test-takers. How these new LR methods perform in DCM-DIF deserves further investigation. Second, there was only one grouping variable with two categories in this study. As studied by Li and Wang (2015), there may be multiple variables underlying DIF with more than two categories, and these variables can be not only categorical but continuous (e.g., age) or latent (e.g., problem solving strategy). Future studies should adapt the new methods to accommodate these more complex situations. Third, the methods focused on modeling DIF effects in dichotomous items. It is of great interest to assess DIF in polytomous items. The GDM (von Davier, 2008) which is not only applicable for both binary and polytomous responses, but also include many common DCMs as special cases (von Davier, 2013, 2014), can be used as the framework to assess DIF in the future. Finally, we identified only four types of DIF in this study. Future studies should aim to identify other types of DIF and evaluate how the new methods will perform in detecting these types.

Acknowledgement This study was sponsored by the General Research Fund, Research Grants Council (No. 18604515).

References

Bozard, J. (2010). *Invariance testing in diagnostic classification models (Unpublished master thesis)*. Athens, Georgia: The University of Georgia.

Chen, C.-T., & Hwu, B.-S. (2017). Improving the assessment of differential item functioning in large-scale programs with dual-scale purification of Rasch models: The PISA example. *Applied Psychological Measurement*, 1–15. https://doi.org/10.1177/0146621617726786

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement, 38*, 18–36. https://doi.org/10.1177/0146621613488643

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633–665.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

de la Torre, J., & Lee, Y. S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*, 115–127.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373–393. https://doi.org/10.1177/0013164406294781

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation*. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–201. https://doi.org/10.1007/s11336-008-9089-5

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis modeling: Applying Wald test to investigate DIF for DINA model. *Journal of Educational Measurement, 1*, 98–125. https://doi.org/10.1111/jedm.12036

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272. https://doi.org/10.1177/01466210122032064

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*, 59–81. https://doi.org/10.1111/j.1745-3984.2011.00160.x

Li, F. M. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning. Unpublished doctoral dissertation*. Athens, Georgia: University of Georgia.

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*, 28–54. https://doi.org/10.1111/jedm.12061

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Madison, M. J., & Bradshaw, L. P. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement, 75*, 491–511. https://doi.org/10.1177/0013164414539162

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute, 22*, 719–748.

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (Expanded edition)*. Chicago, IL: University of Chicago Press. (Original work published 1960).

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96. https://doi.org/10.1177/0013164407301545

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 147–169). Hillsdale, NJ: Lawrence Erlbaum.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM) (research report no. RR-14-40)*. Princeton, NJ: Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1109308.pdf

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221–261. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement, 9*, 387–408.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72*, 687–708. https://doi.org/10.1177/0013164411426157

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713–731. https://doi.org/10.1177/0013164409332228

Wang, W.-C., & Su, Y.-H. (2004). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450–480. https://doi.org/10.1177/0146621604269792

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*, 625–649. https://doi.org/10.1007/s11336-015-9471-z

Zhang, W. (2006). *Detecting differential item functioning using the DINA model. Unpublished doctoral dissertation*. Carolina, NC: University of North Carolina at Greensboro.

Chapter 19
Bifactor MIRT as an Appealing and Related Alternative to CDMs in the Presence of Skill Attribute Continuity

Daniel M. Bolt

Abstract For virtually all tests analyzed using CDMs, low-dimensional compensatory item response theory (IRT) models with continuous abilities appear to provide an equivalent or better statistical fit, as noted in a recent commentary by von Davier and Haberman (Psychometrika, 79:340–346, 2014). We examine these issues using both simulation and real data analyses. We suggest that the results motivate consideration of bifactor MIRT models as an attractive alternative for diagnostic measurement, particular in cases where skill attribute continuity is suspected or can be confirmed. The potential usefulness of bifactor MIRT for diagnostic scoring is also based on other considerations. For example, bifactor MIRT reflects a tendency for items to measure primarily one of the required conjunctively interacting skill attributes (the most difficult of the required attributes), and also makes it possible to address the estimation limitations of MIRT models of high dimensionality (Cai L, Psychometrika, 75(4):581–612, 2010). Additionally, the bifactor MIRT model uses orthogonal statistical dimensions, making it easier to quantify the incremental contribution provided by attending to specific factors that can provide the foundation for diagnosis.

The primary applications of CDMs are their use toward the scoring of item score patterns. As CDMs typically make use of a large number of discrete (typically binary) skills, the models often yield many potential skill attribute patterns for diagnostic classification. At the same time, a recent commentary by von Davier and Haberman (2014) notes that for virtually all tests analyzed using CDMs, low-

D. M. Bolt (✉)
Department of Educational Psychology, University of Wisconsin – Madison, Madison, WI, USA
e-mail: dmbolt@wisc.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_19

dimensional compensatory item response theory (IRT) models with continuous abilities appear to provide an equivalent or better statistical fit. Such findings naturally raise questions regarding CDM assumptions about the discrete nature of skill attribute mastery/nonmastery, as well as the meaningfulness of skill attributes as statistical dimensions in the data. In this chapter we examine the possibility that these findings might also be attributed to the tendency for items that in theory require multiple conjunctively interacting skills to in actuality only statistically discriminate with respect to a small number (perhaps as few as one) of such attributes. Such an explanation seems especially likely in the presence of a higher-order factor structure underlying the attribute correlations, where the mastery status of the "most difficult" (i.e., least frequently mastered) of the required skill attributes is often the most informed by performance on an item.

The potential for a low dimensional compensatory IRT model to approximate a high dimensional non-compensatory model is informed by a previous study by Bolt and Lall (2003). Non-compensatory IRT models typically associate a separate difficulty parameter with each ability dimension in the data (see e.g., Sympson, 1978; Whitely, 1980; Embretson, 1984). Bolt and Lall (2003) demonstrated that when a non-compensatory interaction exists between continuous and positively correlated abilities, the relative difficulties across ability dimensions in the non-compensatory model function analogously to how item discrimination parameters function in compensatory models. Specifically, for a given item, dimensions associated with higher difficulty in a non-compensatory model are more discriminating (i.e., have higher loadings) for comparable dimensions in a compensatory model. Certain necessary skills for an item in a non-compensatory model are often only trivially, if at all, measured by the item when portrayed in a compensatory model. A similar phenomenon may well be occurring in CDMs.

In this chapter, we examine these issues using both simulation and real data analyses. We suggest that the results motivate consideration of bifactor MIRT models as an attractive alternative for diagnostic measurement, particular in cases where skill attribute continuity is suspected or can be confirmed. The potential usefulness of bifactor MIRT for diagnostic scoring is also based on other considerations. For example, bifactor MIRT makes it possible to address the estimation limitations of MIRT models of high dimensionality (Cai, 2010). Additionally, the bifactor MIRT model uses orthogonal statistical dimensions, making it easier to quantify the incremental contribution provided by attending to specific factors that can provide the foundation for diagnosis.

Using both simulation analysis and the frequently studied fraction subtraction dataset (Tatsuoka, 1990), we examine the nature of the student-level diagnostic information provided when bifactor MIRT is used as a basis for score reporting and compare its results against diagnoses provided by CDMs. We suggest that the results also simultaneously yield important practical suggestions related to the design of tests for purposes of diagnosis with CDMs.

19.1 The DINA and Higher-Order (HO-) DINA Models

For purposes of illustration, we consider the frequently studied DINA ("Deterministic Input Noisy And") models. We assume a Q-matrix where $q_{ik} = 1$ implies that item i (=1, ... I) requires skill attribute k (=1, ... K). Table 19.1 defines the item-attribute relationships for a fraction subtraction dataset (Tatsuoka, 1990) that is frequently analyzed using the DINA and related models and that will be used later in the chapter. The skill attribute assignments are shown for both the 20-item test and a 15-item subset that have been frequently analyzed in past work (de la Torre & Douglas, 2004; de la Torre & Lee, 2010).

For these data, the skill attribute mastery status of a student is typically represented as binary, where $\alpha_{jk} = 1$ denotes mastery and 0 non-mastery, and j (=1, ... J) indexes students. It then becomes possible to characterize a student's expected or "ideal" response to an item, denoted $\eta_{ij} = 0, 1$, as:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{jk}{}^{q_{ik}}$$

implying an expected correct response (i.e., $\eta_{ij} = 1$) if all attributes required for the item have been mastered by the student, and an incorrect response (i.e., $\eta_{ij} = 0$) otherwise. Unexpected ("Error") responses, either an incorrect response for an ideal

Table 19.1 Skill attributes required for the 15-item and 20-item fraction subtraction data analyses (Tatsuoka, 1990; de la Torre & Lee, 2010)

Item	20-item	15-item	Item	20-item	15-item
1	4,6,7	1	11	2,5,7	1,3
2	4,7	1,2,3,4	12	7,8	1,3,4
3	4,7	1	13	2,4,5,7	1,2,3,4
4	2,3,5,7	1,2,3,4,5	14	2,7	1,2,3,4,5
5	2,4,7,8	3	15	1,7	1,2,3,4
6	7	1,2,3,4	16	2,7	NA
7	1,2,7	1,2,3,4	17	2,5,7	NA
8	7	1,2	18	2,5,6,7	NA
9	2	1,3	19	1,2,3,5,7	NA
10	2,5,7,8	1,3,4,5	20	2,3,5,7	NA

Note: The items and attributes do not align across the 15- and 20-item datasets
20-item dataset attributes: *Attribute 1* = convert a whole number to a fraction, *Attribute 2* = separate a whole number from a fraction, *Attribute 3* = simplify before subtracting, *Attribute 4* = find a common denominator, *Attribute 5* = borrow from whole number part, *Attribute 6* = column borrow to subtract the second numerator from the first, *Attribute 7* = subtract numerators, and *Attribute 8* = reduce answers to simplest form
15-item dataset attributes: *Attribute 1* = basic fraction subtraction, *Attribute 2* = simplify/reduce fraction or mixed number, *Attribute 3* = separate whole number from fraction, *Attribute 4* = borrow from the whole number in a given mixed number, *Attribute 5* = convert a whole number to a fraction

correct or a correct response for an ideal incorrect, may occur. These events are accounted for by nonzero slip and guessing probabilities, respectively:

$$s_j = P\left(Y_{ij} = 0|\eta_{ij} = 1\right) \text{ and } g_j = P\left(Y_{ij} = 1|\eta_{ij} = 0\right),$$

such that the resulting probability of correct response on the item conditional upon the attribute mastery pattern is given by:

$$P\left(Y_{ij} = 1|\boldsymbol{\alpha}\right) = \left(1 - s_j\right)^{\eta_{ij}} {g_j}^{1-\eta_{ij}}.$$

The resulting "independence" DINA model, where no particular correlational structure is assumed among the attributes, contains two parameters per item. Added to these are parameters representing the proportion of students associated with each of the 2^K possible attribute mastery patterns, implying $2\ J + 2^K$ -1 parameters in total.

The Higher-Order DINA (HO-DINA) model constrains the DINA model by introducing a correlational factor structure among attributes across students. A probability of attribute mastery is defined in relation to a higher-order factor(s) θ as:

$$P\left(\alpha_k = 1|\boldsymbol{\theta}\right) = \frac{\exp\left(\lambda_{0k} + \lambda_{1k}\boldsymbol{\theta}\right)}{1 + \exp\left(\lambda_{0k} + \lambda_{1k}\boldsymbol{\theta}\right)}$$

where λ_{0k}, λ_{1k} define an intercept and slope vector related to attribute k. Relative to an independence model, the HO-DINA has been found to provide a better comparative fit to actual test data using information-based model comparison indices (de la Torre & Douglas, 2004), as skill attributes in most realistic settings are expected to correlate. The HO-DINA also substantially reduces the number of model parameters.

19.2 Simulation Study

As noted above, it might be theorized that in the presence of a higher-order factor against which attributes can be ordered in terms of difficulty, DINA items that theoretically require multiple attributes may in fact measure far fewer. We explore this issue initially through a small simulation study. Specifically, we simulate HO-DINA item response data from a series of conditions manipulated with respect to several factors: (1) the strength of a higher-order factor, as reflected by the magnitude of higher-order factor loadings; (2) the dispersion of attribute mastery thresholds, (3) the overall number of items; and (4) the slip/guessing probabilities.

Table 19.2 illustrates the Q matrix used in the simulation for a 15-item condition. We use a Q matrix in which each combination of the four attributes are assessed once across items. A 30-item condition was created by replicating the Q-matrix. A

Table 19.2 Q matrix, simulation study, 15-item condition

Item	Att1	Att2	Att3	Att4
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
5	1	1	0	0
6	1	0	1	0
7	1	0	0	1
8	0	1	1	0
9	0	1	0	1
10	0	0	1	1
11	1	1	1	0
12	1	1	0	1
13	1	0	1	1
14	0	1	1	1
15	1	1	1	1

higher-order factor is introduced in which λ_1 is manipulated as a factor at constant values of 1, 2, or 3 across attributes; the λ_0s were set at levels of -2, -1, 0, 1 across attributes, respectively; or at -1, $-.5$, 0, .5, so that the attributes were always increasing in difficulty from attribute 1 (easiest) to attribute 4 (hardest). As the λ_1 slope parameters relate the higher order factor to the binary latent skill attributes, the chosen values represent conditions of weak to strong higher order-dependence; in the fraction subtraction analysis reported by de la Torre and Douglas (2004), the slopes ranged from 1 to 5. The manipulation of the λ_0s reflects a high versus a low threshold dispersion condition.

Finally, the generating guessing/slip parameters are set equal across items but were varied as a factor at levels of .1/.1 and .25/.25.

To evaluate how well the items measure the individual skill attributes, a "hold-one-item-out" logistic regression approach was used in which for each item on the test, a separate HO-DINA is first fit to all but one studied item so as to estimate the attributes for each examinee. We then model performance on the studied item using the posterior probabilities of attribute mastery for each of the four attributes, as defined by HO-DINA, as predictors. The logistic regression coefficient estimates were recorded. We used 10 replications per condition. The resulting regression coefficients for each analysis reflect how well the item held out actually measures each of the attributes it is theorized to measure.

Tables 19.3 and 19.4 illustrate by item the mean logistic regression coefficient estimates with respect to each of the four attributes under conditions where $\lambda_1 = 2$ and the slip/guessing probabilities were .1/.1. We observed only small effects related to the λ_1 and slip/guessing parameter factors, so the resulting tables display effects associated with the other two manipulated factors (the number of items and threshold dispersion factors). Boldface identifies the most "difficult" (highest threshold) attribute of the attributes required by the item. Italics identify the

Table 19.3 Means (standard deviations) of estimated logistic regression coefficients predicting item scores as a function of HO-DINA based probability of attribute mastery, $\lambda_1 = 2$; $\lambda_0 = -2, -1, 0, 1$; slip/guess parameters = .1/.1 (10 replications)

	15 item				30 item			
Item	Att1 (SD)	Att2 (SD)	Att3 (SD)	Att4 (SD)	Att1 (SD)	Att2 (SD)	Att3 (SD)	Att4 (SD)
1	***1.32(.12)***	−0.08(.06)	−0.03(.03)	−0.02(.02)	***0.94(.05)***	−0.03(.02)	−0.02(.02)	0.00(.03)
2	−0.06(.02)	***1.05(.03)***	−0.08(.02)	−0.05(.02)	−0.01(.02)	***0.85(.01)***	−0.02(.02)	−0.01(.01)
3	−0.02(.02)	−0.03(.01)	***0.92(.02)***	−0.05(.02)	−0.02(.02)	0.00(.01)	***0.82(.02)***	−0.01(.01)
4	0.00(.01)	−0.01(.02)	−0.02(.01)	***0.93(.01)***	0.00(.02)	0.00(.02)	0.00(.02)	***0.81(.02)***
5	*0.18(.05)*	***0.50(.03)***	0.08(.02)	0.05(.02)	*0.29(.03)*	***0.66(.02)***	0.02(.02)	0.02(.02)
6	*0.10(.04)*	0.07(.03)	***0.59(.03)***	0.07(.02)	*0.11(.03)*	0.00(.02)	***0.78(.01)***	0.01(.02)
7	*0.05(.03)*	0.04(.02)	0.08(.03)	***0.52(.03)***	*0.03(.04)*	0.00(.01)	0.01(.01)	***0.79(.02)***
8	0.14(.02)	*0.18(.05)*	***0.52(.03)***	0.08(.02)	−0.03(.03)	*0.21(.03)*	***0.68(.01)***	0.03(.02)
9	0.03(.02)	*0.07(.03)*	0.08(.02)	***0.49(.03)***	−0.03(.03)	*0.06(.01)*	0.00(.01)	***0.75(.02)***
10	0.01(.04)	0.03(.02)	*0.17(.02)*	***0.40(.03)***	−0.02(.03)	−0.01(.02)	*0.14(.01)*	***0.59(.02)***
11	*0.05(.03)*	*0.18(.03)*	***0.53(.05)***	0.09(.03)	*0.02(.04)*	*0.21(.03)*	***0.66(.02)***	0.03(.01)
12	*0.01(.03)*	*0.04(.03)*	0.01(.01)	***0.75(.06)***	*0.01(.02)*	*0.07(.02)*	0.03(.01)	***0.71(.02)***
13	*−0.01(.04)*	−0.02(.02)	*0.14(.03)*	***0.61(.05)***	*−0.01(.04)*	−0.01(.03)	*0.14(.02)*	***0.59(.02)***
14	−0.01(.05)	*0.03(.04)*	*0.15(.03)*	***0.54(.06)***	−0.02(.03)	*0.01(.02)*	*0.13(.01)*	***0.57(.01)***
15	*−0.03(.02)*	*0.00(.04)*	*0.14(.02)*	***0.58(.07)***	*0.00(.02)*	*0.04(.02)*	*0.14(.02)*	***0.54(.02)***

Table 19.4 Means (standard deviations) of estimated logistic regression coefficients predicting item scores as a function of HO-DINA based probability of attribute mastery, $\lambda_1 = 2$; $\lambda_0 = -1, -.5, 0, .5$; slip/guess parameters = .1/.1 (10 replications)

	15 item				30 item			
Item	Att1 (SD)	Att2 (SD)	Att3 (SD)	Att4 (SD)	Att1 (SD)	Att2 (SD)	Att3 (SD)	Att4 (SD)
1	***1.18(.05)***	−0.12(.04)	−0.09(.03)	−0.08(.02)	***0.92(.02)***	−0.02(.01)	−0.02(.01)	−0.01(.01)
2	−0.03(.02)	***1.14(.06)***	−0.16(.04)	−0.14(.04)	−0.01(.02)	***0.88(.02)***	−0.03(.01)	−0.02(.02)
3	−0.03(.02)	−0.03(.02)	***1.07(.06)***	−0.19(.05)	−0.01(.01)	−0.01(.01)	***0.84(.01)***	−0.02(.01)
4	−0.04(.01)	−0.04(.02)	−0.05(.02)	***1.10(.07)***	−0.01(.02)	−0.01(.02)	−0.01(.02)	***0.82(.01)***
5	*0.19(.05)*	***0.26(.06)***	0.05(.03)	0.04(.03)	*0.34(.03)*	***0.57(.02)***	0.03(.01)	0.04(.02)
6	*0.16(.02)*	0.03(.02)	***0.29(.09)***	0.04(.03)	*0.22(.02)*	0.01(.02)	***0.66(.04)***	0.03(.02)
7	*0.12(.02)*	0.03(.02)	0.03(.02)	***0.28(.09)***	*0.14(.03)*	−0.01(.02)	0.01(.02)	***0.72(.02)***
8	0.04(.02)	*0.18(.02)*	***0.25(.08)***	0.04(.02)	0.00(.03)	*0.30(.02)*	***0.55(.02)***	0.04(.02)
9	0.04(.01)	*0.15(.02)*	0.05(.02)	***0.25(.09)***	0.00(.02)	*0.17(.03)*	0.01(.02)	***0.64(.03)***
10	0.03(.02)	0.03(.02)	*0.17(.03)*	***0.22(.05)***	−0.01(.02)	−0.01(.02)	*0.26(.02)*	***0.53(.02)***
11	*0.11(.02)*	*0.14(.03)*	***0.20(.08)***	0.04(.02)	*0.10(.02)*	*0.27(.02)*	***0.49(.01)***	0.05(.02)
12	*0.07(.02)*	*0.08(.01)*	0.00(.02)	***0.42(.08)***	*0.08(.03)*	*0.16(.02)*	0.02(.01)	***0.58(.02)***
13	*0.03(.03)*	−0.01(.01)	*0.23(.04)*	***0.34(.06)***	*0.05(.02)*	0.00(.01)	*0.24(.01)*	***0.49(.02)***
14	0.02(.04)	*0.11(.02)*	*0.14(.04)*	***0.25(.08)***	−0.02(.02)	*0.10(.02)*	*0.22(.02)*	***0.45(.01)***
15	*0.05(.04)*	*0.11(.03)*	*0.16(.03)*	***0.24(.07)***	*0.02(.02)*	*0.08(.02)*	*0.20(.02)*	***0.44(.02)***

attributes that were required for the item as defined by the Q matrix. Table 19.3 displays results under the high threshold dispersion condition; Table 19.4 the low dispersion condition.

From both tables, it is consistently observed that the required attributes with highest thresholds have the highest regression coefficients, and thus are best measured by each item. Along these lines, those items measuring a single attribute emerge by far as the most discriminating across items, a result that has been observed and discussed previously (e.g., Bradshaw & Madison, 2016). The differential discrimination between single-attribute and multiple-attribute items appears to decrease as the number of items increases. At the same time, the relative contribution of the most difficult attribute appears to increase as the number of items increases. These two effects appear relatively stable across the two threshold dispersion conditions, as seen in comparing the coefficients of Tables 19.3 and 19.4.

The results of the simulation have several potential implications. The first concerns the design of tests for CDMs, and the desirability of including for each attribute at least some items for which the attribute is the most difficult of the required skill attributes. Most of the information obtained for attribute classification appears to be due to such items. Second, despite the desirability of complex (i.e., multi-attribute) items in educational tests, the role such items play in diagnostic classification appears substantially diminished, especially for less difficult attributes. Third, the results clarify that certain attributes may as a whole be better measured than others, and that such results have a lot to do with the difficulty (threshold) of the attribute. The assumption that the presence of an attribute in the Q-matrix makes it a measured attribute is often questionable, as the requirement of an attribute by an item should not be taken to imply that the attribute is well-measured. Finally, and of greatest relevance to this chapter, it would appear not much is lost with respect to the diagnostic information provided by an item if it is statistically allowed to inform with respect to only one attribute. The majority of information provided by an item appears to occur with respect to the coefficients identified in bold in Tables 19.3 and 19.4. This suggests the potential to build a bifactor MIRT model as an effective alternative to CDMs, particularly where simultaneous concern may exist over the presence of continuously distributed skill attributes.

19.3 The Bifactor MIRT Model

The results observed in the simulation suggest the potential diagnostic value of a model that is able to attend to just one attribute per item as opposed to all implied by the Q-matrix. In MIRT applications, such conditions can be reflected in a bifactor MIRT model, where the probability of correct response to an item is given by:

$$P\left(U_j = 1|\theta_{GEN}, \theta_{SP1}, \ldots, \theta_{SPK}\right) = g\left(\alpha_{jGEN}\theta_{GEN} + \alpha_{jSP1}\theta_{SP1} + \cdots + \alpha_{jSPK}\theta_{SPK} + \gamma_j\right),$$

where $U_j = 1$ denotes a correct response to item j, θ_{GEN}, θ_{SP1}, ... , θ_{SPK} denote a general ability dimension and K specific (group) dimensions, α_{jGEN}, α_{jSP1}, ... α_{jSPK} are the corresponding discrimination parameters, γ_j is an item threshold parameter, and g defines a link function, typically a logistic or probit function. Under the bifactor MIRT model, each item has a nonzero discrimination ($\alpha_{jSP_*} \neq 0$) parameter for no more than one specific factor. Further, the θ_{GEN}, θ_{SP1}, ... , θ_{SPK} are mutually uncorrelated. As a result, the bifactor MIRT model effectively provides a decomposition of an item's shared latent item variance with respect to a dimension measured by all items (a general proficiency dimension) and a dimension specific to the item group to which the item belongs (a specific factor).

Bifactor MIRT models (Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937) and related testlet models (Wainer, Bradlow & Wang, 2007) have received much attention in the recent psychometric literature, due both to their frequently good empirical fit to data as well as advances related to their estimation (Rijmen, 2009; Cai, 2010) that have made the models estimable even in cases of high dimensionality. The high dimensionality problem is noted as a primary limitation to the use of MIRT in comparison to CDMs (Rupp & Templin, 2008).

From the simulation study, it would appear that if specific factors are defined so as to represent only the most informative skill attribute measured by an item, very little may be lost with respect to diagnosis. Moreover, due to the continuity of factors in the model, a bifactor MIRT approach may have the benefit of addressing misspecification related to continuity of the skill attributes. As psychometric models, the bifactor and higher-order factor models possess strong similarities (Reise, 2012) in which the general factor of the bifactor model often assumes an interpretation similar to the higher-order factor. Of course one other difference between the bifactor approach and the HO-DINA approach is the presence of a compensatory interaction between the specific and general factors, as opposed to the conjunctive (i.e., "non-compensatory") interaction between skill attributes used in the HO-DINA model. While the psychological distinction between these alternative forms of interaction is clear, the practical relevance is questionable (e.g., van der Linden, 2012). Prior work has found compensatory models to provide a close fit to response data that might be theoretically believed to be conjunctive (non-compensatory) in nature (e.g., Bolt & Lall, 2003).

19.4 Real Data Illustration – Fraction Subtraction Dataset

We consider the fraction subtraction data of Tatsuoka (1990) here because it likely represents a setting where the specificity of skills is high and CDMs are seemingly most applicable. We initially replicated logistic regression analyses of the form applied in the simulation. As a first step we fit the HO-DINA to the 20-item and 15-item datasets using Q-matrices implied by Table 19.1. Table 19.5 shows the parameter estimates related to the higher-order factor model obtained using the *cdm* R package (Robitzsch et al., 2017).

Table 19.5 HO-DINA parameter estimates for higher-order factor model, 20-item and 15-item fraction subtraction analyses

	20 item analysis		15 item analysis	
Attribute	$\widehat{\lambda}_0$	$\widehat{\lambda}_1$	$\widehat{\lambda}_0$	$\widehat{\lambda}_1$
1	.14	4.15	−2.52	2.84
2	−1.67	1.58	−2.10	2.80
3	−1.39	.53	−3.38	4.04
4	−.55	2.04	−.02	1.37
5	.07	2.09	−.08	1.35
6	−1.88	1.78		
7	−3.26	3.39		
8	−1.32	1.41		

Table 19.6 Items for which skill attribute is most difficult, 20-item and 15-item HO-DINA analysis of fraction subtraction data (Tatsuoka, 1990)

Skill Attribute	Items for which attribute is most difficult of required attributes (20 item analysis)	Items for which attribute is most difficult of required attributes (15 item analysis)
1	7,15,19	1,3
2	9,14,16	8
3	None	5, 9, 11
4	1,2,3,5	2,6,7,12,13,15
5	4,10,11,13,17,18,20	4,10,14
6	None	NA
7	6,8	NA
8	12	NA

For the 20-item analysis, we observe the following ordering of the attributes based on their estimated marginal skill probabilities (proportion of masters in parentheses): 1 (.486), 5 (.489), 4 (.596), 8 (.778), 2 (.814), 7 (.822), 6 (.822) and 3 (.891). For the 15-item analysis the order is: 5 (.474), 4 (.496), 2 (.757), 3 (.793), and 1 (.800). Based on the Q-matrix, we can then identify for each item the required skill attribute of highest "difficulty", here taken to be the required attribute that is the least frequently mastered. Table 19.6 reports the corresponding items for each attribute.

Note that for the 20-item analysis, attribute 5 is the most difficult of the required attributes for seven of the twenty items. Attributes 4, 1, and 2 follow with the next highest frequencies, four, three and three items, respectively. For the 15-item analysis, attribute 4 is the most difficult of the required attributes for six of the items, followed by attributes 3 and 5 with three items each.

In mimicking the logistic regression analyses of the simulation, we fit the HO-DINA model twenty (or fifteen) times, each time holding out one studied item, and then use the posterior probabilities of attribute mastery as predictors in a logistic regression model of the binary score on the studied item.

Tables 19.7 and 19.8 report results. As before, bold coefficients identify the required attribute that is of greatest difficulty for the item. Largely consistent with

Table 19.7 Hold-one-item-out logistic regression results, 20-item fraction subtraction analysis

	Skill attributes							
Item	Att1	Att2	Att3	Att4	Att5	Att6	Att7	Att8
1	0.02	−0.10	−0.10	***0.76***	−0.09	*3.47*	*−3.05*	−0.06
2*	0.07	−0.03	0.00	***0.88***	0.05	−0.11	*0.14*	−0.02
3*	0.14	0.08	−0.04	***0.77***	−0.01	0.08	*0.04*	−0.03
4	0.03	*−0.24*	*1.13*	−0.02	***0.64***	−0.05	*−0.06*	−0.19
5*	0.17	*0.31*	0.01	***0.34***	0.13	0.03	*−0.17*	*0.23*
6*	−0.01	−0.20	0.03	−0.03	−0.04	0.03	***1.09***	0.02
7*	***1.36***	*−0.01*	−0.03	−0.15	−0.21	−0.08	*0.00*	−0.05
8*	0.10	0.11	0.28	0.07	0.26	−0.04	***0.06***	−0.02
9*	0.14	***2.54***	−0.05	−0.02	−0.01	0.04	−1.55	0.14
10*	0.06	*−0.03*	−0.58	0.02	***0.98***	−0.08	*−0.08*	*0.34*
11*	0.12	*−0.17*	0.05	−0.05	***0.81***	0.02	*0.02*	0.15
12	0.02	0.14	0.06	−0.02	0.08	0.12	*0.32*	***0.12***
13*	0.06	*−0.09*	−0.08	*0.19*	***0.45***	−0.07	*−0.04*	0.22
14	0.09	***0.28***	−0.10	0.01	0.09	−0.01	*0.49*	0.11
15*	***0.56***	−0.02	−0.09	0.13	0.20	0.02	*0.00*	0.11
16*	0.02	***0.42***	−0.02	0.03	0.11	−0.05	*0.32*	0.15
17*	0.15	*0.10*	0.02	−0.06	***0.72***	−0.02	*0.02*	−0.06
18*	0.12	*0.10*	0.02	0.05	***0.51***	*0.02*	*0.04*	0.13
19*	***0.52***	*0.16*	*−0.96*	0.13	*0.44*	−0.18	*0.39*	0.12
20*	0.04	*0.00*	*0.35*	−0.09	***0.77***	−0.41	*0.23*	0.03

*Items for which most difficult required attribute is also the most discriminating; italics imply the attribute is required by the item, bold identifies the most difficult of the required attributes for the item

the simulation results, in the 20-item analysis, for 16 of the items the attribute of highest difficulty also emerges as the attribute that is most highly discriminated by the item. For the 15-item analysis, this result is observed in 12 of the 15 items. In one of the items (item 1 of the 20-item analysis), the failure to see the most difficult attribute emerging appears to reflect a multicollinearity problem related to attributes 6 and 7, whose posterior probabilities of mastery show a very high correlation (.98) in the HO-DINA analysis.

Importantly, even for items where the most difficult attribute is not the most discriminating, the results need not be viewed as detrimental to the application of bifactor MIRT. By assuming a continuous representation of attributes, under the bifactor MIRT model the "difficulty" of an attribute can naturally be viewed as varying by item. The notion of skill attributes as continuous as opposed to discrete carries an implication that for certain of the items, a higher level of mastery may be required, implying that the difficulty order of required attributes could vary across items. For example, application of the subtraction of numerators (attribute 7) could naturally be more difficult if the numerators are larger as opposed to smaller numbers. In this respect, we might view the Q matrix as ultimately defining

Table 19.8 Hold one item out logistic regression results, 15 item fraction subtraction analysis

	Skill attributes				
Item	Att1	Att2	Att3	Att4	Att5
1*	***0.82***	−0.11	−0.08	0.39	0.19
2*	*0.35*	*0.71*	*−0.58*	***0.74***	−0.01
3*	***0.37***	−0.35	0.41	0.00	0.00
4*	*0.16*	*0.01*	*0.12*	*0.20*	***1.06***
5*	−0.81	−1.89	***2.75***	0.09	0.21
6*	*0.15*	*0.43*	*0.00*	***0.80***	−0.03
7*	*0.12*	*0.53*	*−0.08*	***0.92***	−0.02
8	*0.29*	***0.25***	0.16	NA	−0.02
9	*0.34*	0.21	***0.19***	−0.04	0.01
10	*0.33*	0.42	*−0.04*	*0.33*	***0.31***
11*	*−0.07*	−0.15	***0.98***	−0.03	−0.01
12*	*0.09*	0.16	*0.24*	***0.65***	0.30
13*	*0.27*	*0.22*	*0.22*	***0.42***	0.01
14*	*0.10*	*0.14*	*0.12*	*0.33*	***0.82***
15*	*0.12*	*0.15*	*0.19*	***0.73***	0.21

*Items for which most difficult required attribute is also the most discriminating; italics imply the attribute is required by the item, bold identifies the most difficult of the required attributes for the item

constraints as to which of the skill attributes has the potential to emerge as the most difficult, and thus accommodate settings in which the most relevant attribute for an item need not be the attribute that is on average the most difficult.

Returning to Table 19.6 and the attribute difficulties as defined by HO-DINA, it appears that for the 20-item analysis, two of the attributes have either one or no items for which the attribute is most difficult, namely attributes 3, 6, and 8. The results would suggest attending to attributes 1, 2, 4, 5, and possibly 7, where better overall measurement of the attributes would be anticipated. For the 15-item analysis, attributes 3, 4, and 5, and possibly 1, seem most relevant. However, as noted previously, bifactor MIRT models based on these specifications can be altered, either based on closer inspection of the items and/or statistical criteria, such as modification indices under the bifactor MIRT analyses. Some modification seems likely in the current analysis given the very close mastery proportions of certain attributes as estimated using HO-DINA, and the potential for an attribute with a lower average difficulty to emerge as the most difficult for a given item.

Inspection of modification indices led us to see that for the 20-item analysis, there seemed to be greater value in attending to attribute 7 as opposed to attribute 2. A reason for this is that the attribute is both the most frequently invoked across items and appears to be an attribute of moderate difficulty. Thus attribute 7 might be expected to emerge as the most difficult of the required attributes for a significant number of items. This result appears to a large extent consistent with Table 19.5, where items such as item 14 appear to be affected more by attribute 7 than attribute 2.

Following this re-specification, we are ultimately led to a bifactor MIRT model with four specific factors for the 20-item analysis corresponding to attributes 1, 4, 5, and 7. For comparison purposes, we also considered a bifactor MIRT model with three specific factors, dropping the specific factor related to attribute 1. For the 15-item analysis, we retained models with specific factors as defined by the item/factor correspondences in Table 19.6 (i.e., attributes 1, 3, 4 and 5), as well as a model with 3 specific factors that drops attribute 1. An important feature of all the bifactor MIRT models considered is that the specified relationship between items and specific factors is consistent with the corresponding Q-matrix. Items that do not require any of the attended-to attributes load only on the general dimension and no specific factors.

For comparison purposes, we also considered for both the 20-item and 15-item datasets the independence DINA, the HO-DINA, as well as unidimensional IRT and exploratory MIRT models. The bifactor MIRT model, unidimensional IRT and exploratory MIRT models were all fit using the R *mirt* package (Chalmers, 2012); the DINA and HO-DINA models were fit using the R *cdm* package (Robitzsch et al., 2017).

As seen in Table 19.9, the bifactor MIRT models consistently outperform the DINA and HO-DINA models with respect to traditional information criteria (AIC, BIC). Moreover, the bifactor MIRT models appear to fit better than exploratory MIRT models, particularly with respect to the BIC criterion. For simplicity, hereafter we focus on the 20-item analysis. Table 19.10 displays the discrimination estimates for both the general and specific factors for each of the items for the 20-item analysis.

For the 20-item analysis, consistent with prior findings (e.g., de la Torre & Douglas, 2004), it would appear that item 8 is not useful. The problems with this item, "2/3 – 2/3 =?", have been noted elsewhere (DeCarlo, 2011), and are here seen from its negative loading on its specific factor (although note that the item is still informative in regard to the general factor). As noted earlier, another appealing aspect of the bifactor routine in *mirt* is the potential to allow items to load only on the general factor. It is conceivable that certain items may not provide additional

Table 19.9 Model comparison results, fraction subtraction data (N = 536; Tatsuoka, 1990)

	20 items/8 attributes				15 items/5 attributes			
Model	Loglik	#pars	AIC	BIC	Loglik	#pars	AIC	BIC
DINA	−4402.4	295	9395	10,659	−3455.8	91	7034	7295
HO-DINA	−4423.3	56	8959	9198	−3456.0	40	6992	7163
3 Bifactor	−4415.7	56	8943	9183	−3369.5	42	6823	7003
4 Bifactor	−4406.9	59	8932	9185	−3365.9	44	6820	7008
Uni-IRT	−4641.0	40	9362	9533	−3452.9	30	6966	7094
2D MIRT	−4454.5	59	9027	9280	−3368.6	44	6825	7014
3D MIRT	−4378.6	77	8911	9241	−3320.6	57	6755	6999
4D MIRT	−4346.4	94	8881	9283	−3311.7	69	6761	7057

Table 19.10 Bifactor slope estimates, bifactor MIRT models with 4 specific factors, 20-item fraction subtraction data (Tatsuoka, 1990)

Item	G	S1(Att1)	S2(Att4)	S3(Att5)	S4(Att7)
1	0.789	0.000	0.511	0.000	0.000
2	0.845	0.000	0.510	0.000	0.000
3	0.820	0.000	0.520	0.000	0.000
4	0.661	0.000	0.000	0.365	0.000
5	0.592	0.000	0.247	0.000	0.000
6	0.802	0.000	0.000	0.000	0.383
7	0.854	0.430	0.000	0.000	0.000
8	0.662	0.000	0.000	0.000	−0.188
9	0.467	0.000	0.000	0.000	0.000
10	0.832	0.000	0.000	0.342	0.000
11	0.804	0.000	0.000	0.488	0.000
12	0.747	0.000	0.000	0.000	0.441
13	0.864	0.000	0.000	0.170	0.000
14	0.804	0.000	0.000	0.000	0.433
15	0.876	0.299	0.000	0.000	0.000
16	0.770	0.000	0.000	0.000	0.443
17	0.840	0.000	0.000	0.397	0.000
18	0.792	0.000	0.000	0.276	0.000
19	0.924	0.216	0.000	0.000	0.000
20	0.820	0.000	0.000	0.505	0.000
SS loadings	12.34	.32	.85	1.01	.76
Prop Var	.62	.02	.04	.05	.04

specific information regarding any of the specific factors despite their measurement of the general factor. For example, item 9 is handled this way in the 20-item analysis. A respecification of the model might treat item 8 similarly.

A second advantage of the bifactor parameterization relates to its ability to clarify the overall capacity of a test to discriminate with respect to measurement of the specific attributes. It becomes clear from inspecting the columns of Table 19.10 how well individual specific factors are being measured by the test. This distinction of bifactor MIRT and its implications for student-level diagnostic reports will be discussed further in the next section.

19.5 Applications of Bifactor MIRT for Student-Level Diagnostic Assessment

A comparison of CDMs and bifactor MIRT naturally leads to questions regarding the similarities and differences of the diagnostic information provided by each approach. First, we examine at the individual student level the change in fit

associated with each of the bifactor MIRT and HO-DINA models in comparison to a unidimensional IRT model. Specifically, we examined the log-likelihoods of student-level score patterns conditional on proficiency estimates for (1) the 2PL, (2) HO-DINA, and (3) the bifactor MIRT models, focusing on the change in log-likelihood provided by the HO-DINA and bifactor MIRT in comparison to the 2PL. For each model we define the loglikelihood for an observed response pattern conditional upon the MAP proficiency estimates as well as the estimated item parameters for each model. Across student response patterns, these 2PL- and HO-DINA-based log-likelihoods correlated at .756, while the 2PL- and bifactor-MIRT-based log-likelihoods correlated at .893. When quantifying the change in loglikelihood between the 2PL and HO-DINA versus the 2PL and bifactor MIRT, we observe a correlation between the log-likelihood differences of .778, suggesting both HO-DINA and bifactor MIRT are attending to similar information in improving fit over the 2PL. However, for 411 of the 536 students, the log-likelihood based on the bifactor MIRT analysis resulted in a higher log-likelihood value than that based on the corresponding HO-DINA.

Despite the similarities between the HO-DINA and bifactor MIRT results against the 2PL, the nature of the diagnostic information provided by the general and specific factor estimates of bifactor MIRT is different. As noted above, the general factor estimate of the bifactor model tends to reflect overall performance, while the mutually uncorrelated specific factor estimates identify a type of profile. As each of the factors also has a mean of 0, specific factor estimates different from 0 imply performances on the specific factor items that are below (negative) or above (positive) expectations based on the level of the general factor. We consider below two ways in which such estimates might be used for score reporting purposes. The first reports the continuous factor estimates themselves; the second provides mastery/nonmastery binary skill attribute reports similar to CDMs.

Given the bifactor MIRT model item parameter estimates, maximum a priori (MAP) estimates of the general and specific factors can be obtained for a given response pattern. One of the appealing aspects of this form of score reporting is that it becomes possible to selectively report diagnostic information (i.e., in the form of specific factor estimates) where statistical evidence supports it. A basis for such a report could attend either to the change in likelihood observed at the student level, or alternatively a Wald test applied to the specific factor estimates. Such an approach makes apparent the need for collecting sufficient empirical data when making a diagnosis; interestingly in virtually none of the cases with the 20-item fraction subtraction data do we obtain such evidence, likely owing to the relatively small number of items based on those factor estimates.

A second possibility would use the bifactor MIRT general and specific factor estimates to produce a binary skill attribute classification analogous to CDMs by imposing thresholds of mastery applied to suitable linear combinations of the general and specific factor estimates. The thresholds and linear combinations implied by a HO-DINA classification provide one possibility, but others could be chosen (or alternatively, imposed based on a previous analyses). As an illustration, a logistic regression analysis was applied in which the HO-DINA MAP estimate

of a single attribute was the outcome, and the bifactor MIRT factor estimates were entered as predictors. Conveniently, this approach also provides a way of examining the consistency of classification results between HO-DINA and bifactor MIRT. The logistic regression results can subsequently be used to classify students as masters and nonmasters for each of the attributes using the logistic-regression-based probabilities from the logistic regression (>=.50 implies mastery, <.50 nonmastery). Using this approach, we find very close agreement between mastery/nonmastery classification results for each of attributes 1 (98%), 4 (100%), 5 (99%), and 7 (97%) specifically attended to in the bifactor MIRT analysis of the 20-item dataset. Interestingly, even for the unattended-to attributes (2, 3, 6, and 8), if we wished to make a prediction of attribute mastery based only on the general factor estimate, we still observe quite consistent results. The classification consistencies results for attributes 2 (93%), 3 (99%), 6 (96%), and 8 (93%) are quite high. Of course, in the presence of attribute continuity under bifactor MIRT, it also is possible to use the same regression weights with adjusted thresholds so as to yield higher or lower numbers of masters of a particular attribute.

Tables 19.11a and 19.11b provide some examples of observed response patterns in the 20-item fraction subtraction data. Each row of Table 19.11a corresponds to a different student, identified by case number and item response pattern. The subsequent columns identify the MAP proficiency estimates based on HO-DINA, the 2PL, and the bifactor MIRT, as well as the loglikelihood of the pattern at the corresponding estimates for each model. Table 19.11b provides a comparison of the HO-DINA and bifactor MIRT-based binary attribute classifications for each of the same response patterns using the logistic regression approach described above.

We first consider examples of some of the 125/536 patterns for which HO-DINA displayed a higher log-likelihood than bifactor MIRT. As might be expected, several of these patterns involve conditions where atypical results emerge with respect to attributes 2, 3, 6, or 8. Examples include student IDs 3 and 373. For student 3, the higher loglikelihood under HO-DINA can be explained by the non-mastery status assigned to attribute 8 (predicted as a master in the bifactor model); for case ID 373, a similar effect is observed for a student identified as a non-master on both attributes 6 and 8 (both estimated as masters based on bifactor MIRT estimates). For other patterns on which HO-DINA has the higher loglikelihood, the result appears not due to a unique attribute mastery classification, specifically, but rather the unique functioning of slip/guessing parameters as sources of stochasticity in the HO-DINA model. For example, case ID 105 yields a higher loglikelihood due to the fact that incorrect responses occur on the items with the highest slip parameters.

As noted above, there are many cases (411/536) where the bifactor MIRT model produces the higher log-likelihood. In certain instances, this appears to be related to the capacity of the bifactor MIRT model to provide a continuous representation of the general and specific factors. For example, case ID 234 represents a student that despite overall poor performance on attribute 4 items (3 of the 5 items requiring attribute 4 in the Q-matrix are answered incorrectly) is nevertheless classified as a master under HO-DINA; however the $\widehat{\theta}_{SP2}$ (attribute 4) of -1.37 under bifactor MIRT clearly identifies the attribute as an area of relative weakness for the student.

Table 19.11a Classification results

Case	Item response pattern	HO-DINA $\widehat{\boldsymbol{\alpha}}$	2PL $\widehat{\boldsymbol{\theta}}$	Bifactor MIRT $\widehat{\theta}_{GEN}$	$\widehat{\theta}_{SP1}$(a1)	$\widehat{\theta}_{SP2}$(a4)	$\widehat{\theta}_{SP3}$(a5)	$\widehat{\theta}_{SP4}$(a7)	2PL Loglik	HO-DINA Loglik	Bifactor MIRT Loglik
239	00010000011101001101	01101111	−.03	−.39	−.15	−.50	2.01	−.37	−19.02	−9.94	−8.33
119	00010111111101110111	11101111	.42	.48	.80	−1.48	.27	.10	−12.92	−4.55	−5.82
38	11111110000001110000	11110111	.08	.08	.93	.66	−.64	−.52	−10.70	−5.59	−6.70
285	01100011111011101011	11111111	.48	.43	.87	−.47	.48	−1.97	−20.55	−13.82	−12.46
3	01110111000001110000	11110110	−.05	−.02	1.08	.15	−.51	−.42	−11.30	−4.69	−8.71
373	01100101100001010100	01110010	−.21	−.16	−.30	.35	−.37	−.27	−9.01	−4.02	−7.82
105	01111111111111101000	11111111	.46	.40	.41	−.35	.09	−.89	−11.44	−6.06	−10.07
234	01010111011111111111	11111111	.78	.77	.42	−1.37	.69	.04	−9.63	−10.33	−5.28
171	00000000000100011101	00100000	−.39	−.60	−.07	−.36	1.44	−.17	−15.65	−15.90	−11.40
28	00000000000000000000	00100000	−1.79	−1.68	−.00	−.08	−.07	−.22	−1.35	−2.19	−1.16
48	00000100001101001100	01001111	−.26	−.44	−.13	−.47	1.07	.14	−12.55	−8.14	−8.75
220	11010101011101001011	11111111	.35	.19	−.13	−.26	.93	−.69	−14.05	−15.56	−11.63
431	11111111111111111111	11111111	1.44	1.31	.07	.07	.26	.00	−.71	−3.15	−.62
44	00000100000100000000	00100111	−1.05	−1.15	−.01	−.16	−.17	.45	−4.70	−3.20	−3.69
414	00000000000101010000	01100111	−.86	−1.02	−.01	−.19	−.21	.88	−6.33	−7.53	−4.58
200	10100100000100100000	10111111	−.45	−.52	.87	.63	−.47	−.33	−11.51	−6.72	−10.31
237	11110110100100110000	11110111	.01	−.10	1.20	.68	−.41	−.29	−12.12	−9.13	−8.65

Table 19.11b Classification results

	HO-DINA	Bifactor MIRT
Case	$\widehat{\boldsymbol{\alpha}}$	$\widehat{\boldsymbol{\alpha}}$
239	01101111	00101111
119	11101111	11101111
38	11110111	11110111
285	11111111	11111110
3	11110110	11110111
373	01110010	01110111
105	11111111	11111111
234	11111111	11111111
171	00100000	00100111
28	00100000	00100000
48	01001111	01100111
220	11111111	11111111
431	11111111	11111111
44	00100111	01100110
414	01100111	01100111
200	10111111	01110110
237	11110111	11110111

For ID 171, the relative strength of the student on attribute 5 is identified (based on the $\widehat{\theta}_{SP3}$ of 1.44), while for HO-DINA this student is viewed as a master of only attribute 3. Finally, it is important to note that through the general factor, the bifactor MIRT model is able to highlight differences in overall performance beyond that explained by the binary classification. In particular, the response pattern of all correct (see case ID 431; of which there are 30 in the entire dataset) can be accounted for by a high general factor estimate ($\widehat{\theta}_{GEN}=$ 1.31) that exceeds those of other students who may have as few as 12 items correct (e.g., case ID 220) but were nevertheless also classified as masters under HO-DINA of all eight attributes.

As noted above, in most cases, bifactor MIRT leads to similar diagnostic classification results. Examples from Table 19.9 include IDs 119 and 38. As seen in Table 19.11b, if using the logistic regression approach based on bifactor MIRT estimates to obtain a binary classification, we obtain equivalent results for these (and many other) response patterns.

A deeper understanding of the differences between HO-DINA and bifactor MIRT should attend to the attributes not explicitly modeled in the bifactor MIRT analysis and for which the classification results were less consistent. Noticeably, some of the more atypical diagnostic classifications observed under HO-DINA involved the status of these attributes. Consider, for example, ID 28 (of which there are 13 identical cases). Despite having answered all of items incorrectly, the examinee is declared a "master" of attribute 3. This result contrasts with that of examinee 48, who despite having answered 6 items correctly, is declared a non-master of attribute 3. The result provides an example of a type of "paradox" discussed by Hooker, Finkelman, and Schwartzman (2009) in the context of compensatory MIRT

models, whereby better overall performances can result in a decline in certain ability estimates. The problem arguably becomes even greater with binary skills models having well-ordered attributes, as knowledge gained regarding mastery of certain attributes is often necessary before evidence can be accumulated regarding the mastery status of others. We might view this as implying a type of "conditional" mastery assessment. In this case, the mastery/non-mastery status of attribute 3 cannot be determined when all other attributes appear to be non-mastered. As noted by DeCarlo (2011), the classification as a master of attribute 3 is consequently only a reflection of the prior.

Similarly peculiar results are observed for IDs 44, 414 and 200. For these example respondents, a disproportionately large number of attributes are declared mastered despite very few overall items having been answered correctly. Not surprisingly, such attribute mastery patterns frequently entail mastery classifications for the attributes (i.e., 3, 6 and 8) that from Table 19.6 had no items for which the attribute was most difficult. Consequently, it would again appear that the mastery designation is often not based on evidence in support of mastery, but rather the lack of evidence implying non–mastery. Another example of the Hooker et al. paradox is seen in comparing ID 200 to ID 237. While the response pattern for ID 200 is entirely nested within ID 237 (with five additional items being answered correctly in ID 237) ID 200 is declared a master of attribute 5 while ID 237 is a non-master. In this case there would also appear to be a "conditional" mastery assessment issue, now in regard to attribute 5.

Beyond being a curiosity, such results arguably have the potential to create confusion, particularly in longitudinal applications. A student displaying overall gains in test performance from one year to the next may nevertheless observe a transition from a mastery to non-mastery status on certain attributes. Likewise, a non-mastery diagnosis may simply imply a lack of evidence in support of mastery as opposed to clear evidence implying non-mastery, such as when a student may have mastered a more difficult attribute, but the mastery is unseen because all items requiring the relevant attribute require other attributes the student may not have mastered.

19.6 Modeling Advantages of Bifactor MIRT

The previously discussed issues in regard to student scoring help identify a couple of potential advantages to the use of bifactor MIRT for diagnostic scoring. The first concerns the potential use of bifactor MIRT in designing tests that eliminate the occurrence of Hooker's paradox. Note that for the two examples of Hooker's paradox in the previous section, the paradox is resolved when using the bifactor MIRT estimates. Importantly, such paradoxes can still occur in the context of bifactor MIRT. However, bifactor models make it possible to design tests so as to eliminate their potential for occurrence. Space limitations preclude further

discussion of this issue here, but we refer the interested reader to Hooker and Finkelman (2010) and van Rijn and Rijmen (2015) for details.

A second advantage concerns the potential for equating/linking calibrations so as to accommodate comparisons of students administered different tests. While in theory CDMs possess invariance properties, the presence of skill attribute continuity often coincides with violations of parameter invariance for which there are not clear methods that can be applied to preserve invariance of the skill attribute metrics (Bolt, 2017). By contrast, the use of linking/equating methods to create common metrics across populations that may differ substantially in ability is more straightforward in MIRT (see e.g., Weeks, 2015).

Third and finally are advantages related to bifactor MIRT as a generalization of unidimensional IRT. One advantage of this feature that was previously discussed concerns the capacity to statistically evaluate, either at the level of the student or entire sample, whether attending to the specific factors appears statistically justified. A second relates to the capacity of bifactor MIRT to resolve the challenges associated with simultaneous application of CDM and unidimensional IRT models to the same data, as happens in computerized adaptive testing applications, for example (Wang, Zheng, & Chang, 2017). In such settings, both overall performance and diagnostic evaluations are often of interest. The simultaneous use of two different statistical models is not just an inconvenience but requires practical resolution in relation to item selection algorithms, for example. Use of bifactor MIRT, as seen above, can allow both elements of performance to occur within the same statistical model.

19.7 Discussion

In this chapter, the bifactor MIRT model is suggested as an appealing alternative to CDMs, especially under conditions where attribute continuity is suspected or can be confirmed. A separate paper (Bolt, 2017) shows attribute continuity to be clearly present even in Tatsuoka's fraction subtraction data set, for which CDMs have been extensively applied. The results in this chapter remind that the assumption that a particular skill is required in solving an item should not always be taken to imply that the skill is statistically measured by the item. The applicability of bifactor MIRT follows in part from the tendency for items measuring multiple conjunctively interacting skill attributes to primarily distinguish only with respect to the most difficult of the required skill attributes, especially when a higher order factor underlies the skill attributes. The result observed by Bolt and Lall (2003) regarding noncompensatory MIRT models appears to apply also to CDMs, namely that attributes (traits) of lower difficulty seem to be not well measured, if at all. In addition, the computational cost of a bifactor MIRT model is minimal, even in the presence of high dimensionality.

The findings of the simulation also speak to aspects of test design for CDMs. On the one hand, the results help clarify why items that measure just one skill attribute

are frequently found to be most useful in measurement of an attribute (Bradshaw & Madison, 2016). For optimal measurement of an attribute, it also appears important to have items designed for which the attribute emerges as the most difficult of the required skill attributes for some items. Skill attributes having no such items can still contribute to diagnoses, but often do so in asymmetric and conditional ways. For example, the ability to see a student as a non-master on a particular attribute may require that the student have mastered a more difficult attribute. This feature of the HO-DINA may become problematic when diagnostic assessments occur repeatedly over time (such as within an academic year or across years), as improvements in overall performance can nevertheless lead students to move from mastery to non-mastery diagnoses on certain attributes.

Our results suggest a close correspondence between dimensions defined in CDMs and those of MIRT. As seen in the fraction subtraction data analyses, the skills that emerge as dimensions in the analysis of fraction subtraction data appear to conceptually be the same as the skill attributes of CDMs; there are simply fewer of them that emerge under the bifactor MIRT analysis. In terms of student-level diagnoses, the results also suggest a closer relationship between the diagnostic information provided by the bifactor and CDM approach than might have otherwise been anticipated. Where there are differences, bifactor MIRT appears to afford some value in how the general factor can distinguish performances among students that are otherwise all declared as masters on the studied attributes. HO-DINA appears to offer some advantages in making use of attributes that fail to emerge as statistically meaningful in the bifactor MIRT analyses.

From a score interpretation perspective, seeing how items frequently vary (often quite substantially) in their measurement of skill attributes within CDMs is important. There can naturally be a perception that for a given skill attribute all items identified by the Q matrix as measuring the attribute contribute equally; this is clearly not the case. It seems likely that a given student's mastery status on an attribute could well have been informed by largely just one item, in which case the item might be scrutinized more carefully for validity and any consequential decisions based on that diagnosis might be evaluated more cautiously.

We acknowledge some limitations in our current analyses. Naturally, there exist other criteria by which the bifactor MIRT and CDM models could be compared empirically, both to the fraction subtraction data as well as other datasets. There also exists a wide range of simulation conditions against which the modeling approaches could be compared, not just in terms of overall fit, but also in the scoring of individual examinees. We suggest that the combination of results from the studies above motivate further considerations of bifactor MIRT models in contexts that frequently motivate consideration of CDMs.

Acknowledgements The author would like to thank Nana Kim and the two assigned reviewers for their review and comments on an earlier version of this chapter.

References

Bolt, D. M. (2017). *Parameter invariance and skill attribute continuity in the DINA model.* Unpublished manuscript.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*(6), 395–414.

Bradshaw, L. P., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing, 16*(2), 99–118.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581–612.

Chalmers, R. P. (2012). MIRT: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*. www.jstatsoft.org

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*(1), 115–127.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49*, 175–186.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57*, 423–436.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54.

Hooker, G., & Finkelman, M. (2010). Paradoxical results and item bundles. *Psychometrika, 75*(2), 249–271.

Hooker, G., Finkelman, M., & Schwartzman, A. (2009). Paradoxical results in multidimensional item response theory. *Psychometrika, 74*(3), 419–442.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696.

Rijmen, F. (2009). Efficient full information maximum likelihood estimation for multidimensional IRT models. *ETS Research Report Series*, i-31.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). CDM: Cognitive diagnosis modeling. *R package version*, 3-1.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219–262.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82–98). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

van der Linden, W. J. (2012). On compensation in multidimensional response modeling. *Psychometrika, 77*(1), 21–30.

van Rijn, P., & Rijmen, F. (2015). On the explaining-away phenomenon in multivariate latent variable models. *British Journal of Mathematical and Statistical Psychology, 68*(1), 1–22.

von Davier, M., & Haberman, S. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'Diagnostic' classification models – A commentary. *Psychometrika, 79*, 340–346.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge, UK: Cambridge University Press.

Wang, C., Zheng, C., & Chang, H.-H. (2017). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement, 51*, 358–380.

Weeks, J. P. (2015). Multidimensional test linking. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 406–434). New York, NY: Routledge.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479–494.

Part III
Applications

Chapter 20
Utilizing Process Data for Cognitive Diagnosis

Hong Jiao, Dandan Liao, and Peida Zhan

Abstract Process data, different from item responses, essentially shows the interactions between test-takers, item presentation including stems and options, technology-enhanced help features, as well as the computer interface. With the availability of process data in addition to product data, additional auxiliary information from the response process can be utilized to serve different assessment purposes such as enhancing accuracy in ability estimation, facilitating cognitive diagnosis, and aberrant responding behavior detection. Response time (RT) is the most frequently studied process data contained in log files in current psychometric modeling, although other process data is available such as the number of clicks, the frequency of use of help features, frequency of answer changes, and data collected using eye-tracking devices. Process data is worthy of exploration and the integration with product data can enhance our evidence base for assessment purposes. This chapter will focus on the use of RT as one important type of process data in cognitive diagnostic modeling.

H. Jiao (✉)
Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, USA
e-mail: hjiao@umd.edu

D. Liao
American Institutes for Research, Washington, DC, USA
e-mail: dliao@air.org

P. Zhan
Department of Psychology, College of Teacher Education, Zhejiang Normal University, Zhejiang, China
e-mail: zhan@zjnu.edu.cn

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_20

20.1 Introduction

Cognitive diagnosis intends to provide fine-grained information about the strengths and weaknesses in learning. Such information can help remedial instruction and learning adapted to learners' deficiency. Cognitive diagnosis can be conducted using different theoretical frameworks such as latent trait based models (e.g., de la Torre & Douglas, 2004; de la Torre & Song, 2009; Embretson, 2015; Embretson & Yang, 2013; Haberman & Sinharay, 2010; Yao & Boughton, 2007), latent class based models (e.g., de la Torre, 2011; Macready & Dayton, 1977; Maris, 1995, 1999; von Davier, 2008; von Davier & Yamamoto, 2004) such as the deterministic inputs, noisy "and" gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977), and item-feature related models (e.g., Embretson & Yang, 2013). There are multiple generalized frameworks such as the general diagnostic model (GDM; von Davier, 2008), the generalized DINA model (G-DINA; de la Torre, 2011), and the loglinear CDM model (LCDM; Henson, Templin, & Willse, 2009). Among the three, the GDM can be viewed as the most general approach as it allows both continuous latent traits and discrete attributes as latent variables. All these models utilize item response data, which are item scores based on the answer to each item in a test.

With the latest advances in computer technology, computer-based assessment is becoming common practice in large-scale testing and classroom assessment. The use of computers in assessment makes it easy to collect more data in addition to item responses. Rupp, Gushta, Mislevy, and Shaffer (2010) defined two types of data collected in game-based assessment, process data and product data. Process data is related to the interactions of test-takers with other test-takers and computers or games during the process of assessment or game play while product data is the outcome of the assessment tasks or items such as item scores. Both types of data are on a regular basis collected in computer-based assessment.

Process data, different from product data or item responses, essentially shows the interactions between test-takers, item presentation including stems and options, technology-enhanced help features, as well as the computer interface. With the availability of process data in addition to product data, additional auxiliary information from the response process can be utilized to serve different assessment purposes such as enhancing accuracy in ability estimation, facilitating cognitive diagnosis, and aberrant responding behavior detection. Response time (RT) is the most frequently studied process data contained in log files in current psychometric modeling, although other process data is available such as the number of clicks, the frequency of use of help features, frequency of answer changes, and data collected using eye-tracking devices. Process data is worthy of exploration and the integration with product data can enhance our evidence base for assessment purposes. This chapter will focus on RT as one important type of process data.

RT, as a continuous variable, contains information that may not be directly decoded from discrete product data like item responses, which are often scored dichotomously representing correct and incorrect answers or polytomously repre-

senting different degrees of correctness. Given the same outcome, an item score, different RTs may reveal the psychometric properties of the task or the level of cognitive challenge a test-taker may face with. RT can further provide information about the working speed of test-takers. For example, when test-takers do not have enough time to complete items on a test, their RTs may display a different pattern compared to test takers who have enough time to fully engage in problem-solving. Similarly, test-takers may not be motivated enough in a low-stakes test. Thus, they may be engaged in speeded responding behaviors (e.g., Klein Entink, van der Linden, & Fox, 2009; Locke, 1965; Logan, Medford, & Hughes, 2011). Further, test-takers with prior knowledge of items or those who are cheating on a test may exhibit shorter RT than other test-takers (e.g., Qian, Staniewska, Reckase, & Woo, 2016). Thus, information in RTs can potentially be used as collateral information in addition to item responses to serve additional psychometric purposes.

RT data has been utilized to deal with different psychometric issues and challenges. For instance, van der Linden, Klein Entink, and Fox (2010) used RT as collateral information for IRT parameter estimation. Gaviria (2005) explored using RT to increase model parameter estimation precision in computer-based tests. Ranger and Kuhn (2012) conducted a similar exploration in psychological tests. Other studies explored the use of RT in detecting abnormal response behaviors (e.g., Holden & Kroner, 1992; Lee & Wollack, 2017; van der Linden & Guo, 2008) or for better understanding response behaviors (e.g., Schnipke & Scrams, 2002). Others also developed new person fit indexes based on RT (e.g., Fox & Marianti, 2017; Man, Jiao, & Ouyang, 2016; Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014). Recently, researchers proposed different models for response times (e.g., Klein Entink, van der Linden, et al., 2009; Ranger & Kuhn, 2013; Ranger & Ortner, 2012; van der Linden, 2006; Wang, Chang, & Douglas, 2013) and the joint modeling of responses and response times (e.g., Bolsinova, De Boeck, & Tijmstra, 2016; Klein Entink, Fox, & van der Linden, 2009; Loeys, Rosseel, & Baten, 2011; Maris & van der Maas, 2012; Molenaar, Bolsinova, Rozsa, & De Boeck, 2016; Molenaar, Oberski, Vermunt, & De Boeck, 2016; Molenaar, Tuerlinckx, & van der Maas, 2015; Ranger & Kuhn, 2014; van der Linden, 2007; Wang, 2006; Wang, Fan, Chang, & Douglas, 2013; Wang & Hanson, 2005).

Joint modeling of responses and RT accounts for the dependence between accuracy and speed. However, some empirical data analyses indicated RT depends on item responses (e.g., Bolsinova & Maris, 2016; Bolsinova & Tijmstra, 2016; Glas & van der Linden, 2010; van der Linden & Glas, 2010) and item characteristics (Bolsinova, Tijmstra, & Molenaar, 2017; Goldhammer, Naumann, & Greiff, 2015; Liao, 2018) after controlling for the relation between latent ability and speed. Further, as observed in real data analyses, the relationship between accuracy and speed is not invariant across diverse groups of test-takers. Some studies reported positive relation while others reported negative correlation between accuracy and speed. Such differential or conditional differences have been explored in several studies (e.g., Bolsinova et al., 2016, 2017; Fox & Marianti, 2016; Jiao, Zhan, Liao, & Man, 2017; Liao, 2018; Meng, Tao, & Chang, 2015; Molenaar, Bolsinova, et al.,

2016; Molenaar, Bolsinova, & Vermunt, 2018; Wang & Xu, 2015) using conditional modeling, multigroup structure or mixture modeling. However, all these modeling approaches were developed within the item response theory framework.

RT has been applied in adaptive testing as well for increasing the precision of parameter estimation and for detecting aberrant responses (e.g., van der Linden & van Krimpen-Stoop, 2003) or response time patterns (e.g., van der Linden & Guo, 2008; van Rijn & Ali, 2017). Several researchers (e.g., Minchen, 2017; Minchen & de la Torre, 2016; Zhan, Jiao, & Liao, 2018) proposed using RT or joint modeling of response and response time for cognitive diagnosis.

20.2 Joint Modeling of Responses and Response Time for Cognitive Diagnosis

Minchen and de la Torre (2016) first proposed to use RT to improve ability estimation in cognitive diagnosis models (CDMs). They followed the hierarchical framework proposed by van der Linden (2007) for joint modeling response and RT. To model item responses, a higher order attribute distribution (de la Torre & Douglas, 2004) with the DINA model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) was used while a lognormal model was used for RT with speed intensity parameters. Zhan et al. (2018) proposed a similar perspective for integrating responses and RTs for cognitive diagnosis. However, these two studies differ in that Zhan, Jiao, and Liao's model takes into account the dependency between item accuracy and speed parameters which was ignored in Minchen and de la Torre's formulation.

Following what Zhan et al. (2018) presented in their paper for the joint RT-DINA model, the subsequent sections present the graphical and the numerical representations of the joint model. Bayesian estimation of the model parameters are discussed as well.

20.2.1 The Joint RT-DINA Model

Like the hierarchical modeling framework for joint modeling of responses and RT (van der Linden, 2007) for accuracy and speed, item responses, $Y_{ji,}$ and log response time, $\log(T_{ji})$, are separately modeled at level 1 in the joint RT-DINA model. Level 2 contains two correlational structures to take into account dependencies among item parameters and among person parameters, respectively.

20.2.1.1 Level 1 Models: RT and DINA Models

The Lognormal RT Model

The lognormal RT model (van der Linden, 2006) is one of the most commonly used RT models. Thus, this model is used for illustration of joint modeling of responses and RT for cognitive diagnosis. Let T_{ji} be the observed item RT for person j ($j = 1, \ldots, J$) and item i ($i = 1, \ldots, I$). The logarithm of RT is used to transform the positively skewed RT, and a normal distribution for this transformation is assumed. That is,

$$\log\left(T_{ji}\right) = \zeta_i - \tau_j + \varepsilon_{ji}, \quad \varepsilon_{ji} \sim N\left(0, \sigma^2_{\varepsilon_i}\right), \tag{20.1}$$

where τ_j is the person speed parameter representing the average speed of person j on a test; ζ_i is the time-intensity parameter representing the population-average time needed to complete item i; ε_{ji} is the normally distributed error term indicating that this is a lognormal RT model. Thus, $\log\left(T_{ji}\right) \sim N\left(\zeta_i - \tau_{\mathrm{j}}, \sigma^2_{\varepsilon_i}\right)$.

Equation (20.1) can be extended to include a slope parameter for speed as a time-discrimination parameter (Klein Entink, Fox, et al., 2009; Klein Entink, van der Linden, et al., 2009). Further, a person-specific growth parameter can be included to allow for variable working speed throughout a test (Fox & Marianti, 2016). In addition, other methods could be used to model RT such as the Box-Cox transformation by Klein Entink, Fox, et al. (2009) and Klein Entink, van der Linden, et al. (2009) or a linear transformation model by Wang, Chang, et al. (2013).

The DINA Model

The DINA model is one of the commonly used and frequently studied models for cognitive diagnosis. Let Y_{ji} be the observed response of person j to item i. Equation 20.2 presents the relationship among attributes and the probability of an observed response.

$$P\left(Y_{ji} = 1\right) = g_i + (1 - s_i - g_i) \prod\nolimits_{k=1}^{K} \alpha_{jk}^{q_{ik}}, \tag{20.2}$$

where $P(Y_{ij} = 1)$ is the probability of a correct response by person j to item i; the two parameters, s_i and g_i are the slipping and guessing probability for item i respectively, indicating the item-level aberrant response probabilities; $(1 - s_i - g_i)$ is the item discrimination index (IDI_i; de la Torre, 2008) indicating item quality; the higher the value, the more discriminating an item is. α_{jk} is the mastery status of attribute k ($k = 1, \ldots, K$) for person j, with 1 if person j masters attribute k, and 0 otherwise. The Q-matrix (Tatsuoka, 1983) is a I-by-K matrix with elements q_{ik} indicating whether attribute k is required to answer item i correctly; it takes a value of 1 if the attribute is required for item i, and 0 otherwise.

Two reparameterizations are applied to transform the binary variable, α_{jk}, s_i, and g_i, so that multivariate normal distributions can be assumed at level 2 to consider the dependency of parameters on item and person sides respectively. The two item parameters, s_i and g_i, can be reparameterized from the probability scale to the logit scale (DeCarlo, 2011; Henson et al., 2009; von Davier, 2014) as follows.

$$\beta_i = \text{logit}\left(g_i\right), \tag{20.3}$$

$$\delta_i = \text{logit}\left(1 - s_i\right) - \text{logit}\left(g_i\right), \tag{20.4}$$

where $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$. Thus, Eq. (20.2) can be reformulated as:

$$\text{logit}\left(P\left(Y_{ji} = 1\right)\right) = \beta_i + \delta_i \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}, \tag{20.5}$$

where β_i and δ_i are the item intercept and interaction parameter respectively in this reparameterized DINA model (DeCarlo, 2011).

Further, to account for the dependency among assessed attributes that are often conceptually related and statistically correlated, a higher-order latent structure (de la Torre & Douglas, 2004) could be formulated, as follows.

$$\text{logit}\left(P\left(\alpha_{jk} = 1\right)\right) = \gamma_k \theta_j - \lambda_k, \tag{20.6}$$

where $P(\alpha_{jk} = 1)$ is the probability of person j mastering attribute k; θ_j is a general (higher-order) ability of person j; γ_k and λ_k are the slope and intercept parameter for attribute k, respectively. The higher the θ value, the higher the mastery probability of attribute k (assuming a positive slope). The use of a higher-order structure alleviates the computational burden by reducing the number of model parameters to be estimated, explains the correlations among attributes, and estimates an overall performance for every test-taker in addition to an attribute profile of mastery. Other options for a more parsimonious skill distribution are given by Xu and von Davier (2008a, 2008b).

20.2.1.2 Level 2 Models: Correlational Structures

Two correlational structures are formulated at level 2, one for item parameters and the other for person parameters. For the joint RT-DINA model, item parameters are assumed to follow a trivariate normal distribution with the mean vector and variance and covariance matrix specified as follows.

$$\boldsymbol{\Psi}_i = \begin{pmatrix} \beta_i \\ \delta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \end{pmatrix}, \boldsymbol{\Sigma}_i\right). \tag{20.7}$$

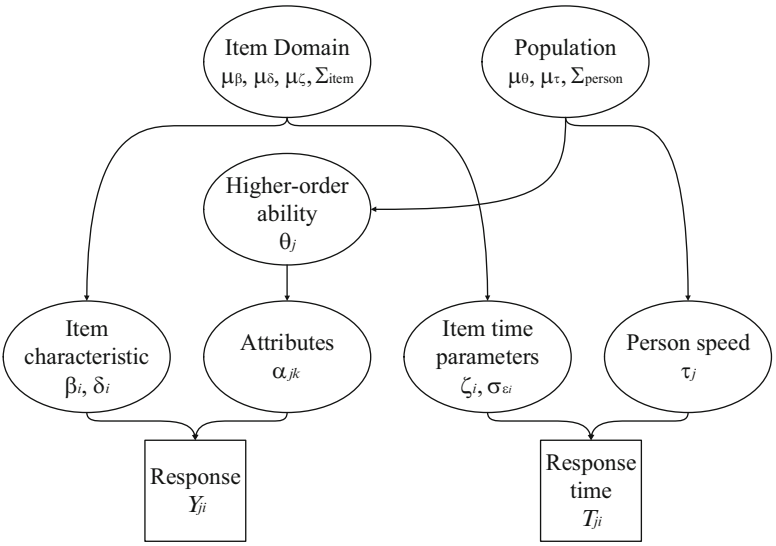


Fig. 20.1 A graphical representation of the joint RT-DINA model

Modeling correlations among item parameters captures the often-ignored relationship among guessing and slipping parameters in the DINA model (see Lee, de la Torre, & Park, 2012; Zhan, Jiao, Liao, & Bian, 2017). The error variance, $\sigma^2_{\varepsilon_i}$, is assumed to be independently distributed, thus it is not included in $\boldsymbol{\Psi}_i$.

Similarly, person parameters of the joint RT-DINA model are assumed to follow a bivariate normal distribution:

$$\boldsymbol{\Theta}_j = \begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\tau \end{pmatrix}, \boldsymbol{\Sigma}_j \right), \quad \boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma^2_\theta & \rho_{\theta\tau}\sigma_\theta\sigma_\tau \\ \rho_{\theta\tau}\sigma_\theta\sigma_\tau & \sigma^2_\tau \end{pmatrix}. \tag{20.8}$$

To sum up, Eqs. (20.1), (20.2), (20.3), (20.4), (20.5), (20.6), (20.7), and (20.8) together are the joint RT-DINA model. A graphical representation of the joint RT-DINA model is presented in Fig. 20.1 following van der Linden (2007).

For scale identification, three aspects need to be considered for this joint model: the identifiability between θ_j and τ_j, the identifiability between θ_j and attribute intercept (λ_k) and attribute slope (γ_k), and the identifiability between τ_j and ζ_i. Three constraints are set as $\mu_\theta = 0$, $\sigma_\theta = 1$, and $\mu_\tau = 0$, fixing the location and variation of the scale of the latent ability and the location of τ_j. The first two constraints set the scale and the third centers the scale for τ_j allows ζ_i to vary freely across items solving the third identifiability issue. In addition, four local independence assumptions are imposed: (a) latent attributes, α_{jk}, are conditionally independent given θ_j; (b) item responses, Y_{ji}, are conditionally independent given $\boldsymbol{\alpha}_j$; (c) the log response time, $\log(T_{ji})$, are conditionally independent given τ_j; (d) Y_{ji} and $\log(T_{ji})$ on item i are conditionally independent given all person parameters.

Some constraints could be further imposed such as $\gamma_k > 0$ indicating higher ability leads to higher probability of mastery; constraints such as $\delta_i > 0$, i.e., $g_i < 1 - s_i$ or $\text{IDI}_i > 0$ (e.g., Culpepper, 2015; DeCarlo, 2012; Henson et al., 2009; Junker & Sijtsma, 2001) need further exploration.

20.2.2 Bayesian Parameter Estimation

Bayesian estimation using the Markov chain Monte Carlo (MCMC) method can be used to estimate parameters in the joint RT-DINA model. In Bayesian estimation, prior distributions of model parameters and the observed data likelihood functions lead to a joint posterior distribution for the model parameters. Zhan et al. (2018) used the *JAGS* and the *R2jags* package (Version 0.5–7; Su & Yajima, 2015) in *R* (Version 3.3.1 64-bit; R Core Team, 2016) to estimate parameters. Model specification is illustrated as follows.

First, Y_{ji}, $\log(T_{ji})$, and α_{jk} are assumed conditionally and independently distributed and specified as follows:

$$Y_{ji} \sim \text{Bernoulli}\left(P\left(Y_{ji}=1\right)\right), \log\left(T_{ji}\right) \sim \text{Normal}\left(\zeta_i - \tau_j, \sigma^2_{\varepsilon_i}\right), \alpha_{jk} \sim \text{Bernoulli}\left(P\left(\alpha_{jk}=1\right)\right)$$

The priors of item parameters are assumed to follow a multivariate normal distribution with

$$\begin{pmatrix} \beta_i \\ \delta_i \\ \zeta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\beta \\ \mu_\delta \\ \mu_\zeta \end{pmatrix}, \boldsymbol{\Sigma}_i\right), \sigma^2_{\varepsilon_i} \sim \text{InvGamma}\,(1, 1)$$

where the hyper priors are specified as:

$$\mu_\beta \sim \text{Normal}\,(-2.197, 2)\,, \mu_\delta \sim \text{Normal}\,(4.394, 2) \quad I\,(\mu_\delta > 0)$$

$$\mu_\zeta \sim \text{Normal}\,(3, 2)\,, \boldsymbol{\Sigma}_i \sim \text{InvWishart}\,(\mathbf{R}, 3)$$

where $\mathbf{R}$ denotes a three-dimensional identity matrix. Hyper priors specified above are on a logit scale for both β and δ. The mean guessing effect is set at 0.1, which is approximately equivalent to a logit value of -2.197 for μ_β. With a standard deviation of $\sqrt{2}$ on the logit scale for μ_β, the assumed mean guessing effects range from 0.026 to 0.314. The mean slipping effect is also set at 0.1, indicating that μ_δ would approximately be 4.394 on the logit scale. With a standard deviation of $\sqrt{2}$ on the logit scale for μ_δ, the assumed mean slipping effects range from 0.007 to 0.653. The hyper prior specified above is on a log scale for ζ. Then, the mean RT is set at 20.086, which is equivalent to a log value of 3 for μ_ζ. With a standard deviation of $\sqrt{2}$ on the log scale for μ_ζ, the simulated mean RTs range from 4.883 s to 82.617 s.

In addition, the priors of person parameters are set as

$$\begin{pmatrix} \theta_j \\ \tau_j \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_j\right).$$

Please note that an inverse-Wishart prior cannot be used for the $\boldsymbol{\Sigma}_j$, because the variance of θ_j is set to 1 for identifiability. To solve this problem, $\boldsymbol{\Sigma}_n$ can be reparameterized in terms of its Cholesky decomposition as $\boldsymbol{\Sigma}_n = \boldsymbol{\Delta}_n \boldsymbol{\Delta}_n'$, where

$$\boldsymbol{\Delta}_n = \begin{pmatrix} 1 & 0 \\ \varphi & \psi \end{pmatrix}$$

is a lower triangular matrix with positive entries on the diagonal and unrestricted entries below the diagonal, and $\boldsymbol{\Delta}'_n$ is the conjugate transpose of $\boldsymbol{\Delta}_n$. Thus, the priors of elements in $\boldsymbol{\Delta}_n$ are specified as φ Normal(0, 1), ψ Gamma(1, 1).

Then, the priors of higher-order structure parameters are specified as:

$$\gamma_k \sim \text{Normal}\,(0, 4) \quad I\,(\gamma_k > 0)\,, \lambda_k \sim \text{Normal}\,(0, 4)$$

Finally, as a categorical value, the posterior mode of $\widehat{\alpha}_{jk}$ is treated as the estimated value, following Zhan et al. (2018). As an alternative, the posterior mean of $\widehat{\alpha}_{jk}$ (i.e., $\widehat{\alpha}_{jk} \equiv 1$ if $\widehat{\alpha}_{jk} > 0.5$, and $\widehat{\alpha}_{jk} \equiv 0$ otherwise) suggested by de la Torre and Douglas (2004) can also be used.

Given the priors specified above and each sampled model parameter given in S as follows:

$$S = \left\{\alpha_j, \theta_j, \tau_j, \lambda_k, \gamma_k, \beta_i, \delta_i, \zeta_i, \mu_\beta, \mu_\delta, \mu_\zeta, \boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j, \sigma^2_{\varepsilon_i}\right\},$$

The joint posterior probability for the joint RT-DINA model can be expressed as follows:

$$\begin{aligned} P\,(S|\mathbf{Y}, \log(\mathbf{T})) \propto\; & L\left(\mathbf{Y}, \log(\mathbf{T})\,|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\zeta}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}}\right) \\ & \times P\,(\boldsymbol{\alpha}|\boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\theta})\, P\,(\boldsymbol{\lambda})\, P\,(\boldsymbol{\gamma})\, P\left(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{0}, \boldsymbol{\Sigma}_j\right) P\left(\boldsymbol{\Sigma}_j\right) \\ & \times P\left(\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\zeta}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) P\left(\boldsymbol{\mu}_\beta\right) P\left(\boldsymbol{\mu}_\delta\right) P\left(\boldsymbol{\mu}_\zeta\right) P\,(\boldsymbol{\Sigma}_i)\, P\left(\boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}}\right), \end{aligned} \tag{20.9}$$

where $L\left(\mathbf{Y}, \log(\mathbf{T})\,|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\zeta}, \boldsymbol{\tau}, \boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}}\right) = \prod_{j=1}^{J}\prod_{i=1}^{I} P\left(Y_{ji}|\boldsymbol{\alpha}_j, \beta_i, \delta_i\right) f\left(\log\left(T_{ji}\right)\right.$ $\left.|\zeta_i, \tau_j, \sigma^2_{\varepsilon_i}\right)$ is the likelihood of the joint RT-DINA model.

20.3 Potential Extensions

The previous section demonstrated how RT can be combined with item responses for cognitive diagnosis using the lognormal RT model and DINA model as examples. As reviewed, RT could be modeled in multiple ways following different distributions such as Exponential, Gamma, Weibull, the Box-Cox normal model, Cox-proportional transformation, and linear transformation model; further, different CDMs such as DINA, DINO, and general CDMs (GDM, generalized DINA, and loglinear CDMs), could be formulated for different item response structures. Thus, it is quite possible that different models for RT and item responses could be used jointly to better reflect the nature of specific RT and item responses in an assessment. Further, the current demonstration is for dichotomous item responses. An extension to polytomous item response data (e.g., Ma & de la Torre, 2016; von Davier, 2008) awaits further exploration.

Recently, innovative assessment has been increasingly used in testing practice to measure higher-order thinking skills. Items in these assessments are often embedded in different contexts such as scenarios, passages, and graphs or tables; local item independence is likely to be violated. Different researchers (e.g., Jiao, Liao, & Zhan, 2018; Zhan, Liao, & Bian, 2018) have explored extended models taking into account local item dependence from one or multiple sources in joint modeling of RT and item responses for cognitive diagnosis.

As Embretson and Yang (2013) demonstrated, using a multicomponent or non-compensatory multidimensional latent trait model with linear logistic test model (LLTM) modeling item features to cognitive diagnosis, a multidimensional latent trait model with RT could be another potential extension to provide cognitive diagnosis based on latent trait-based model and/or item feature-based model for cognitive diagnosis.

20.4 Summary and Discussion

It is expected that the inclusion of RT in addition to item responses in psychometric modeling provides additional sources of information which may improve the latent trait estimation and attribute estimation in cognitive diagnosis. Both Minchen and de la Torre (2016) and Zhan et al. (2018) indicate that when a test has an adequate test length containing items of good quality (i.e., with low guessing and slipping effects) and the Q-matrix is identifiable, the use of RT did not provide added value in improving the precision in model parameter estimation and the accuracy for attribute mastery for cognitive diagnosis. This is not beyond expectation as the information containing in item responses would be sufficient to provide accurate cognitive diagnosis. However, when a test was not ideally designed for cognitive diagnosis, i.e., the Q-matrix is not identified, the test length is short, and items are of low quality with high guessing and slipping effects, the improvement in

model parameter estimation is evident. The joint RT-DINA model provides a good starting point for future exploration in how to utilize RT in CDM and using RT and item responses simultaneously in aberrant responding behavior detection such as a mixture version of joint RT-CDM.

Though the use of process data and RT could be beneficial to cognitive diagnosis, the added value of integrating process data strongly depends on whether the process data are substantiated and meaningful from a theoretical perspective. Goldhammer et al. (2014) reported that the interpretation of RT is intricate, and the statistical considerations should not be separated from substantial considerations (task-related, domain-specific, psychological).

Modeling RT is not a simple task. Modeling RT and item responses jointly takes into account the dependency between RT and item responses. This is one type of within-subject dependency. Many researchers have already noticed the challenges in RT modeling. Though different distributions or models have been proposed for RT, some of these perspectives simplify the relationship between RT and item responses in real-world applications. One of the challenges is related to the within-subject differential speed effects. That is, some test-takers may demonstrate different patterns of speed effects depending on ability and/or item difficulty. Different models and approaches to dealing with this within-subject dependency have been explored (e.g., Bolsinova et al., 2016, 2017; Fox & Marianti, 2016; Liao, 2018; Molenaar, Bolsinova, Rozsa, et al., 2016; Molenaar, Oberski, et al. 2016; Molenaar, Bolsinova, & Vermunt, 2018). However, none of the research integrating RT and responses has integrated this more complex within-subject dependency into cognitive diagnostic modeling. This could be a future exploration.

Further, other estimation methods such as the maximum likelihood estimation method for the joint RT-CDM could be explored in future studies. The application of RT-CDM in computerized adaptive test Finkelman, et al. (2014) would be worthy of more extensive investigation. Using process data other than RT awaits further exploration for cognitive diagnosis. Further, how to integrate data from multiple sources for cognitive diagnosis in linear and adaptive test delivery algorithms deserves more investigation.

Recently, other methods have been proposed for analyzing process data in large-scale assessments. For instance, Liu and Cheng (2018) explored support vector machine (SVM), a popular supervised learning method to conduct cognitive diagnosis given a training dataset and found that SVM provided as least comparable attribute and profile classification accuracy with small sample sizes as the traditional CDMs. Another example, He and von Davier (2015, 2016) conducted a case study for one problem-solving item using the response data and log files from PIAAC 2012. They treated consecutive actions in the log files as a sequence and utilized two n-gram model and two feature selection methods, chi-square statistics and weighted log-likelihood ratio test to identify sequences of features from process data for comparisons among performance groups and across countries. Though this method has not been used for cognitive diagnosis, it is worthy of exploration. Given the advantages of artificial intelligence, more and more researchers explore to apply machine (or deep) learning algorithms for large-scale assessment data analysis. In

the near future, as the availability and types of process data grow, the joint modeling approach may not be the only method in handling the computational complexity. Thus, machine learning algorithms are likely to bloom for cognitive diagnosis.

Overall, while taking advantage of the auxiliary information provided in process data, caution should be exercised in that whether the integration of multiple data sources from assessment really serves the assessment purpose better or it introduces more noise in the decoding of assessment data awaits further exploration. More empirical investigation and theoretical justification of integrating item response data and process data should be provided to address the validity considerations when using process data such as RT as auxiliary information.

References

Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology, 69*(1), 62–79. https://doi.org/10.1111/bmsp.12059

Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics, 41*(2), 123–145. https://doi.org/10.3102/1076998616631746

Bolsinova, M., & Tijmstra, J. (2017). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology, 71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2016). Modeling conditional dependence between response time and accuracy. *Psychometrika*, 1–23.https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology, 70*(2), 257–279. https://doi.org/10.1111/bmsp.12076

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics, 40*, 454–476. https://doi.org/10.3102/1076998615595403

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*, 620–639. https://doi.org/10.1177/0146621608326423

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26. https://doi.org/10.1177/0146621610377081

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*, 447–468. https://doi.org/10.1177/0146621612449069

Embretson, S. E. (2015). The multicomponent latent trait model for diagnosis: Applications to heterogeneous test domains. *Applied Psychological Measurement, 39*(1), 16–30. https://doi.org/10.1177/0146621614552014

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*, 14–36. https://doi.org/10.1007/s11336-012-9296-y

Finkelman, M., Kim, W., Weissman, A., & Cook, R. (2014). Cognitive diagnostic models and computerized adaptive testing: Two new item-selection methods that incorporate response times. *Journal of Computerized Adaptive Testing, 2*(3), 59–76. http://dx.doi.org/10.7333/1412-0204059

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128

Fox, J. P., & Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54*(2), 243–262. https://doi.org/10.1111/jedm.12143

Gaviria, J. L. (2005). Increase in precision when estimating parameters in computer assisted testing using response time. *Quality & Quantity, 39*(1), 45–69. https://doi.org/10.1007/s11135-004-0437-y

Glas, C. A., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology, 63*(3), 603–626. https://doi.org/10.1348/000711009X481360

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*(3), 608–626.

Goldhammer, F., Naumann, J., & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence, 3*(1), 21–24. https://doi.org/10.3390/jintelligence3010021

Haberman, J. S., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*, 331–354. https://doi.org/10.1007/s11336-010-9158-4

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In *Quantitative psychology research* (pp. 173–190). Cham, Switzerland: Springer.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In *Handbook of research on technology tools for real-world skill development* (pp. 750–777). Hershey, PA: IGI Global.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using loglinear models with latent variables. *Psychometrika, 74*, 191–210. https://doi.org/10.1007/s11336-008-9089-5

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment, 4*(2), 170–173.

Jiao, H., Zhan, P., Liao, M., & Man, K. (2017, November). *A joint multigroup testlet model for responses and response time accounting for differential item and speed functioning*. Presented at the fifth conference on the statistical methods in psychometrics. New York: Columbia University.

Jiao, H., Liao, M., & Zhan, P. (2018, April). *Cognitive diagnostic modeling using responses and response times for items embedded in multiple contexts*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New York City, NY.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Klein Entink, R. H., van der Linden, W. J., & Fox, J. P. (2009). A Box–Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*(3), 621–640. https://doi.org/10.1348/000711008X374126

Lee, S. Y., & Wollack, J. (2017). *Use of response time for detecting security threats and other anomalous behaviors.* Paper presented at the Timing Impact on Measurement in Education conference, Philadelphia, PA.

Lee, Y.-S., de la Torre, J., & Park, Y. S. (2012). Cognitive diagnosticity of IRT-constructed assessment: An empirical investigation. *Asia Pacific Education Review, 13*, 333–345.

Liao, D. (2018). *Conditional joint modeling of response time and response accuracy for speed-accuracy-difficulty interaction*. Unpublished doctoral dissertation, University of Maryland, College Park.

Liu, C., & Cheng, Y. (2018). An application of the support vector machine for attribute-by-attribute classification in cognitive diagnosis. *Applied Psychological Measurement, 42*, 58–72. https://doi.org/10.1177/0146621617712246

Locke, E. A. (1965). Interaction of ability and motivation in performance. *Perceptual and Motor Skills, 21*, 719–725. https://doi.org/10.2466/pms.1965.21.3.719

Loeys, T., Rosseel, Y., & Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika, 76*(3), 487–503. https://doi.org/10.1007/s11336-011-9211-y

Logan, S., Medford, E., & Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. *Learning and Individual Differences, 21*, 124–128. https://doi.org/10.1016/j.lindif.2010.09.011

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*, 253–275. https://doi.org/10.1111/bmsp.12070

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99–120. https://doi.org/10.3102/10769986002002099

Man, K., Jiao, H., & Ouyang, Y. (2016, April). *Response time based nonparametric person fit index for aberrant response behavior detection in large-scale assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, DC.

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523–547. https://doi.org/10.1007/BF02294327

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212. https://doi.org/10.1007/BF02294535

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615–633. https://doi.org/10.1007/s11336-012-9288-y

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement, 52*, 1–27. https://doi.org/10.1111/jedm.12060

Minchen, N. (2017). *Continuous response in cognitive diagnosis models: Response time modeling, computerized adaptive testing, and Q*-Matrix *validation*. Unpublished doctoral dissertation. Rutgers, The State University of New Jersey.

Minchen, N. D., & de la Torre, J. (2016, April). *Using response time in cognitive diagnosis models*. Poster presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research, 50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684

Molenaar, D., Bolsinova, M., Rozsa, S., & De Boeck, P. (2016). Response mixture modeling of intraindividual differences in responses and response times to the Hungarian WISC-IV Block Design test. *Journal of Intelligence, 4*(3), 10–29. doi:10.3390/jintelligence4030010

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research, 51*(5), 606–626. https://doi.org/10.1080/00273171.2016.1192983

Molenaar, D., Bolsinova, M., & Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 205–228. https://doi.org/10.1111/bmsp.12117

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice, 35*, 38–47. https://doi.org/10.1111/emip.12102

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ranger, J., & Kuhn, J. T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement, 36*(3), 214–231. https://doi.org/10.1177/0146621612439796

Ranger, J., & Kuhn, J. T. (2013). Analyzing response times in tests with rank correlation approaches. *Journal of Educational and Behavioral Statistics, 38*(1), 61–80. https://doi.org/10.3102/1076998611431086

Ranger, J., & Kuhn, J. T. (2014). An accumulator model for responses and response times in tests based on the proportional hazards model. *British Journal of Mathematical and Statistical Psychology, 67*(3), 388–407. https://doi.org/10.1111/bmsp.12025

Ranger, J., & Ortner, T. (2012). A latent trait model for response times on tests employing the proportional hazards model. *British Journal of Mathematical and Statistical Psychology, 65*(2), 334–349. https://doi.org/10.1111/j.2044-8317.2011.02032.x

Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment, 8*(4). Retrieved [2019-01-29] from https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1623

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Hillsdale, NJ: Lawrence Erlbaum Associates.

Su, Y.-S., & Yajima, M. (2015). *R2jags: Using R to run 'JAGS'*. R package version 0 (pp. 5–7). Retrieved from http://CRAN.R-project.org/package=R2jags

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204. https://doi.org/10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika, 75*(1), 120–139. https://doi.org/10.1007/s11336-009-9129-9

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika, 68*(2), 251–265. https://doi.org/10.1007/BF02294800

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement, 34*(5), 327–347. https://doi.org/10.1177/0146621609349800

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology, 70*(2), 317–345. https://doi.org/10.1111/bmsp.12101

von Davier, M. (2005). *A general diagnostic model applied to language testing data (ETS research rep. No. RR-05-16)*. Princeton, NJ: ETS.

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*, 49–71. https://doi.org/10.1111/bmsp.12003

von Davier, M., & Yamamoto, K. (2004). *A class of models for cognitive diagnosis*. Presented at the Fourth Spearman Conference, Philadelphia, PA.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68*, 456–477. https://doi.org/10.1111/bmsp.12054

Wang, C., Chang, H.-H., & Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*, 144–168. https://doi.org/10.1111/j.2044-8317.2012.02045.x

Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics, 38*(4), 381–417. https://doi.org/10.3102/1076998612461831

Wang, T. (2006). *A model for the joint distribution of item response and response time using a one-parameter Weibull distribution* (Center for Advanced Studies in Measurement and Assessment Research Report, no. 20). Iowa City, IA: University of Iowa.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323–339. https://doi.org/10.1177/0146621605275984

Xu, X., & von Davier, M. (2008a). *Fitting the structured general diagnostic model to NAEP data* (RR-08-27, ETS Research Report).

Xu, X., & von Davier, M. (2008b). *Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model* (RR-08-35, ETS Research Report).

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83–105. https://doi.org/10.1177/0146621606291559

Zhan, P., Jiao, H., Liao, M., & Bian, Y. (2018). Bayesian DINA modeling incorporating within-item characteristics dependency. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621618781594

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*, 262–286. https://doi.org/10.1111/bmsp.12114

Zhan, P., Liao, M., & Bian, Y. (2018). Joint testlet cognitive diagnosis modeling for paired local item dependence in response times and response accuracy. *Frontiers in Psychology, 9*, 609. https://doi.org/10.3389/fpsyg.2018.00607

Chapter 21
Application of Cognitive Diagnostic Models to Learning and Assessment Systems

Benjamin Deonovic, Pravin Chopade, Michael Yudelson, Jimmy de la Torre, and Alina A. von Davier

Abstract Over the past few decades, cognitive diagnostic models have generated a lot of interest due in large part to the call made by the No Child Left Behind Act of 2001 (No Child Left Behind, Act of 2001 Public Law No. 107–110, § 115. Stat, 1425, 2002) for more formative assessments in learning systems. In this chapter, we provide an overview of learning and assessment systems, including the rise in popularity of online and personalized learning systems; we contrast the role of summative and formative assessments in learning systems; and we provide a review of cognitive diagnostic models and the challenges of retrofitting models to data not designed for cognitive diagnostic models.

21.1 Introduction

Learning, broadly defined as the acquisition of knowledge, skills, values, beliefs, and habits through experience, study, or instruction, takes place within a variety of frameworks. The framework of interest in this chapter is a learning system, which also takes shape in various forms and formats. Traditional examples include schools and textbooks, but computers and online forums can also be utilized as learning systems. Over the past few decades, cognitive diagnostic models have generated a lot of interest due in large part to the call made by the No Child Left Behind Act of 2001 (No Child Left Behind, 2002) for more formative assessments in learning systems. In this chapter, we provide an overview of learning and

B. Deonovic · P. Chopade · M. Yudelson · A. A. von Davier (✉)
ACTNext ACT Inc., Iowa City, IA, USA
e-mail: Benjamin.Deonovic@act.org; Pravin.Chopade@act.org; Michael.Yudelson@act.org; Alina.vonDavier@act.org

J. de la Torre
Division of Learning, Development and Diversity, University of Hong Kong, Hong Kong, China
e-mail: j.delatorre@hku.hk

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_21

assessment systems, including the rise in popularity of online and personalized learning systems; we contrast the role of summative and formative assessments in learning systems; and we provide a review of cognitive diagnostic models and the challenges of retrofitting models to data not designed for cognitive diagnostic models. We conclude the chapter with a novel, personalized, online learning system developed by ACTNext, the ACTNext Educational Companion App, which utilizes concepts from cognitive diagnostic models.

21.2 Learning Systems

21.2.1 Smart Learning

Zhang and Chang (2016) highlight the rise in interest of personalized learning and assessment models citing the push made by industry leaders such as IBM to utilize technology to help shape the future of education (Palmisano, 2008; Rudd, Davia, & Sullivan, 2009). The concept of driving innovation in education by utilizing technology has been referred to as 'Smart Education' and has been described by Zhu and Shen (2013) as "creating a smart environment for learning that promotes the development of both the learner's well-roundedness and specialized competency, which, ultimately, will create value for the entire society". Advances in technology have played a pivotal role in providing personalized and equitable learning to everyone.

21.2.2 Online Learning

One technological advancement that has been a great advantage to personalized learning in the past few decades is online learning. Online learning is a learning system that is entirely virtual. In secondary school education, Khan Academy (Dijksman & Khan, 2011), a free online resource that offers instructional videos and interactive tasks, has been a significant contributor to online learning resources. Online learning, as offered by the Khan Academy, and blended learning, a combination of traditional and online learning systems, are extensively employed in the USA. Another company that provides a repository for online learning material is OpenEd. OpenEd is the only K-12 educational resource library focused on aligning resources to learning objectives and a variety of standards including Common Core State Standards (CCSS) and Next Generation Science Standards (NGSS) by utilizing machine learning algorithms. According to Journell, McFadyen, Miller, and Brown (2014), more than one million secondary students take online courses from resources such as Khan Academy and OpenEd as part of their curriculum each year. The need to provide high-quality public education at a low cost is one of the key drivers of this phenomenon.

In higher education, the Massachusetts Institute of Technology (MIT) was one of the pioneering institutes in creating online learning systems. As early as 2001, MIT provided access to audio lectures and slides on their open courseware (OCW) website (Abelson, 2008). Many institutions around the world followed suit including Rice University, which created the Connexions Project in 1999, now known as OpenStax (OpenStax, 2018) and Carnegie Mellon University, which created the Open Learning Initiative (Open Learning Initiative, 2018). This sparked the open education resources (OER) movement, a movement to increase the availability of teaching, learning, and research materials in any medium (Atkins, Brown, & Hammond, 2007).

21.2.3 Learning and Assessment

Assessments are utilized to ascertain whether learners have indeed learned the objectives set out by a learning system, whether that is a traditional learning system or an online/adaptive learning system. An assessment is an instrument designed to observe performance in a learner and produce data used to draw inferences about the material which the learner has learned. Research in assessment focuses on ensuring that assessments are reliable (produce similar results under consistent conditions) and valid (the extent to which the assessment measures the underlying construct of interest).

21.2.3.1 Connection Between Learning and Assessment

Historically, there has been a lack of connection between assessments and learning, due to the use of assessments for ranking individuals, which is a different goal than that of learning systems, and due to the lack of ability to collect learning data. This led to a focus of psychometrics on models for assessment. Classical models in the assessment literature were primarily concerned with measuring an individual's latent ability without connecting this to a model for the content the learners were learning. In recent years, with the advances and affordances of technologies, there has been a renewed interest to connect assessment to learning. This idea of connecting assessment with learning is not new though. As far back as 1957, there have been aspirations to incorporate theories of cognition and learning into assessment (Cronbach, 1957). The past few decades especially have seen a revival of this concept (Embretson & Gorin, 2001; Gorin, 2006; Leighton, 2004; Leighton & Gierl, 2007; NRC, 2001; Pellegrino, Baxter, & Glaser, 1999). In 2001 the National Research Council (NRC) released a report entitled "Knowing what Students Know" (NRC, 2001) in which they assert the need to rethink the fundamental scientific principles of current approaches to assessment, and to broaden the assessment framework to incorporate advances in cognitive sciences as well as apply the expanded capabilities in psychometrics.

21.2.3.2 The Psychometric Perspective: Formative Assessment

Formative assessments have seen increased interest in recent years. Formative assessments monitor learning and provide the learner and/or the instructor with information to help guide learning/instruction. Examples include weekly quizzes or homework in a traditional learning system. At the other end of the spectrum are summative assessments. Summative assessments are used to evaluate how much an individual knows at a particular point in time. These assessments are often high stakes tests, such as a college readiness exam.

21.2.3.3 The Perspective of the Educational Data Mining and Learning Analytics & Knowledge Communities: Domain Modeling and Knowledge Modeling

While psychometricians focused on modeling assessment data, researchers in educational data mining (EDM) and learning analytics & knowledge (LAK) communities focused on modeling data collected during the process of learning directly. Although superficially the models built by psychometricians for assessment and the models built by researchers in the learning fields share the goal of ascertaining what the learner has learned, they have diverged significantly and are entirely different in practice.

The EDM/LAK tradition provides models for tracking learning and models for the learning content itself. This is referred to in a 2017 review of the field by Pelanek (2017) as knowledge modeling and domain modeling, respectively, and made explicit in the knowledge-learning-instruction (KLI) framework of Koedinger, Corbett, and Perfetti (2012). This framework connects learning processes such as fluency building, induction, refinement, understanding, and sense-making to knowledge components (i.e., skills/attributes). By defining models for the domain and for learning, researchers in these fields can make explicit the link between the domain and knowledge model, achieving the same goals as those set up by formative assessments, that is, to ascertain what the learner knows.

Here we briefly describe the learning progressions, knowledge maps, and the Q-matrix as three approaches to domain modeling. The Q-matrix representation has been mostly used in psychometric modeling.

21.2.3.4 Learning Progressions

Once a framework is in place to connect the domain and knowledge model, the question remains of how to help a learner through the steps necessary to learn new skills. One answer to this question is the application of learning progressions. The concepts underlying learning progressions are relatively old, but the specific term was coined in the 2005 (NRC, 2005) NRC report, was featured in the 2007 report (NRC, 2007), and utilized soon after in an application to describe possible levels

of student development of skills and concepts for science assessment (Corcoran, Mosher, & Rogat, 2009).

A learning progression refers to the sequencing of learning materials and resources across time (e.g., developmental stages, ages, or grade levels). It dictates what skills should be taught to learners at a particular point in time based on the learners' ability and the skills they already have mastered. In the context of assessments, learning progressions can be incorporated into a Rasch/IRT (Item Response Theory) model using Wright maps (Wilmot, Schoenfeld, Wilson, Champney, & Zahner, 2011; DiBello & Stout, 2007). A Wright map is a graphic which overlays the latent ability distribution of the learners with the locations of items (Wilson, 2005). This organization suggests a particular traversal of the items based on a person's ability.

21.2.3.5 Knowledge Maps

A concept similar to learning progressions in the field of EDM/LAK is the idea of knowledge maps or domain models (Pelanek, 2017). Knowledge maps reflect the assignment of individual items to particular knowledge components and with the modeling of the relationship between different knowledge components (e.g., prerequisite information). A simple construction of a knowledge map is one in which the knowledge concepts are considered independent, disjoint sets of items. Pelanek (2017) extends this idea in three main directions: multiple knowledge concepts per item, a hierarchy of knowledge concepts (which allows for capturing skills of different granularity), and a directed graphical representation of the knowledge concepts that capture their prerequisite structure. These formations of knowledge maps can be modeled using Bayesian networks (Millan, Loboda, & Perez-de-la-Cruz, 2010; Conati, Gertner, & Vanlehn, 2002; Käser, Klingler, Schwing, & Gross, 2014; Carmona, Millán, Pérez-de-la-Cruz, Trella, & Conejo, 2005) or knowledge space theory (Doignon & Falmagne, 2012).

21.2.3.6 Q-Matrix and Multidimensionality of Latent Ability

Psychometric models used to analyze summative assessments tend to utilize unidimensional latent variable models, in which the latent variable represents the learner's ability. Such models are suitable to analyze summative assessments, but may be less useful for formative assessments, and for other learning systems, in which one is interested in observing and manipulating learning. A unidimensional measure of a person's ability does not lend itself to inferring what aspects of the material the individual has mastered or not mastered. For a formative assessment to be successful, it needs a statistical model that is capable of identifying specific aspects of the material particularly difficult for the learner.

Psychometric models suitable for analyzing formative assessments are cognitive diagnostic models (CDMs). An integral aspect of the CDM is the Q-matrix (Embret-

son, 1984; Tatsuoka, 1985). A Q-matrix is a mapping which skills or attributes are tested by an item or which skills or attributes are required to successfully complete an item on an assessment. In this way, it serves a similar function as the domain models and knowledge maps of EDM/LAK.

21.3 Models for Learning Systems

Statistical models can be used to track and ascertain the learning that occurs in a learning system. The models described in this section can be used to track both the current state of student knowledge as well as the process of knowledge progression (learning) by administering assessments to the learner. In the following section, the models mentioned above that are utilized in learning system are described. The two models including Bayesian Knowledge Tracing (BKT), the de-facto standard approach used in intelligent tutoring systems (ITSs) and CDMs, a model suggested by educational measurement research for (formative) assessments are discussed. Concepts from both of these models are utilized by the ACTNext Educational Companion App that is illustrated later in the chapter.

21.3.1 Overview of Bayesian Knowledge Tracing Used for Learning

The first paper describing Bayesian Knowledge Tracing (BKT) was published in 1995 (Corbett & Anderson, 1995). Since then, this approach has been widely used in the field of ITS. BKT uses a formalism of Hidden Markov Models (HMM) (Levinson, Rabiner, & Sondhi, 1983) to model student learning. For every student-skill combination, there is a separate BKT instance that relates an observed sequence of successful/failed attempts to apply a skill (binary variable) to a hidden binary variable capturing whether a student mastered this skill. One of BKT's main goals is to provide an estimate of mastery because mastery is a latent variable.

An illustration of a BKT model for a single skill is shown in Fig. 21.1. Here, unobserved states of the skill mastery at various time slices are shown as empty circles; observed performance nodes are shown as shaded circles. As per the Markovian assumption, the current level of skill mastery depends only on the previous state. Students' performance depends only on the current state of skill mastery.

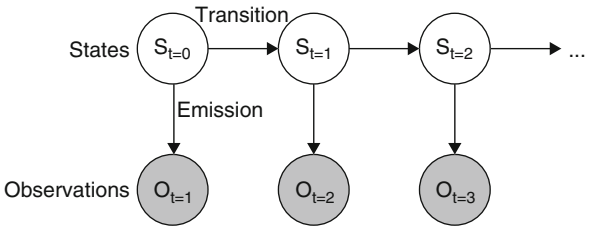


Fig. 21.1 An unrolled view of a BKT model

The standard BKT model has the following parameters.

- p-init or $p(L_0)$ – is the probability the skill was mastered a priori.
- p-learn or p(T) – is the probability the skill will transition into the mastered state after a practice attempt.
- p-forget or p(F) – is the probability that the skill will transition into the un-mastered state after a practice attempt. Traditionally, p(F) is set to zero – there is no forgetting in standard BKT.
- p-slip or p(S) – is the probability that a student fails to apply a mastered skill.
- p-guess or p(G) – is the probability that an un-mastered skill will be applied correctly.

Given that it is assumed that the skill mastery is not forgotten, that is p(F) is set to zero, there are four parameters for every student-skill combination in standard BKT. In addition to the parameters described above, there are also two more parameters that are frequently mentioned. First, the running estimate of the probability that student has mastered the skill: p(L) or p-mastery. Second, the expected value of student responses being correct: p-correct, or p(C). The parameters of the BKT model are estimated using the EM algorithm or a brute force grid search. The BKT model is described by Eqs. (21.1a), (21.1b), (21.1c), (21.1d) and (21.1e).

$$p\left(L_1\right) = p\left(L_0\right) \tag{21.1a}$$

$$P\left(L_{t+1}|e = correct\right) = \frac{p\left(L_t\right) \cdot \left(1 - p(S)\right)}{p\left(L_t\right) \cdot \left(1 - p(S)\right) + \left(1 - p\left(L_t\right)\right) \cdot p(G)} \tag{21.1b}$$

$$P\left(L_{t+1}|e = wrong\right) = \frac{p\left(L_t\right) \cdot p(S)}{p\left(L_t\right) \cdot p(S) + \left(1 - p\left(L_t\right)\right) \cdot \left(1 - p(G)\right)} \tag{21.1c}$$

$$p\left(L_{t+1}\right) = p\left(L_{t+1}|e\right) + \left(1 - p\left(L_{t+1}|e\right)\right) \cdot p(T) \tag{21.1d}$$

$$p\left(C_{t+1}\right) = p\left(L_t\right) \cdot \left(1 - p(S)\right) + \left(1 - p\left(L_t\right)\right) \cdot p(G) \tag{21.1e}$$

Equation (21.1a) sets the running estimate of the mastery to the prior in the beginning. Equations (21.1b) and (21.1c) define conditional probability given evidence (e) of student being right or wrong. Equation (21.1d) shows how to update the running estimate of mastery using the conditional probability. Finally, Eq. (21.1e) is a conversion of the running estimate of latent mastery into the probability of a correct response on the next opportunity to apply the skill.

21.3.2 Overview of Traditional CDM for (Formative) Assessment

One class of models used to identify mastery of skills/attributes in assessment data are CDMs. CDMs are multivariate, discrete latent variable models developed primarily to identify the mastery, or lack thereof, of skills (or more generically, *attributes*) measured in a particular domain. Two features distinguish CDMs when compared to traditional item response models, namely, the finer-grained nature of the inferences that can be derived from the models, and the interpretability and relevance of these inferences to the student learning process.

At its core, we use CDMs to summarize the relationship in the response vector $\boldsymbol{X}_i = (X_{i1}, \ \ldots \ , X_{ij}, \ \ldots \ , X_{iJ})$ using a lower dimensional discrete latent variable $\boldsymbol{A}_i = (A_{i1}, \ \ldots \ , A_{ik}, \ \ldots \ , A_{iK})$ where X_{ij} represents the response of the ith individual to the jth item and A_{ik} is a discrete latent variable for individual i for latent dimension k. Specifically, we use CDMs to set up the model $P(\boldsymbol{X}_i|\boldsymbol{A}_i)$, the conditional probability of observing $\boldsymbol{X}_i$ given $\boldsymbol{A}_i$. By assuming conditional independence, we can write this probability as given in Eq. (21.2),

$$P\left(\mathbf{X}_i|\mathbf{A}_i\right) = \prod_{j=1}^{J} P\left(X_{ij}|\mathbf{A}_i\right) \tag{21.2}$$

where $P(X_{ij}|\boldsymbol{A}_i)$ represents a particular CDM of the response of individual i to item j. The specific form of the CDM depends on the assumptions we make regarding how the elements of $\boldsymbol{A}_i$ interact to produce the probabilities of response X_{ij}.

A wide range of CDMs have been proposed in the psychometric literature, Chaps. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 in this volume cover many of the more commonly applied CDMs. Although various ways of classification exist, particularly for CDMs that involve binary attributes, these models can be differentiated based on whether or not they assume a particular process that underlies how the attributes interact with the item characteristics to produce the item responses. CDMs that do model the interaction of attributes and item characteristics are referred to as reduced, or specific CDMs; those that do not are referred to as saturated or general CDMs. Examples of the former are the *deterministic inputs,*

noisy "and" gate (DINA) (Haertel, 1989; Junker & Sijtsma, 2001; de la Torre, 2009) model, the *deterministic input, noisy "or" gate* (DINO) (Templin & Henson, 2006; Templin, 2016) model, the *linear logistic model* (LLM) (Hagenaars, 1993; Maris, 1999), the *reduced reparametrized unified model* (rRUM) (Hartz, 2002; Dibello, Roussos, & Stout, 2007), the *additive-CDM* (A-CDM) (de la Torre, 2011); examples of the latter are the *generalized DINA* (G-DINA) (de la Torre, 2011) model, the *general diagnostic model* (GDM) (von Davier, 2005), and the *log-linear CDM* (LCDM) (Henson, Templin, & Willse, 2009). In addition to accommodating a wider range of underlying processes, constraining general CDMs reduces them to specific CDMs. There are advantages to using one type of CDM over the other. For example, general CDMs require fewer assumptions, so they are more likely to fit the data, however, they are also more like to exhibit identification issues. On the other hand, reduced CDMs can be used with smaller sample sizes and provide for more straightforward interpretation (Huebner, 2010).

In practice, instead of picking a general CDM or one particular reduced CDM, a reasonable compromise is to consider using different CDMs for different items. However, deciding a priori correct CDM to use for an item is a difficult, if not thorny, undertaking. Rather, the task can be accomplished empirically once test data become available. The Wald test (de la Torre, 2011; de la Torre & Lee, 2013) is a statistical procedure that can be used to determine the appropriate item level CDM. Compared to the likelihood ratio test, Akaike Information Criterion (AIC), or Bayesian Information Criterion (BIC), which are used for test-level comparisons, the Wald test is an item-level procedure that determines the best CDM for an item by statistically comparing the fit of the G-DINA model with those of a number of reduced CDMs. Consequently, the Wald test allows multiple CDMs to be used within a single test simultaneously. Ma, Iaconangelo, and de la Torre (2016) have shown that employing that Wald test to determine the most appropriate CDM for each item can result in a higher classification accuracy compared to fitting a single general or potentially incorrect reduced CDM to all items.

21.3.2.1 Q-Matrices for the CDM

An implicit, yet integral component of a CDM specification is the Q-matrix. The jth row of the Q-matrix identifies which specific elements of $\boldsymbol{A}_i$ are involved in answering item j. In typical CDM applications, Q-matrices are constructed by domain experts. As such, Q-matrix construction involves some degree of subjective judgment, resulting in potential misspecifications. These misspecifications can affect the quality of item parameter estimates, and ultimately the examinee classification accuracy (de la Torre, 2008; Rupp & Templin, 2008). For this reason, expert-based Q-matrices need to be empirically validated to engender a greater degree of confidence in the inferences derived from CDMs.

21.3.2.2 Q-Matrix Validation

One empirical Q-matrix validation procedure that can be used with a wide class of CDMs is the method based on the G-DINA discrimination index (GDI) proposed by de la Torre and Chiu (2016). Given a provisional Q-matrix and item response data, the procedure searches for the correct q-vector for each item. A q-vector is deemed correct if it is the simplest q-vector and the proportion of variance accounted for (PVAF) by the q-vector is high relative to the maximum GDI of the item. The mesa plot of the GDIs, which combines quantitative and graphical GDI information, can be used to complement the PVAF (de la Torre & Ma, 2016). Another approach to Q-matrix validation using regularized ML can be found in Chap. 12 of this volume.

21.3.2.3 Q-Matrix Retrofitting

The optimal use of CDMs is in conjunction with cognitively diagnostic assessments (i.e., assessments specifically designed using a CDM framework; de la Torre & Minchen, 2014). However, for various reasons, some applications necessitate retrofitting CDMs to existing test data. A number of challenges can arise from such applications, particularly when the breadth of the domain is relatively wide. These challenges include the attributes with coarse granularity (or equivalently, a large number of the attributes), items of poor quality, the q-vectors lacking variability, and the identifiability of model parameters under the validated or retrofitted Q-matrix. When retrofitted data are involved, sufficient care needs to be taken to ensure that a minimum examinee classification accuracy is attained for individual attributes or attribute vectors. As likely would be the case, additional information (e.g., extra test items, ancillary variables) that can supplement test data, at least for some examinees, may be needed to ensure that every examinee is reliably classified and any subsequent actions that would be taken are sufficiently warranted. These challenges are discussed in Haberman and von Davier (2006) and von Davier and Haberman (2014). The issue of identifiability is discussed in detail in Zhang (2014).

21.4 Applications

21.4.1 ACTNext Educational Companion App

A promising trend in learning systems is the move from a generalized, discrete, fixed time/place delivery method towards a personalized, continuous, mobile anytime/anywhere approach. Universities, school classrooms, and educational technology companies are offering mobile solutions that allow students to consume lecture materials, hold learning sessions, and take tests – all from the palm of their hand. However, many of these solutions are simply mobile-friendly applications of

web portals providing access to data stored in backend school/institution systems (grade books, learning management systems, etc.). Mobile devices hold the promise for richer, intelligent interactions that can be tailored adaptively to each learner. ACTNext, an innovative research arm of ACT, Inc. is working on a new mobile app called the Educational Companion App (ECA) to change that status quo. ECA makes the promise to deliver an integrated, comprehensive mobile learning experience. Students can review a fused perspective of their abilities drawn from a range of inputs including formative and high-stakes test results, social-emotional assessments, skill practice, and targeted quizzes. These results are seamlessly linked to a suite of *Skill Up* activities and open educational resources (OER) from ACT's OpenEd that allows students to practice the skills they have yet to master. The ECA leverages the ACT Holistic Framework (HF; Camara, O'Connor, Mattern, & Hanson, 2015) – a publically released, hierarchical set of skills and skill areas that covers all aspects of development: cognitive, emotional, cross-cutting and navigational. The ECA is in the pilot stage, and the overarching goal is to demonstrate an advanced adaptive technology that could drive significant scholastic improvement for the students. The ECA framework consists of six functional modules and is described in Fig. 21.2.

1. The Learning Analytics Platform (**LEAP**) is a data storage repository that holds a vast amount of student/learner data from a variety of platforms in its native format until it is needed.
2. LEAP leverages student/learner metadata to perform **data matching.** To deliver feedback to individual learners, we identify and connect all the data we have about them. Doing this effectively is a fundamental capability for the ACT ecosystem of products and services.
3. A **diagnostic model** based on a CDM is developed to identify a student's areas of weakness in the HF based on their available data in LEAP (Chopade et al., 2017; von Davier et al., 2017).
4. A **feedback model** is constructed on top of the diagnostic model. The feedback model uses information provided by the diagnostic model to provide learners with actionable feedback: What do they do next? For example, the feedback model provides a user with a *Skill Up* activity that would allow the user to

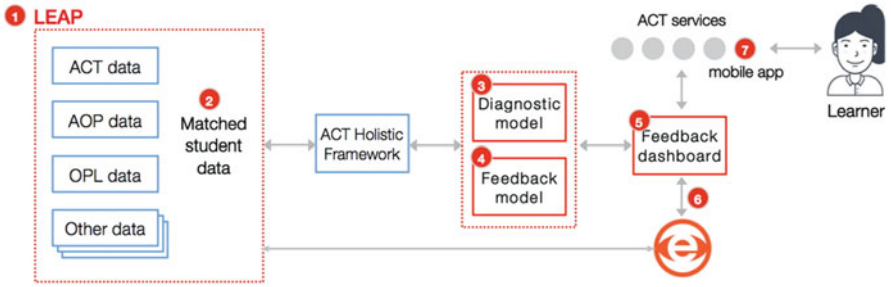


Fig. 21.2 ECA framework. (Chopade et al., 2017; von Davier et al., 2017)

practice the skills they have yet to master. This work is built directly on the psychometric and statistical development of the diagnostic model and on the HF.

5. A **feedback dashboard** provides users with the readable/understandable output. The dashboard appears in various services, platforms, modes, or apps as shown in Fig. 21.3, and it will go beyond mere reporting to include new information and perspectives on what an individual might explore or do next. Good user experience design is critical to this module because the learner must stay engaged and revisit as often as they need to for the app to be effective and help the student achieve mastery.
6. **Link feedback to ACT resources.** The OpenEd platform is an excellent basis for this because the content is well curated, and we can link the results of our feedback model to the appropriate resources. To our knowledge, only ACT is able to provide this level of integration.
7. **ACTNext ECA Mobile Interface.** This app demonstrates an advanced adaptive technology that could drive significant scholastic improvement for students using it. Figure 21.3 shows several ECA screenshots.

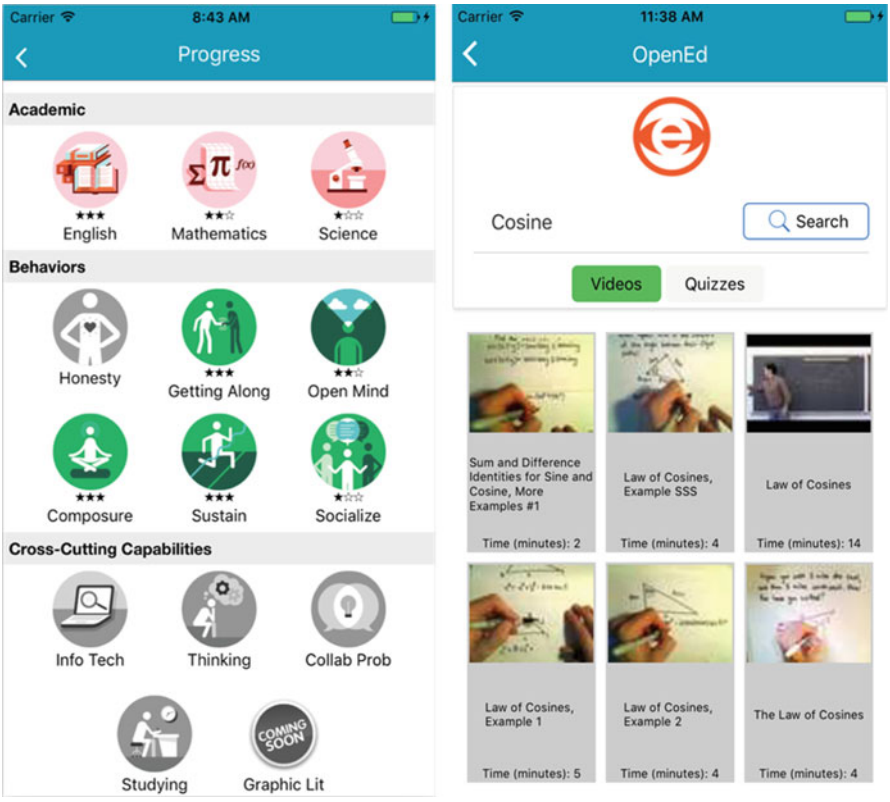


Fig. 21.3 ECA screenshots. (Chopade et al., 2017; von Davier et al., 2017)

The ECA faced three major research challenges which are described below. The first challenge was how to leverage a large bank of assessment data which has been tagged and associated with multiple sets of attributes. This is the mapping problem. The second challenge, after mapping and aligning the data properly, was designing a model capable of drawing inference from the data available to the ECA about users' skills and attributes. The solution utilized by the ECA is an extension of the standard IRT model which takes into account concepts from CDMs. The final challenge was how to validate this model. To validate the approach taken by the ECA we performed an intensive analysis of the data using the standard CDM approach.

21.4.2 Data

The ECA leverages rich test item metadata and student test-taking data. Test-taking data are comprised of tens of thousands of student test attempts across a set of testing solutions available from several testing instruments. The volume of test-taking data allows a large sample of students to be leveraged to build the skill-diagnostic model for the app.

Test item metadata comes in the form of indexing by attributes of, ideally, a single overarching taxonomy of attributes that covers the full spectrum of the knowledge students would be tested on. The HF is such a taxonomy that, in addition to accounting for traditional competency-level constructs, identifies non-cognitive and behavioral constructs (e.g., socio-emotional). This taxonomy allows for the construction of a Q-matrix for the items with skills/attributes corresponding to the HF taxonomy.

The HF is still relatively new and the metadata data that is most commonly available for the old items is from a legacy taxonomy that is a less fine-grained approach for indexing test items. To make HF indexing more widely available, three lines of activities were set up. First, some test items were manually indexed with HF attributes. Second, a crosswalk document, connecting the older legacy taxonomy to new HF taxonomy, was created. Third, the doubly-indexed test items and the crosswalk document were used to create a machine learning solution to automatically map the legacy taxonomy attributes to the HF attributes.

21.4.3 Mapping Problem

We formulated the mapping from the legacy taxonomy to HF as a bipartite mapping problem. Namely, we have graph $G = (U, V, E)$, where nodes U are all possible attributes from the legacy taxonomy, nodes V are all possible attributes from the

HF, and E are hyper-arcs (arcs that connect n nodes to m nodes, rather than 1 to 1 node) connecting one set of vertices from V and one set of vertices from U, i.e., the ith edge, e_i is of the form $e_i = \left(\mathbf{v}_{e_i}, \mathbf{u}_{e_i}\right)$, where $\mathbf{v}_{e_i}$ is the set of legacy taxonomy attributes that are mapped via edge e_i to $\mathbf{u}_{e_i}$ which is the set of all HF attributes mapped to by edge e_i Here, i is the index of all edges that runs from 1 to I, where I is the number of all edges.

The attributes from the HF and legacy taxonomy are structured hierarchically. For example, a legacy taxonomy attribute ADA-MAT-BOM denotes *Advanced algebra* ➔ *Matrices* ➔ *Basic operations on matrices*. The hierarchical nature of the attribute is coded in three nodes denoting ADA, ADA-MAT, and ADA-MAT-BOM – all three levels of the attribute hierarchy. The HF also has hierarchical attributes that are treated the same way.

As a result, we need to solve the equation $Y = F(X)$. Here, Y is a matrix of 0 s and 1 s of the size $I \times \ |U|$ that codes $\mathbf{u}_i$ parts of the hyper-arcs (from the HF), X is a matrix of 0 s and 1 s of the size $I \times \ |V|$ that codes $\mathbf{v}_i$ parts of the hyper-arcs (legacy attribute taxonomy), where I is the number of all cases from the training data (doubly-coded items and crosswalk document specifications), $|U|$ and $|V|$ is the total number of HF and legacy attributes, respectively. Every ith row of Y has 1's in positions denoting the elements of the HF attribute that are associated with the elements of the legacy attribute taxonomy specified by the 1's in the ith of matrix X.

The mapping function F is what we wanted to "learn." From the available machine learning approaches, we chose three: multinomial logistic regression (mlr), Support Vector Machines (SVM), and k-nearest neighbors. We used the mlr (Bischl et al., 2016) package from the CRAN R (R Core Team, 2017) statistical library. All three models predict a single binary variable (such as a HF tag), so, to achieve prediction of multiple binary variables (the whole pallet of HF tags) a form of multi-label classifier called a binary multi-label relevance learner was constructed using standard mlr functionality (Tsoumakas & Katakis, 2006). A binary relevance multi-label relevance learner creates multiple binary classifiers one for each different attribute in a taxonomy. It transforms the original data set of taxons U into |U| data sets that contain all examples of the original data set. The accuracy of the joint multi-label classifier is estimated as an average of the accuracies of each individual binary classifier.

The accuracy of the multinomial regression achieved 91.2% correct classification of training data. The accuracy of Support Vector Machine mapping was a little lower and reached 87.9%, and finally, k-nearest neighbors achieved the highest mapping accuracy of 93.8% with respect to correctly classification of the cases.

Due to the small size of the available doubly-coded set of items and the severely skewed density of the attribute counts the Y and X matrices were very sparse. As a result, we could not cross-validate our fitting models and our current models are overly specific (and could be slightly overfit). More work is being conducted at the time of writing this chapter by adding more doubly coded items.

21.4.4 Linear Logistic Test Model

To convert student test-taking records into student proficiency records, we devised a linear logistic test model (LLTM) by extending the Rasch model (Rasch, 1960). The formulation of the model is given below by Eqs. (21.3) and (21.4).

$$m_{ij} = 1 + \theta_i + \beta_j + \sum_{k=1}^{K} q_{jk} \cdot w_k \tag{21.3}$$

$$p_{ij} = \frac{1}{1 + e^{-m_{ij}}} \tag{21.4}$$

The model captures three classes of parameters: unidimensional student ability parameter θ_i, item difficulty (we defined it as item easiness) β_j, and w_k skill (HF tag) easiness; also q_{jk} captures the elements of the Q-matrix – the tagging of every assessment item to a set of skills. The probability of student i correctly responding to item j is expressed as p_{ij}. The model is set up as a logistic regression with dependent variable m_{ij} being the logit-transformation on student's binary (Bernoulli) response. Student abilities, item and skill easiness (a measure of how easy an item requiring a particular skill is relative to student ability) are treated as random factors based on the fact that they are sampled from the universe of possible conceptualizations of skills and items and the universe of available students where $\theta_i \sim N(0, \sigma_\theta)$, $\beta_j \sim N(0, \sigma_\beta)$, and $w_k \sim N(0, \sigma_w)$. We are continuing our work in this direction to increase the variance of student performance accounted for. In particular, we are working on extending the conceptualization of the skill vocabulary.

21.4.5 Utilizing CDMs to Validate the ECA

This section examines the extent to which cognitive diagnostic and test models can be used, in conjunction with large-scale summative tests that students have taken in the past, to diagnostically inform learning and assessment systems. We look at two components of these systems, namely, attribute and Q-matrix development and the CDM model selection process, as they pertain to a large scale mathematics assessment. This work serves as validation of the linear logistic test model used in the ECA. This project is still a work in progress, and here we describe the considerations made thus far.

As mentioned in the previous section, the Q-matrix, which maps the test items and skills, is an important component of CDM formulation because they determine the integrity of any subsequent actions. The Q-matrix development process for four forms of the large-scale mathematics assessment test started with the recognition that the test forms are aligned with the HF and CCSS. Thus, content experts defined attributes and developed Q-matrices that aligned with the HF and the standards. In

doing so, information provided to each student based on his/her performance on the test can be readily linked to available resources (i.e., the OpenEd platform), which makes the feedback directly actionable. To enhance the reliability and validity of the scores derived from the large-scale assessment test, the Q-matrices developed by experts were empirically validated; moreover, the most appropriate skills for each item was also selected. Below are the details of the Q-matrix development and validation, and model selection processes.

The Q-matrices for the math test were developed by content experts. Four test forms each with 60 items were considered in this study. Twenty-four attributes were defined across the three domains: ten skills for Operations, Algebra, & Functions (OAF); five skills for Geometry (G); and nine skills for Number (N). Due to a large number of attributes, each domain was analyzed separately. The domain-specific skills in a target domain were the focus of the analysis; irrelevant skills were collapsed into coarser nuisance attributes. The total numbers of target and nuisance attributes for OAF, G and N were 13, 8, and 12, respectively. The definitions of the target and nuisance attributes are given in Table 21.1.

The number of times that each attribute was measured across four forms and across three domains is shown in Fig. 21.4. Ignoring the nuisance attributes (i.e., Attributes 11–13 for Domain 1, Attribute 6–8 for Domain 2, and Attributes 10–12 for Domain 3), it can be observed that the number of times the target attributes were measured are quite disparate: 7–24 times for Attributes in Domain 1, 3–18 times for attributes in Domain 2, and 5–24 times for attributes in Domain 3. Cross-tabulations of the number of the target against the number of nuisance attributes measured are given in Tables 21.2, 21.3, and 21.4 for Domains 1–3, respectively. Most of the items measure one to three target attributes and zero or one nuisance attributes in Domain 1, one or two target attributes and one or two nuisance attributes in Domain 2, and one to three target attributes and zero or one nuisance attribute in Domain 3.

It can be gleaned from Tables 21.2, 21.3, and 21.4 that, for a particular target domain, not all 60 items were relevant in that some items solely measure nuisance attributes. The number of relevant items averaged across four test forms for each domain was approximately 54. By computing the row totals of Tables 21.2, 21.3, and 21.4, we can also examine the distribution of the number of target attributes required by each of the items. Across the three domains, the majority of the items measured one to three attributes; specifically, 91% of the items in Domain 1 and 95% of the items in Domain 3.

The G-DINA model (de la Torre, 2011; Ma & de la Torre, 2017) was fit to response data using a subset of $N = 5000$ examinees. A Q-matrix validation was then conducted using the data-driven approach based on the GDI and mesa plot. A mesa plot shows the PVAF for some possible q-vectors for a given item. It always starts with all-zero q-vector. The cutoff for a q-vector to be considered appropriate was set at PVAF = .85. The validation results given in Table 21.5 show that the attribute-wise agreement between the provisional and suggested Q-matrices across all test forms and domains was very high: the minimum was 93% and the average was 95%. After validating the Q-matrix using GDI, the attribute-wise agreements

Table 21.1 Attributes defined for three domains

Domain	Operations, Algebra, & Functions (OAF)	Geometry (G)	Number (N)
Target attributes	Addition & Subtraction, Multiplication & Division with Whole Numbers (ASMD.WN)	2-D and 3-D Figures & Their Properties – 2-Dimensional Figures and Their Properties; Planes, Points, Lines, and Angles; Circles; 3-Dimensional Figures (PF.2DFP, PF.PPLA, PF.C, 3DF)	Understanding Signed Numbers, Operations with Signed Numbers (N.SN, OAF.ASMD.SN)
	Addition & Subtraction, Multiplication & Division with Fractions and Decimals (ASMD.F,D)	Plane Figures – Coordinate Plane (PF.CP)	Understanding Real & Complex Numbers, Operations with Real & Complex Numbers (N.RCN, OAF.ORCN)
	Addition & Subtraction, Multiplication & Division with Signed Numbers (ASMD.SN)	Plane Figures – Perimeter, Area (PF.P, PF.A)	Addition & Subtraction, Multiplication & Division with Whole Numbers (ASMD.WN)
	Multistep Problem (MP)	Congruence, Similarity, and Transformations; Right Triangles; Trigonometry (CST, RT, T)	Addition & Subtraction, Multiplication & Division with Fractions and Decimals (ASMD.F,D)
	Ratio & Proportion (RP)	Other Geometry domain topics not already included in the attributes (Geometry other)	Multistep Problem (MP)
	Expressions (EX)		Expressions (EX)
	Function Concepts (FC)		Function Concepts (FC)
	Equations & Inequalities (EI)		Equations & Inequalities (EI)
	Operations with Real and Complex Numbers (ORCN)		Other Number domain topics not already included in the attributes (Number other)
	Other Operations, Algebra, & Functions domain topics not already included in the attributes (OAF other)		
Nuisance attributes	Number domain (Number)	Number domain (Number)	Other Operations, Algebra, & Functions domain topics not already included in the attributes (OAF other)
	Geometry domain (Geometry)	Operations, Algebra, & Functions domain (OAF)	Geometry domain (Geometry)
	Statistics & Probability domain (STAT PRB)	Statistics & Probability domain (STAT PRB)	Statistics & Probability domain (STAT PRB)

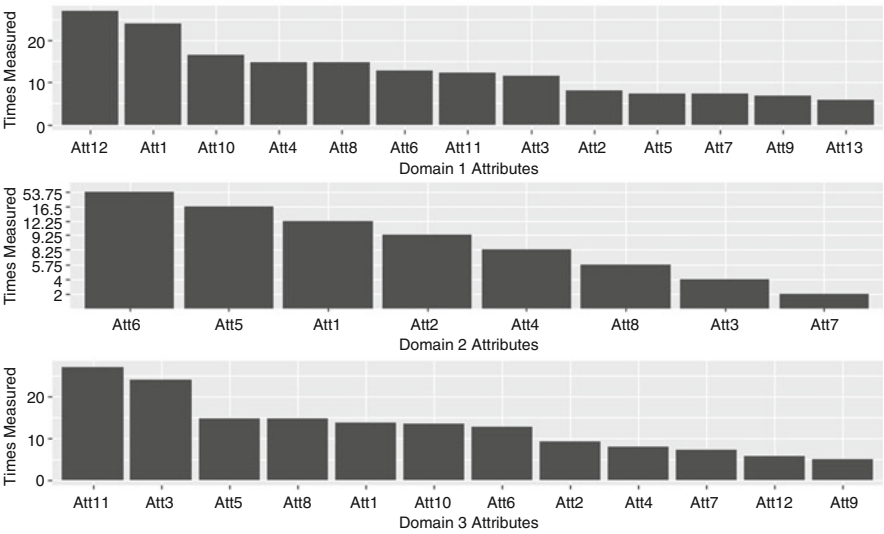


Fig. 21.4 Frequency that attributes are measured by items for three domains

Table 21.2 Cross table of number of target and nuisance attributes, domain 1

		# Nuisance attributes			
		None	One	Two	Three
# Target attributes	None	0	19	5	1
	One	14	43	4	0
	Two	20	41	6	0
	Three	32	30	6	0
	Four	5	7	5	0
	Five	1	0	1	0

Table 21.3 Cross table of number of target and nuisance attributes, domain 2

		# Nuisance attributes			
		None	One	Two	Three
# Target attributes	None	0	74	58	1
	One	10	50	14	0
	Two	6	18	6	0
	Three	1	1	1	0

Table 21.4 Cross table of number of target and nuisance attributes, domain 3

		# Nuisance attributes			
		None	One	Two	Three
# Target attributes	None	0	22	2	0
	One	14	40	9	0
	Two	33	49	11	1
	Three	30	17	2	0
	Four	5	4	1	0

Table 21.5 Percent attribute-wise agreement between provisional and suggested Q-matrices across domains and test forms

	OAF				G				N			
Form	1	2	3	4	1	2	3	4	1	2	3	4
Agreement	93	97	97	98	94	94	94	95	95	94	96	97

Table 21.6 Models selected by the Wald test for the three domains – Form A

	CDM					
Domain	G-DINA	DINA	DINO	LLM	rRUM	A-CDM
OAF	26	2	0	10	11	4
G	14	3	0	15	6	4
N	28	3	1	10	9	1

across all test forms and domains reach 95% in average, so the suggested Q-matrices are deemed to be reliable.

The Wald test was used in conducting item-level comparisons of the G-DINA model and a number of reduced CDMs, namely, the DINA model, DINO model, LLM, rRUM, and A-CDM, to find the optimal set of CDMs for a given test. The models selected for each domain are given in Table 21.6.

Comparing three domains (OAF, G, and N) on math test form A, as shown in Table 21.6, we found the G-DINA model was selected as the best model most frequently, especially in Domain 1: OAF and Domain 3: N. There are 10 target attributes and 3 nuisance attributes in Domain 1. Across four forms of tests, approximately 26 items favor the G-DINA model, 10 items favor LLM and 11 items favor rRUM. Note that, some reduced models were also selected frequently including LLM and rRUM. All of the above models relax the constraint of fixing the number of latent classes that an item could distinguish into only two latent classes to different extents. Finally, few items were suggested to be DINA, DINO or A-CDM. Similar to the results from Domain 1, three models: the G-DINA model, LLM, and rRUM are selected as the three best models for most of the items for Domain 2:G. The A-CDM and the DINA model are selected by fewer items. Again, for Domain 3: N, the G-DINA model with LLM and rRUM are selected as the best models most frequently, and the A-CDM and the DINA model are selected less frequently. Interestingly, we did find only one item that favored the DINO model in Form A. The DINO model also assumes a restricted number of latent classes. Besides it also assumes mastering one of the required attributes suffice to achieve the highest probability of success.

As shown in Table 21.5, results showed the suggested Q-matrices, and the provisional Q-matrices have approximately 95% element-wise agreement rate (EAR) and 76% vector-wise agreement rate (VAR) across all forms and options, which provides evidence to the construct validity of the Q-matrices developed by content experts.

21.5 Discussion

In this chapter, we discussed similarities and differences between the learning systems and assessments and described the traditional models that have been used by different research communities to identify the gaps in the students' knowledge. We emphasized the similarities of the Q-matrix used in CDMs and the skill/knowledge components from the BKT. We also discussed briefly the role of domain modeling (attributes, learning progressions, learning maps, knowledge components) and students' skill modeling. In doing so, we attempted to bring the psychometric and EDM/LAK literature together.

In our application, the ACTNext Educational Companion, we applied both classes of methodologies, the CDMs, and the EDM-inspired models. In this chapter, we reported some of the preliminary results from both approaches. The ML approach of matching taxonomies seems promising, mainly because it can handle multiple test forms and large datasets simultaneously. Nevertheless, more work is needed to validate the approach. The CDM works well, and we validated it; however, it remains an open question on how to create an automatic CDM that can be scaled-up across (parallel) test forms and very large data sets.

What we illustrated here is an example of computational psychometrics (CP), where traditional psychometric models are blended with machine learning algorithms (see von Davier, 2017). In future work, we will focus on validation and scalability of both methodologies. We will also aim for a better integration of the various approaches.

Acknowledgements The authors wish to thank Terry Ackerman, Former Lindquist Research Chair, ACT Inc., Yu Fang, Principal Psychometrician, Psychometric Research, ACT Inc. for providing insightful comments and feedback for this chapter. We sincerely acknowledge Melanie Rainbow-Harel, Former Assessment Designer and David Carmody Principal Assessment Specialist, ACT Inc. for their contribution towards design of attributes for three Math domains. We are thankful to Andrew Cantine- Communications and Publications Manager, ACTNext for editing this work. We are also thankful to ACT, Inc. for their ongoing support as this chapter took shape.

Glossary

A-CDM	The additive-CDM
ADA-MAT-BOM	Algebra-Matrices-Basic operations on matrices.
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BKT	Bayesian Knowledge Tracing
CCSS	Common Core State Standards
CDM	Cognitive Diagnostic Models

CP	Computational Psychometrics
DINA	The deterministic inputs, noisy "and" gate
DINO	The deterministic input, noisy "or" gate
DM	Data Mining
EAR	Element-wise agreement rate
ECA	Educational Companion App
EDM	Educational Data Mining
EM	Expectation Maximization
G	Geometry
G-DINA	The generalized DINA
GDI	G-DINA discrimination index
GDM	The general diagnostic model
HF	The ACT Holistic Framework
HMM	Hidden Markov Model
IRT	Item Response Theory
ITS	Intelligent Tutoring System
KC	Knowledge Component
KLI	The Knowledge-learning-instruction
LAK	Learning Analytics & Knowledge
LAS	Learning at Scale
LCDM	The log-linear CDM
LEAP	The Learning Analytics Platform
LLM	The log-linear model
LLTM	Linear Logistic Test Model
ML	Machine Learning
NGSS	Next Generation Science Standards
NRC	The National Research Council
OAF	Operations, Algebra, & Functions
OCW	Open Courseware
OER	Open Education Resources
OpenEd	Open educational resources (OER) from ACT's OpenEd that allows students to practice the skills they have yet to master.
PFA	Performance Factors Analysis
PVAF	The proportion of variance accounted for
Q-matrix	A Q-matrix is a mapping which identifies which skills or attributes are tested by an item or which skills or attributes are required to successfully complete an item on an assessment.
rRUM	The reduced reparametrized unified model
SRL	Self-regulated Learning
VAR	Vector-wise agreement rate
1PL	One-parameter Logistic Model

References

Abelson, H. (2008). The creation of OpenCourseWare at MIT. *Journal of Science Education and Technology, 17*(2), 164–174.

Atkins, D., Brown, J., & Hammond, A. (2007). *A review of the open educational resources (OER) movement: Achievements, challenges, and new opportunities*. San Francisco, CA: Creative Common, The William and Flora Hewlett Foundation.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... Jones, Z. (2016). mlr: Machine learning in R. *Journal of Machine Learning Research, 17*(170), 1–5. http://jmlr.org/papers/v17/15-066.html

Camara, W., O'Connor, R., Mattern, K., & Hanson, M. A. (2015). *Beyond academics: A holistic framework for enhancing education and workplace success*. Iowa City, IA: ACT, Inc.

Carmona, C., Millán, E., Pérez-de-la-Cruz, J., Trella, M., & Conejo, R. (2005). Introducing prerequisite relations in a multi-layered Bayesian student model. In *International conference on user modeling* (pp. 347–356). Berlin/Heidelberg, Germany: Springer.

Chopade, P., von Davier, A., Polyak, S., Peterschmidt, K., Yudelson, M., Greene, J., & Blum, A. (2017). Introducing the ACTNext educational companion: An intelligent, personalized guide for mobile learning. *Poster session presented at Educational Technology and Computational Psychometrics Symposium (ETCPS) organized by ACTNext, ACT Inc at The Englert Theatre*, November 15–16, 2017 (pp. 21). Iowa City, IA: ACTNext ACT Inc.

Conati, C., Gertner, A., & Vanlehn, K. (2002). Using Bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction, 12*(4), 371–417.

Corbett, A., & Anderson, J. (1995). Knowledge tracing: Modeling the acquisition of procedural know ledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Corcoran, T., Mosher, F., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Philadelphia, PA: Consortium for Policy Research in Education.

Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*(11), 671–684.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362.

de la Torre, J. (2009, March). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011, April). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355–373.

de la Torre, J., & Ma, W. (2016). Cognitive diagnosis modeling: A general framework approach and its implementation in R. In *A short course at the fourth conference on statistical methods in psychometrics*. New York, NY: Columbia University.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologa Educativa, 20*, 89–97.

Dibello, L., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*(4), 285–291.

Dibello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, volume 26, psychometrics* (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

Dijksman, J., & Khan, S. (2011). Khan academy: The world's free virtual school. In *APS meeting abstracts*.

Doignon, J., & Falmagne, J. (2012). *Knowledge spaces*. Berlin, Germany: Springer.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Gorin, J. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21–35.

Haberman, S., & von Davier, M. (2006). Some notes on models for cognitively based skills diagnosis. In C. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 1031–1038). Amsterdam, The Netherlands: Elsevier.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.

Hagenaars, J. (1993). *Loglinear models with latent variables*. Thousand Oaks, CA: Sage.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation, 15*(3), 1.

Journell, W., McFadyen, B., Miller, M., & Brown, K. (2014). K-12 online education: Issues and future research directions. In *Handbook of research on emerging priorities and trends in distance education: Communication, pedagogy, and technology* (p. 385). Hershey, PA: Information Science Reference.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Käser, T., Klingler, S., Schwing, A., & Gross, M. (2014). Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *International conference on intelligent tutoring systems* (pp. 188–198). Cham, Switzerland: Springer.

Koedinger, K., Corbett, A., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798.

Leighton, J. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6–15.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK/New York, NY: Cambridge University Press.

Levinson, S., Rabiner, L., & Sondhi, M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *Bell System Technical Journal, 62*(4), 1035–1074.

Ma, E., & de la Torre, J. (2017). *The generalized DINA model framework, package 'GDINA'*. Retrieved February 12.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*, 200–217.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Millán, E., Loboda, T., & Pérez-de-la-Cruz, J. (2010). Bayesian networks for student model engineering. *Computers & Education, 55*(4), 1663–1683.

No Child Left Behind. (2002). Act of 2001 Pubic Law No. 107–110, § 115. Stat, 1425.

NRC. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Research Council/National Academies Press.

NRC. (2005). In M. Wilson & M. Bertenthal (Eds.), *Systems for state science assessments. Committee on test design for K-12 science achievement*. Washington, DC: National Research Council, National Academies Press.

NRC. (2007). In R. Duschl, H. Schweingruber, & A. Shouse (Eds.), *Taking science to school: Learning and teaching science in grades K-8. Committee on science learning, kindergarten through eighth grade*. Washington, DC: National Research Council, National Academies Press.
OpenEd. (n.d.). *Driving blended learning from classroom assessments*. https://www.opened.com/
Open Learning Initiative. (2018). Retrieved from https://oli.cmu.edu/learn-more-about-oli/
OpenStax. (2018). Retrieved from https://openstax.org/about
Palmisano, S. (2008). *A smarter planet: The next leadership agenda*. New York, NY: IBM.
Pelánek, R. (2017, December). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*(3–5), 313–350.
Pellegrino, J., Baxter, G., & Glaser, R. (1999). Addressing the "Two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *In Review of Research in Education, 24*, 307–353.
R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* (Studies in mathematical psychology, Vol. 1). Copenhagen, Denmark: Danmarks Paedagogiske Institut.
Rudd, J., Davia, C., & Sullivan, P. (2009). *Education for a smarter planet: The future of learning*. Redbooks IBM.
Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*(1), 78–96. https://doi.org/10.1177/0013164407301545
Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55–73.
Templin, J. (2016). *Diagnostic measurement: Theory, methods, applications, and software*. NCME training session, Washington, DC. Retrieved April 8.
Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.
Tsoumakas, G., & Katakis, I. (2006). *Multi-label classification: An overview*. Thessaloniki, Greece: Deptartment of Informatics, Aristotle University of Thessaloniki.
von Davier, A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement, 54*, 3–11.
von Davier, A., Polyak, S., Peterschmidt, K., Chopade, P., Yudelson, M., de la Torre, J., & Paek, P. (2017, November). *Systems and methods for interactive dynamic learning diagnostics and feedback*. U.S. Patent Application No. 15/802,404.
Von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS research report RR-05-16). Educational Testing Service. https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2005.tb01993.x
von Davier, M., & Haberman, S. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'Diagnostic' classification models—A commentary. *Psychometrika*, 340–346.
Wilmot, D., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning, 13*(4), 259–291.
Wilson, M. (2005). *Constructing measures: An item response theory approach*. Mahwah, NJ: Lawrence Erlbaum.
Zhang, S., & Chang, H.-H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning*, 67–92.
Zhang, S. S. (2014). *Statistical inference and experimental design for Q-matrix based cognitive diagnosis models*. Doctoral dissertation. Columbia University.
Zhu, Z., & Shen, D. (2013). Learning analytics: The scientific engine for smart education. *E-Education Research, 241*(1), 5–12.

Chapter 22
CDMs in Vocational Education: Assessment and Usage of Diagnostic Problem-Solving Strategies in Car Mechatronics

Stephan Abele and Matthias von Davier

Abstract The aim of this chapter is to use psychometric models including DCMs to assess diagnostic problem-solving strategies and to investigate the usage of these strategies in car mechatronics. The present study not only advances research on the strategies' assessment, but also informs professional and vocational education. From the educational perspective, it is not only important to know how to assess diagnostic problem-solving strategies but also to gather information about the strategies' usage. Such knowledge helps teaching when and under which conditions the strategies are applicable.

22.1 Introduction

The aim of this chapter is to use psychometric models including DCMs to assess diagnostic problem-solving strategies and to investigate the usage of these strategies in car mechatronics. The present study not only advances research on the strategies' assessment, but also informs professional and vocational education. From the educational perspective, it is not only important to know how to assess diagnostic problem-solving strategies but also to gather information about the strategies' usage. Such knowledge helps teaching when and under which conditions the strategies are applicable.

S. Abele (✉)
Institute of Vocational Education and Vocational Didactics, Technische Universität Dresden, Dresden, Germany
e-mail: stephan.abele@tu-dresden.de

M. von Davier
National Board of Medical Examiners (NBME), Philadelphia, PA, USA
e-mail: mvondavier@nbme.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_22

Diagnostic problem-solving strategies are employed to solve diagnosis problems. Diagnosis problems refer to situations in which causes of undesired states (e.g., diseases or machine defects) must be identified. Solving diagnosis problems is relevant in many professional and vocational domains: Faced with a syndrome of symptoms physicians must find the underlying cause or disease; teachers must find reasons for observed learning difficulties to facilitate learning progress; technicians must find causes of malfunctioning machines, and so forth. Most of the research on diagnostic problem solving has been conducted in medical (e.g., Croskerry, 2009; Elstein, Shulman, & Sprafka, 1990; Norman, 2005) and technical domains (e.g., mechanics or electronics; Hoc & Amalberti, 1995; Rasmussen, 1993; Rouse, 1983). In these domains, researchers have often used the terms "clinical reasoning" (Kassirer, Wong, & Kopelman, 2010) or "troubleshooting" (Perez, 2012) instead of "diagnostic problem solving" and also consider treatment options (e.g., repair in technical domains). The emphasis, however, typically is on diagnostic problem solving, that is on finding the cause(s) of undesired states (Jonassen, 2011, p. 78). This study investigates diagnostic problem solving in a technical domain, more specifically, in the domain of car mechatronics, but also draws on research from other professional and vocational domains where diagnostic problem solving is involved. Car mechatronics is a multidisciplinary field integrating mechanical and electrical technologies as well as computer and information systems relevant to troubleshooting and other tasks of auto technicians (Baethge & Arends, 2009).

Diagnostic problem-solving strategies related to car mechatronics are taught in vocational education settings. Vocational education has traditionally played an important role in many European countries and in Australia (Billett, 2011), due to its social impact (e.g., on the national employment rate) and its economic significance, to name but a few reasons. In the USA, vocational education has recently started to gain attention of policy-makers and the public (Lerman, 2016). The relevance of vocational education is also documented by the recent study of Hanushek, Schwerdt, Woessmann, and Zhang (2017), who compared the vocational education system to the general education system, using an interesting yet limited sample of countries and focusing on a specific aspect of educational systems: their labor-market outcomes.

With respect to the assessment of diagnostic problem-solving strategies, the available literature clearly shows some challenges and research desiderata: There is no or at least little consensus on how to operationalize diagnostic problem-solving strategies (Konradt, 1995; Schaper, Hochholdinger, & Sonntag, 2004) and there is little research on how to distinguish between different types of diagnostic problem-solving strategies based on their empirical correlates (i.e., observable diagnostic problem-solving behavior). As will be discussed later, diagnostic problem-solving strategies refer to different approaches that regulate the diagnostic problem-solving process. Consequently, any assessment of these strategies requires diagnosticians' process data before the advent of computer-based assessments commonly collected with think-aloud protocols or interviews (Konradt, 1995). These procedures are time-consuming, which often leads to small sample studies and hence findings that are difficult to generalize. In this assessment context, computer-generated

log-file data could be particularly beneficial: Such process data can be gathered automatically with computer-based assessments and can subsequently be analyzed applying diagnostic classification models.

In the following, we clarify the term "diagnostic problem-solving strategy", delineate the assessment framework and conceptualize the diagnostic problem-solving strategies under investigation. We then discuss the usage of the strategies and introduce research hypotheses regarding the strategies' assessment and usage. Afterwards, we present the results of our empirical study. Finally, we evaluate the research hypotheses relative to our main findings and discuss implications as well as limitations of our study. To test the research hypotheses, a sample of car mechatronic apprentices was drawn from vocational schools and each member was confronted with diagnosis problems of car mechatronics in an authentic (i.e., highly realistic) computer-based assessment.

22.2 Assessment of the Diagnostic Problem-Solving Strategies

22.2.1 Focus: Knowledge-Based Diagnostic Strategies

The focus here is on diagnostic problem-solving strategies that require domain-specific knowledge. Van Merriënboer (2013) differentiates between weak and knowledge-based methods to solve problems. Weak methods such as the trial-and-error strategy, the means-ends-analysis strategy or the heuristic search strategy can be applied to unfamiliar types of problems and domains as they do not rely on domain-specific knowledge. In contrast, knowledge-based strategies are closely related to domain-specific knowledge. Research suggests that diagnostic problem-solving outcomes depend on many factors in vocational contexts (Abele, 2014). With respect to individual characteristics, domain-specific knowledge, however, has turned out to be especially important for successful diagnostic problem solving (Boshuizen & Schmidt, 2008), although general mental abilities (e.g., intelligence and metacognition, Jonassen, 2011, p. 78 ff.), self-regulation, motivation, interest (Rausch, Seifried, Wuttke, Kögler, & Brandt, 2016) and emotion (Sembill, Rausch, & Kögler, 2013) appear to be relevant as well.

Knowledge-based diagnostic strategies imply several problem-solving activities. Imagine a situation in which a diagnostician observes that the car's "check engine" light is on. He hypothesizes (makes an "assumption") that a defective fuel temperature sensor is responsible for this symptom, knowing this is one of many possible causes of the light being on. He decides to test the sensor and obtains an abnormal test result. Drawing on this result, he concludes that the fuel temperature sensor is broken and that the broken sensor causes the symptom. This simple example shows that the solution of a diagnosis problem is in general the result of a sequence of several activities. Some of these activities are overt and directly observable (e.g., the sensor test) and are called diagnostic problem-solving behavior;

some of these activities are not directly observable and are called mental problem-solving activities. Individuals use mental problem-solving activities and observed problem-solving behaviors to manage the diagnostic problem-solving process.

The diagnostic problem-solving process starts with the perception of the diagnosis problem and ends with individuals offering either the (correct or incorrect) problem's solution or giving up. Abele (2017) presents a theory of the diagnostic problem-solving process and distinguishes between the following sub-processes: (1) representing information, (2) formulating diagnostic hypotheses, (3) testing diagnostic hypotheses, and (4) evaluating diagnostic hypotheses.

(1) To begin with, diagnosticians mentally represent problem-related information (e.g., an active "check engine" light). (2) Using this information, they formulate diagnostic hypotheses. A diagnostic hypothesis is an assumption of a potential but untested cause of the undesired state (e.g., a defective fuel temperature sensor might cause the "check engine" light to come on). (3) In order to test diagnostic hypotheses, diagnosticians collect relevant evidence (e.g., by testing the fuel temperature sensor). (4) Afterwards, the diagnostic hypotheses can be evaluated using the evidence from the tests. In case of confirming evidence, the hypothesis can be accepted, and the diagnostician can specify the problem's solution. If evidence refutes the hypothesis, problem solvers must develop and test additional diagnostic hypotheses and so forth. Real diagnostic problem-solving processes can deviate from this ideal chronological sequence, but it seems reasonable to suppose that the four sub-processes reflect main requirements of diagnostic problem solving in vocational and professional domains (Abele, 2017).

Diagnosticians use different knowledge-based diagnostic strategies to solve diagnosis problems (Konradt, 1995). The following knowledge-based diagnostic strategies are considered here and discussed in the next sections: the computer-based strategy, the case-based strategy, and the mental-model-based strategy (in short: model-based strategy). The common ground of these strategies is that individuals apply them to regulate the diagnostic-problem solving process. That is, diagnosticians apply knowledge-based diagnostic strategies to represent problem-related information as well as to formulate, test, and evaluate diagnostic hypotheses. A difference between the knowledge-based diagnostic strategies is that they are associated with different mental problem-solving activities and observable diagnostic problem-solving behavior. These differences facilitate the assessment and empirical distinction of the strategies.

22.2.2 Assessment Framework

This paper provides a first step toward a theory-based framework to assess diagnostic problem-solving strategies using psychometric models. The assessment framework comprises two steps: First, idiosyncratic patterns of the diagnostic problem-solving strategies are defined theoretically. Second, the patterns are used as a search template to scan the log-file data and to find out if and which

individuals showed the respective patterns (i.e., strategies). A central assumption of this framework is that knowledge-based diagnostic strategies can be inferred from patterns of observable problem-solving behavior represented in computer-generated log files. Such log files usually include fine grained records over time of interactions between diagnosticians and the computer-based assessment (Fig. 22.1). Computer-generated log-files can contain many hundreds or more records of human-computer interactions that can be linked to diagnostic problem-solving behavior per individual and diagnosis problem (Abele, 2017). Research on the analysis of computer-generated log files has recently made significant progress (Goldhammer, Kröhne, Keßel, Senkbeil, & Ihme, 2014; Greiff, Wüstenberg, & Avvisati, 2015; He & von Davier, 2015, 2016) but has also clearly shown that log-file analyses benefit from focusing on a selection of log-file entries. It is an open question which entries should be included.

The following problem-solving behavior is included here: critical test behavior and critical information behavior. These behaviors turned out to correlate with the problem-solving success and thus are called "critical" (Abele, 2017). To identify the critical test behavior and critical information behavior, the following three-step procedure must be executed: (1) identification of critical diagnostic hypotheses; (2)

```
<?xml version="1.0"?>
- <activityLog>
    <activity lightState="00" motorState="000" timeStamp="12:15:19" id="1">lesson loaded: Fehler 6</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:00:02" id="2">menu open: SystemSelection</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:01:52" id="3">menu open: Document</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:01:52" id="4">Arbeitsauftrag page: 1</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:02:29" id="5">button close: Arbeitsauftrag</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:02:39" id="6">system loaded: Motorraum</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:02:42" id="7">system loaded: Cockpit</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:02:44" id="8">ignition on: true</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:02:45" id="9">shortCut open: DiagnoseSoftware</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:02:45" id="10">ESITRONIC screen: 00_01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:02:47" id="11">ESITRONIC screen: 00_02</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:02:56" id="12">ESITRONIC screen: g4_00_01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:02:58" id="13">ESITRONIC screen: g4_03_01_01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:03:03" id="14">ESITRONIC screen: g4_03_01_13_01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:03:10" id="15">ESITRONIC screen: g4_03_01_13_m01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:03:13" id="16">ESITRONIC screen: g4_03_01_13_m01_m08_m02_04_01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:03:18" id="17">ESITRONIC Fehlerspeicher lesen:</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:03:22" id="18">ESITRONIC screen: g4_03_01_13_m01_m08_m02_04_02</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:03:57" id="19">ESITRONIC screen: g4_44A3_01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:12" id="20">button close: null</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:21" id="21">module open: Motorabdeckung</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:23" id="22">module zoom in: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:25" id="23">module open: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:27" id="24">module close: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:29" id="25">module open: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:34" id="26">shortCut open: Multimeter</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:42" id="27">Messspitze aufgesetzt: red/01-04-03-01</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:45" id="28">Messspitze aufgesetzt: black/01-04-03-02</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:04:45" id="29">Multimeter measure: 01-04-03-01/01-04-03-02 (5 Udc)</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:05:20" id="30">button close: Multimeter</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:05:22" id="31">shortCut open: DiagnoseSoftware</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:05:27" id="32">ESITRONIC screen: g4_44A3_02</activity>
    <activity lightState="00" motorState="100" timeStamp="+00:05:38" id="33">button close: null</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:05:41" id="34">ignition on: false</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:05:45" id="35">shortCut open: Multimeter</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:05:54" id="36">Messspitze aufgesetzt: black/01-04-03-04</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:05:59" id="37">Messspitze aufgesetzt: red/01-04-03-03</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:05:59" id="38">Multimeter measure: 01-04-03-03/01-04-03-04 (OL R)</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:39" id="39">button close: Multimeter</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:40" id="40">module close: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:42" id="41">module open: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:43" id="42">module close: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:44" id="43">module zoom out: Kraftstofftemperatursensor</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:46" id="44">button close: null</activity>
    <activity lightState="00" motorState="000" timeStamp="+00:06:49" id="45">lesson FINISHED</activity>
  </activityLog>
```

Fig. 22.1 Example of a log file showing an individual's problem-solving behavior exhibited during diagnostic problem solving in the domain of car mechatronics; each line stands for a specific observed problem-solving behavior

using these hypotheses to identify critical information and critical tests; (3) using the critical information and critical tests to identify the critical information behavior and the critical test behavior.

(1) Critical diagnostic hypotheses provide potential causes of undesired observed states that makes sense from a substantive viewpoint and in terms of a specific diagnosis problem. For example, regarding the "check engine" light coming on, it makes sense to hypothesize a *broken* fuel temperature sensor causes the defect. In contrast, it does not make sense to suppose an empty fuel tank is responsible for this symptom. Whereas the first assumption represents a critical diagnostic hypothesis, the second hypothesis does not, as an empty tank is not a plausible explanation of the symptom. We identified the critical diagnostic hypotheses by applying problem-specific expertise to the symptoms (e.g., the check engine light coming on) given in the problem description. A symptom commonly allows for generating (many) critical diagnostic hypotheses. For example, broken sensors, broken cables, and so forth can cause the "check engine" light. Note that each critical diagnostic hypothesis gives a reasonable explanation for the symptoms but only one of these hypotheses relates to the "true" cause, assuming a single cause is responsible for the symptom. (2) Drawing on the critical diagnostic hypotheses, the critical information and the critical tests can be identified. For example, to generate critical diagnostic hypotheses, problem solvers must know (i.e., mentally represent) symptoms of a diagnosis problem and they must reflect on how to test critical diagnostic hypotheses. (3) The critical information behavior and the critical test behavior follow from the critical mental problem-solving activities. For instance, to represent symptoms mentally, problem solvers must retrieve the problem description; to test the critical diagnostic hypotheses, they usually conduct specific tests.

22.3 Knowledge-Based Diagnostic Strategies: Conceptualization and Behavioral Patterns

We consider differences in how critical diagnostic hypotheses are formulated as an important reason for different behavioral patterns relating to three major strategies: With the computer-based strategy, the critical diagnostic hypotheses come from a computer-based expert system, that is, an external source. With the case-based strategy, the critical diagnostic hypotheses originate from previous experiences (cases), that is, from long-term memory. Diagnosticians using the *model-based strategy* build mental models that represent details of the problem-related system (e.g., details of a car's motor) and use these models to derive critical diagnostic hypotheses systematically.

22.3.1 Computer-Based Strategy

When diagnosticians follow the instructions given in a computer-based expert system to solve a diagnosis problem they apply the computer-based strategy. Computer-based expert systems are computer programs that aid individuals to solve (diagnosis) problems. The computer-based strategy requires domain-specific knowledge of how to handle a computer-based expert system and how to execute the system generated instructions. This strategy resembles a specific type of instruction-based strategy, that is, an approach to solve problems following the instructions provided by an external source (a computer-based expert system, a user guide, a human mentor, etc.).

With the computer-based strategy, the critical diagnostic hypotheses are provided by an expert system. Such systems do not necessarily give the critical diagnostic hypotheses explicitly; sometimes they supply instructions that are implicitly connected to critical diagnostic hypotheses only. For example, if a computer-based expert system suggests testing the fuel temperature sensor, it is assumed implicitly that a broken fuel temperature sensor might cause the car's symptom of the check engine light. Computer-based expert systems frequently cover the most likely or most relevant hypotheses but not necessarily all the possible critical diagnostic hypotheses of a diagnosis problem in car mechatronics.

The specific behavioral pattern emerging from the computer-based strategy can be identified by executing the steps described above and focusing on the critical diagnostic hypotheses offered by the computer-based expert system. Diagnosticians employing the "pure" computer-based strategy exclusively show the critical information behavior and critical test behavior related to the critical diagnostic hypotheses suggested by the computer and do not show other critical problem-solving behavior. More specifically, they do *not* exhibit critical information behavior to build a mental model, as they do not formulate critical diagnostic hypotheses independently from the computer-based expert system. Finally, diagnosticians do *not* show critical test behavior related to critical diagnostic hypotheses *not* provided by the computer-based expert system.

22.3.2 Case-Based Strategy

The case-based strategy activates knowledge about a previous case to solve a current diagnosis problem. For example, if a broken fuel temperature sensor was responsible for a "check engine" light in the past, the broken sensor is hypothesized to be responsible for the same symptom in the current situation as well. A previous case reflects a diagnosis problem that was mastered in the past and induces the activation of specific knowledge due to similarities with the current diagnosis problem. The similarity between the previous case and the current diagnosis problem activates

the case and the related piece of knowledge. The symptom (e.g., a check engine light) commonly represents the similarity (i.e., link) between the current diagnosis problem and the previous case.

The label "case-based strategy" was inspired by the study of Jonassen and Hernandez-Serrano (2002, p. 65) who introduced the term "case-based reasoning" to describe situations in which problems "are solved by retrieving similar past experiences". The case-based strategy largely corresponds to the symptomatic strategy identified by Rasmussen (1993) and utilized by many other authors in technical domains (Konradt, 1995; Schaper et al., 2004). The symptomatic strategy starts with exploring and mentally representing symptoms of a diagnosis problem. Then, the mental representation of the symptoms "is used as a search template to find a matching set in a library of symptoms" (Rasmussen, 1993, p. 987); the library of symptoms is stored in long-term memory and connected to other problem-related information.

The dual-process theory (Croskerry, 2009; Schwartz & Elstein, 2008, p. 229) and the study of Norman, Young, and Brooks (2007) suggest that the case-based strategy can be classified as a non-analytical strategy. Non-analytical strategies imply automatic, fast, effortless and, compared to analytical strategies, fewer mental problem-solving activities and problem-solving behaviors. From the perspective of cognitive psychology, the case-based strategy resembles analogical reasoning (van Merriënboer, 2013). Taking the perspective of research on expertise, the case-based strategy draws upon pattern recognition (Norman et al., 2007): If a well-known symptom pattern is recognized, problem solvers retrieve and apply corresponding pieces of knowledge. Boshuizen and Schmidt (2008) termed such pieces of knowledge "illness scripts". An illness script develops in dependence of diagnosing a disease (i.e., a diagnosis problem) and encodes information on symptoms, the cause and other characteristics of the disease as well as a diagnosis procedure (p. 115). An illness script is activated automatically as a whole, meaning that all components are immediately available and applicable when diagnosticians perceive relevant symptoms (p. 115). Semantically, it seems appropriate to replace the term "illness script" by "(car) defect script" in the domain of car mechatronics. Following the previous argumentation, a defect script encodes information on the symptoms of a car's defect, its cause as well as on how to test for and evaluate the cause.

In terms of the case-based strategy, the critical diagnostic hypotheses are available through the defect scripts. Please note that defect scripts link symptoms to causes, that is, (implicitly) provide critical diagnostic hypotheses. To assess the case-based strategy, we introduce an auxiliary assumption: A diagnostician's number of case scripts (i.e., critical diagnostic hypotheses) depends on his diagnostic experience. This study was based on a sample of car mechatronic apprentices nearing the end of their three-year formal training. We assume that these apprentices had only the "very common" case scripts available. In view of a lighting system problem, such scripts relate to a blown fuse or a broken lightbulb.

To identify idiosyncratic pattern of critical information behavior and critical test behavior found when the case-based strategy is applied, the common case scripts

for a specific symptom must be determined a priori. In any case, the "pure" case-based strategy does *not* imply critical information behavior exhibited to retrieve critical diagnostic hypotheses from the computer-based expert system. Furthermore, this strategy is *not* associated with critical information behavior necessary to build mental models.

22.3.3 Model-Based Strategy

Diagnosticians applying the model-based strategy use mental models to solve diagnosis problems. Mental models facilitate to mentally derive critical diagnostic hypotheses systematically. In case of the "check engine" light, problem solvers mentally model relevant parts of the motor, including several components of the system (fuel temperature sensor, engine control unit, cables, etc.) and the dependencies between these components. Based on the mental representation, they specify causes of the "check engine" light in a first step and test them subsequently. For example, they might assume a broken fuel sensor causes the "check engine" light and then test that sensor.

The model-based strategy shares features with the hypothethico-deductive strategy, which is a very prominent concept in medical education (Schwartz & Elstein, 2008): Both strategies stress the role of formulating and testing diagnostic hypotheses. There is, however, an important difference: Whereas the formulation of critical diagnostic hypotheses depending on mental models is a key component of the model-based strategy, the hypothethico-deductive strategy ignores where diagnostic hypotheses come from (Norman et al., 2007). This aspect is covered by the scheme-inductive strategy (Coderre, Mandin, Harasym, & Fick, 2003). When the scheme-inductive strategy is applied, knowledge structures (schemes) are activated. Such schemes can include different causes for certain symptoms (i.e., they imply different critical diagnostic hypotheses, Coderre et al., 2003, p. 703). A scheme is retrieved from memory and does not give information about problem-related systems. The scheme-inductive strategy implies searching through schemes to develop critical diagnostic hypotheses. In contrast, with the mental-based strategy, the critical diagnostic hypotheses are derived from mental models depicting (parts) of the problem-related system. Such mental models typically integrate internal information (e.g., system knowledge) and external information (Perez, 2012). In technical domains, system knowledge encodes topographical (locations of components), structural (functional relations between components) and functional (purpose of components) information about technical systems and their components (Kluwe & Haider, 1990). Relevant external information comes from interactions with the problem environment, that is, critical information behavior and may refer to retrieving wiring diagrams, for example. The topographic strategy also rests upon mental models (Rasmussen, 1981). In case of the topographic strategy, the models, however, focus on the location of components (their *topos*) and neglect the functional relations between the components. Thus, topographical

models usually enable the formulation of only some but not all critical diagnostic hypotheses. Considering the dual-process theory, the model-based strategy mirrors a systematic, analytical, slow and effortful approach to solve diagnosis problems (van Merriënboer, 2013).

The model-based strategy is associated with a complete collection of critical diagnostic hypotheses of a diagnosis problem in principle: An adequate mental model allows generating all critical diagnostic hypotheses related to a specific symptom or set of symptoms. Consequently, a more complete selection of critical information behavior and critical test behavior is to be expected when the model-based strategy is employed: Unlike the computer-based strategy and the case-based strategy, the model-based strategy includes critical information behavior related to build mental models (e.g., retrieving the circuit diagram of the fuel temperature sensor). The model-based strategy is *not* associated with the instructions given in the computer-based expert system.

22.4 Usage of Knowledge-Based Diagnostic Strategies

We assume that the three knowledge-based diagnostic strategies being considered differ in terms of their individual costs and that the strategies' probability of usage depends in part on their individual cost. The individual costs of the model-based strategy are particularly high, as this strategy requires both the availability and the adequate use of deep diagnostic knowledge, that is, system knowledge. Moreover, this analytical strategy is associated with a relatively high degree of mental effort and time for information processing. In comparison to the model-based strategy, the individual costs of the computer-based strategy and the case-based strategy are lower. When applying these two strategies, problem solvers do not need deep knowledge as they do not intend to model and understand problem-related systems. Furthermore, the mental effort for and the time spent on information processing is relatively low as the computer-based expert system and the defect scripts directly provide the critical diagnostic hypotheses and a template to solve the diagnosis problem.

Following Schwartz and Elstein (2008, p. 226), we assume that the probability of usage of the strategies depends on the difficulty of diagnosis problems. The difficulty of diagnosis problems is closely correlated with the familiarity of these problems in the context of car mechatronics (Nickolaus, Abele, Gschwendtner, Nitzschke, & Greiff, 2012), where "familiarity" indicates a (very) high frequency of exposure to a diagnosis problem in practice/real-life. This implies that the case-based strategy should have a high probability of usage with familiar and easy diagnosis problems. A diagnosis problem can be familiar in view of its symptoms and its cause(s). In case of familiar symptoms and *un*familiar causes (i.e., difficult problems), a shift from the case-based strategy to another strategy seems plausible, as the case-based strategy is not successful.

Nickolaus et al. (2012) also found that a computer-based expert system is helpful to solve easy but not difficult diagnosis problems in car mechatronics, if available and applicable. One plausible explanation for this is that computer-based expert systems are especially well suited to provide familiar critical diagnostic hypotheses but less so for unfamiliar ones. What happens if the computer-based strategy is applied to a diagnosis problem and the computer-based expert system does not provide the correct cause of the problem? In that case, the computer-based strategy is not successful and a change in strategy is needed. In line with this argumentation, the probability of usage of the computer-based strategy should be higher for easy than for difficult diagnosis problems.

It is possible that diagnosticians switch from the case-based strategy to the computer-based strategy or vice versa. These changes in strategy, however, are probably not particularly helpful to solve difficult diagnosis problems: As mentioned before, neither the case-based strategy nor the computer-based strategy seem to have a great potential to solve difficult diagnosis problems. If both strategies are not successful, some diagnosticians might give up or stop due to a lack of competence and some diagnosticians might switch to the model-based strategy.

To sum up, the usage probability of the computer-based strategy and the case-based strategy should generally be higher than the usage probability of the model-based strategy because of differences in costs associated with each. The usage probability of the computer-based strategy and the case-based strategy is expected to be higher when faced with easy diagnosis problems than with difficult problems due to a difficulty-induced change to the model-based strategy. At the same time, it can be expected that the usage probability of the model-based strategy is higher for difficult than for easy diagnosis problems.

22.5 Research Hypotheses

Given the reasoning with regard to the three types of distinct strategies developed in the previous section, we investigated the following hypotheses in the context of car mechatronics:

H1: Diagnosticians use the computer-based strategy, the case-based strategy and the model-based strategy to solve diagnosis problems. This hypothesis was examined by applying the three-step procedure described above to identify and score individuals' critical problem-solving behavior as extracted from computer-generated log-files. To identify behavioral patterns, a range of modeling approaches, including diagnostic classification models, item response theory, and latent class analysis, was applied. Empirically, we expected to find patterns of critical information behavior and critical test behavior that can be related to the likelihood of the three knowledge-based diagnostic strategies.

H2: Diagnosticians more often employ computer-based and case-based strategies than the model-based strategy independent of the difficulty of the diagnostic

problem. Empirically, we expected to observe fewer diagnosticians employing the model-based strategy than the other strategies *both* for easy and difficult diagnosis problems.

H3: Diagnosticians apply the model-based strategy more often when facing difficult than when facing easy diagnosis problems. Part of the theoretical foundation of this hypothesis is that (some) diagnosticians may switch strategies from computer-based or case-based to the model-based strategy when they attempt to solve difficult rather than easy diagnosis problems. As the choice of which knowledge-based diagnostic strategy to apply probably does not depend on a problem's difficulty alone but also on other problem characteristics, we investigated this hypothesis using two pairs of diagnosis problems. Each pair consisted of very similar diagnosis problems but varied in difficulty. Empirically, we expected that more diagnosticians would apply the model-based strategy in the difficult problem than in the easy diagnosis problem.

22.6 Method

22.6.1 Sample and Design

To test the research hypotheses, 369 car mechatronic apprentices[1] nearing the end of the third year of formal training were sampled from three German federal states (Baden-Württemberg, Bavaria and Hesse) and 25 classes of vocational schools. The sample size varied slightly in the statistical analyses, mainly due to apprentices that did not complete the full set of problem-solving assessments (from $N = 369$ to $N = 336$). In this line of training for a job in car mechatronics, most apprentices are male in Germany, and consequently almost all the participants were male (96.6%); the apprentices were 20.8 years old on average and their age ranged from 17 to 41 years.

Two pairs of simulated problem scenarios were administered: One pair covered diagnosis problems concerning the fuel temperature sensor (sensor problems), the other two diagnosis problems related to the lighting system (lighting problems). Each pair represented diagnosis problems that were very similar in terms of their symptoms but differed in terms of the symptoms' causes and their difficulty (Abele, Walker, & Nickolaus, 2014, p. 174; Gschwendtner, Abele, & Nickolaus, 2009, p. 573). Within the pairs there was one rather easy and one rather difficult problem. The diagnosis problems were given in an authentic computer simulation, using computer

[1]The occupational field of car mechatronics covers, among other things, troubleshooting, repair and maintenance of cars (Baethge & Arends, 2009, p. 33–47). In Germany, car mechatronic apprentices usually attend a 3.5 years training programme including a school-based and workplace-based training ("dual apprenticeship system"). The training of car mechatronic technicians differs significantly from one country to the next (Baethge & Arends, 2009, p. 34).

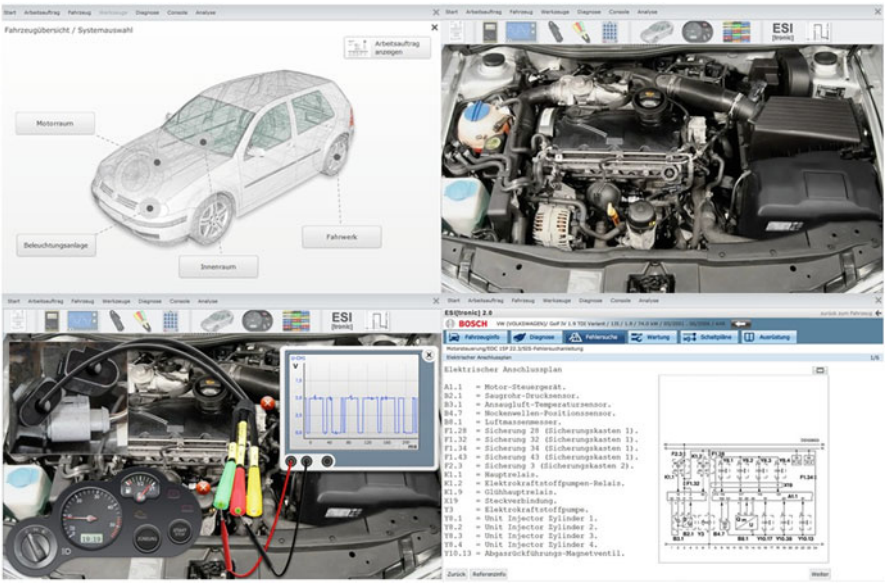


Fig. 22.2 Screenshots of the computer simulation in German (top left: start page giving an overview of the car systems; top right: the upper part shows the icons of the toolbox, below the motor compartment referring to the system "electronic engine management" is shown; bottom left: measurement of a signal using the oscilloscope, cockpit and adapter; bottom right: circuit diagram retrieved from the computer-based expert system)

labs of the vocational schools. The total time provided for the testing session was 85 min (20 min for the easy sensor problem and 25 min for the difficult one, 15 min for the easy lighting problem and 25 min for the difficult one). To control for position effects, problem presentation was rotated in a Latin square design (Frey, Hartig, & Rupp, 2009, p. 45).

22.6.2 Assessment of the Knowledge-Based Diagnostic Strategies

22.6.2.1 Computer-Based Test: An Authentic Computer Simulation for Car Mechatronics

The computer-based assessment of diagnostic problem solving uses authentic graphic material (pictures, screenshots, etc.) and represents the following parts of the work environment of car mechatronics: (1) a selection of car systems, (2) a toolbox and (3) a computer-based expert system (Fig. 22.2).

(1) The simulation covers four systems of a VW Golf, which were identified to be of high practical relevance by experienced car mechatronic technicians, teachers/trainers of car mechatronic apprentices and academic experts (Baethge &

Arends, 2009, p. 16). Here, the "lighting system" and the system "electronic engine management" are relevant. In the "electronic engine system", for example, 17 components (plugs of actuators and sensors, the battery, etc.) are accessible. (2) The toolbox is available in both systems and contains icons representing different work equipment (e.g., problem description, multimeter, fuse box, computer-based expert system). (3) Computer-based expert systems are an integral part of a car mechatronic technicians' work environment. The simulation covers relevant segments of the ESI[tronic] from Bosch, which is an internationally widespread system and is used by a broad range of car manufacturers. It offers a great variety of relevant information material.

The computer simulation provides a large number of authentic diagnostic problem-solving steps: The "electronic engine management" alone allows more than a 1000 different user actions to measure voltage, resistance and signals. A guiding principle when developing the simulation was to allow interactions that closely align with occupational reality of car mechatronics. The computer-based assessment proved to produce valid test score interpretations, that is, measures indicating authentic diagnostic problem-solving skills (Gschwendtner et al., 2009).

To solve a diagnosis problem (i.e., to detect the problems' causes), electrotechnical measurements had to be conducted. The sensor problems allowed using the computer-based expert system to retrieve location diagrams, circuit diagrams and test instructions as well as to read out error-messages of the car's electronic control unit. Test instructions contained information on electrotechnical measurements useful for solving the diagnosis problem. For the lighting problems, the computer-based expert system was not useful as the lighting system is not connected to an electronic control unit, which is a precondition of using the expert system. Instead, electronic copies of relevant circuit diagrams were made available in the computer simulation.

As described in a previous article (Abele, 2017), the standardized instruction for the assessment took 30 min. Initially, the instructor demonstrated the handling of the simulation by means of a video presentation. Afterwards, the apprentices worked individually on standardized tasks concerning the handling of the simulation. In very rare cases, apprentices could not complete a task. Then, the instructor provided explanations to the class using a projector. Finally, the apprentices were instructed how to document their problem-solving results.

22.6.2.2 Strategy Patterns of Critical Information Behavior and Critical Test Behavior

To score critical information behavior and critical test behavior, the three-step procedure described above was applied: Based on the symptoms given in the problem descriptions, critical diagnostic hypotheses were identified. Drawing on these hypotheses, the critical information and the critical tests were determined and used to identify critical information behavior and critical test behavior. To detect whether diagnosticians exhibited the relevant behavior to solve the diagnosis

problem, relevant data was extracted from the computer-generated log-files. Diagnosticians showing relevant behavior received the score "1", the others the score "0".

The symptoms in the sensor problem descriptions were quite general implying many critical diagnostic hypotheses. In this study, only the four critical diagnostic hypotheses directly connected to the fuel temperature sensor were included. Furthermore, some critical diagnostic hypotheses of the lighting problems could not be considered, as the computer simulation did not allow for testing them (e.g., light switch defect). The apprentices received information about this limitation in the test instruction. Five critical diagnostic hypotheses were formulated for the lighting problems.

Tables 22.1 and 22.2 document the strategy patterns of critical information behavior and critical test behavior. These patterns were determined based on theoretical conceptualizations of the knowledge-based diagnostic strategies. As can be seen in the tables, each strategy is associated with a high probability in selecting the problem description, which is very plausible as this behavior is the fundamental requirement to apply any of the strategies.

For the *sensor problems* (Table 22.1), there were no expectations in view of retrieving the location diagram to find the fuel temperature sensor: Information on the location of components is not indicative of any of the strategies. Therefore, the probability of conducting that behavior could be anything, low, middle or high for each of the strategies. The *computer-based strategy* is associated with high probabilities of critical information behavior needed to retrieve information from the computer-based expert system. Moreover, diagnosticians using the computer-based strategy very likely test the fuel temperature sensor: This critical test behavior directly follows from the information given by the expert system. The *case-based strategy* also relates to a high probability in testing the sensor assuming a broken fuel

Table 22.1 The strategies' patterns of critical information behavior and critical test behavior for the sensor problems

Critical problem-solving behavior		Computer-based strategy	Case-based strategy	Model-based strategy
Information behavior	Sensor problem description	+	+	+
	Expert system instruction 1	+	−	−
	Expert system instruction 2	+	−	−
	Sensor circuit diagram	−	−	+
	Sensor location diagram	+/−	+/−	+/−
Test behavior	Sensor test	+	+	+
	Sensor cable test 1	−	−	+
	Sensor cable test 2	−	−	+
	Engine control unit test	−	−	+

Note: +: high probability; −: low probability; +/−: low, middle or high probability depending on knowledge about the location of the sensor

Table 22.2 The strategies' patterns of critical information behavior and critical test behavior for the lighting problems

Critical problem-solving behavior		Case-based strategy	Model-based strategy
Information behavior	Lighting problem description	+	+
	Lighting circuit diagram	−	+
	Fuse diagram	+/−	+/−
Test behavior	Fuse test	+	+
	Lightbulb test	+	+
	Lighting cable test 1	−	+
	Lighting cable test 2	−	+
	Lighting cable test 3	−	+

Note: +: high probability; −: low probability; +/−: high, middle or low probability depending on knowledge about the location of the relevant fuse

temperature sensor occurs frequently in practice and, therefore, many diagnosticians may have activated a corresponding defect script. In this context, it is important to note that the study sample included individuals having a significant yet limited diagnostic problem-solving experience reducing the chance that many respondents acquired additional defect scripts. Other critical test behavior than "sensor test" refer to more specific and rare critical diagnostic hypotheses and are shown less likely by apprentices using the case-based strategy. Turning to the *model-based strategy*, retrieving the circuit diagram was considered very important as viewing this diagram greatly facilitates the development of a mental model needed to systematically derive critical diagnostic hypotheses. As the model-based strategy allows testing all critical diagnostic hypotheses, each of the listed critical test behaviors are expected to have a higher probability compared to the other strategies.

In terms of the *lighting problems*, the computer-based expert system was not applicable: The car's lighting system did not have an electronic control unit, which is a precondition to use the computer-based expert system. We supposed that a substantial number of apprentices had two defect scripts available: The first defect script referred to a broken fuse causing the lighting system defect, the second script referred to a broken headlight lamp. From this it follows, that the *case-based strategy* should have high probabilities in testing the fuse and the respective lamp (Table 22.2). In contrast, the probabilities of the other critical test behavior should be rather low. Apart from "fuse test", we expected low probabilities in retrieving the circuit diagram. "Fuse diagram" represented retrieving the fuse diagram to identify the location of the relevant fuse and was neither indicative of the case-based nor the model-based strategy. For the *model-based strategy*, we anticipated a high probability of critical information behavior (except for "fuse diagram") and the critical test behavior.

22.6.3 Statistical Analysis

The data were analyzed by means of model selection procedures based on estimates obtained using a range of customary psychometric models. This allows an evaluation of whether the assumed multiple strategy model, implemented as a diagnostic classification model, can fit the observed behavioral data better than standard psychometric models that would provide alternative explanations of the data.

The psychometric models used in the analyses are aimed at testing the four research hypotheses described above:

1. Unidimensional item response theory (IRT; e.g., Lord & Novick, 1968; van der Linden, 2018) which is a family of models implementing the assumption of a single underlying continuous latent variable responsible for monotonic differences in the observed data. Respondents with higher levels of the latent variable show a higher propensity of the observed behavior. This would correspond to the assumption that the different strategies can be characterized as ordered levels of proficiency when interacting with diagnosis problems of the type studied here.
2. Latent Class Analysis (LCA; Lazarsfeld & Henry, 1968) is a model that implements the assumption of qualitative differences between respondents without imposing a monotonic relationship between strategy variables and observed behaviors: Respondents that are members of different latent strategy classes may have different, potentially intersecting, profiles of propensities with respect to the set of observed behaviors.
3. Diagnostic classification models (e.g., Junker & Sijtsma, 2001; von Davier, DiBello, & Yamamoto, 2008) are analytic approaches that can be considered constrained latent class models (von Davier, 2009) implementing multiple latent dichotomous or ordinal variables (von Davier, 2005, 2008; von Davier & Rost, 2016) representing an expert generated hypothesis of how multiple skill (here: strategy) variables are related to observed behaviors. The strategy patterns given in Tables 22.1 and 22.2 were used in conjunction with the general diagnostic model, one of the most general frameworks in this domain (GDM; von Davier, 2008, 2013, 2014) as input for the skill-attribute design matrix that is commonly referred to as Q-matrix (Tatsuoka, 1983) in this context.

This range of models, while independently developed over more than half a century, can be represented as special cases of general latent variable models (e.g., McDonald, 1999; Moustaki & Knott, 2000; Skrondal & Rabe-Hesketh, 2004; von Davier, 2008; von Davier & Yamamoto, 2004) and analyzed with toolkits that allow specification of these models within a common framework. The analyses of response data with this range of models was carried out with the software *mdltm* (von Davier, 2008), which is a general stand-alone program that was developed for large scale psychometric analyses including operational psychometric analyses for international large scale studies such as OECD PIAAC and PISA 2015 & 2018 (von Davier et al., 2019; Yamamoto, Khorramdel & von Davier, 2013).

22.7 Results

22.7.1 Descriptive Statistics

Tables 22.3 and 22.4 display descriptive statistics for the sensor problems and lighting problems. As expected, the difficulty of the problems differed remarkably within the pairs. Moreover, the relative frequencies of the critical problem-solving behavior varied considerably depending on problem difficulty.

Table 22.3 Descriptive statistics for the critical problem-solving behavior and the sensor problems

Critical problem-solving behavior	Sensor problems:	Easy Relative (absolute) frequencies	Difficult Relative (absolute) frequencies
Sensor problem description		.98 (351)	.98 (351)
Expert system instruction 1		.41 (146)	.45 (161)
Expert system instruction 2		.31 (111)	.37 (131)
Sensor circuit diagram		.13 (46)	.35 (124)
Sensor location diagram		.34 (121)	.33 (119)
Sensor test		.59 (214)	.49 (177)
Sensor cable test 1		.03 (10)	.15 (52)
Sensor cable test 2		.03 (9)	.12 (42)
Engine control unit test		.01 (5)	.10 (37)
Difficulty		.56 (196)	.16 (54)

Note: $N = 348–360$

Table 22.4 Descriptive statistics of the critical problem-solving behavior and the lighting problems

Critical problem-solving behavior	Sensor problems:	Easy Relative (absolute) frequencies	Difficult Relative (absolute) frequencies
Lighting problem description		.97 (358)	.97 (354)
Lighting circuit diagram		.69 (253)	.59 (214)
Fuse diagram		.25 (92)	.71 (259)
Fuse test		.76 (279)	.79 (286)
Lightbulb test		.75 (278)	.85 (308)
Lighting cable test 1		.08 (29)	.16 (59)
Lighting cable test 2		.03 (12)	.16 (59)
Lighting cable test 3		.02 (8)	.10 (35)
Difficulty		.72 (239)	.12 (41)

Note: $N = 332–369$

22.7.2 Model Selection

Tables 22.5 and 22.6 provide information criteria (AIC, BIC, CAIC) used for model selection, balancing the likelihood of the data under the model, and model complexity (Akaike, 1973; Bozdogan, 1987; Schwarz, 1978). The smaller the information criteria, the better the model is considered to fit the data, taking the number of parameters needed to fit the model into account.

For the sensor problems, it appears that the GDM, that is, the model that explicitly imposes strategy-based probability differences, fits the data best among the models estimated here. While for the 'difficult' version of the sensor problem, the BIC and CAIC point to the GDM and the AIC points to the 4-class LCA solution, the GDM is favored by AIC, BIC, and CAIC in the case of the 'easy' sensor problem.

Given that the AIC tends to over-fit data and favors more complex than necessary solutions (Bozdogan, 1987; Schwarz, 1978), we accept the GDM as the preferred solution in both sensor cases. This has the added value that the GDM allows classifications of each test taker into strategy classes that can be considered direct indicators of the application of the case-based, computer-based and model-based strategies described above.

Tables 22.7 and 22.8 summarize the model selection criteria for the lighting problems. It turns out that the model selection procedures show a less well-determined distinction between the models estimated for these datasets: While 2 separate strategies were assumed, the empirical data do not clearly support this hypothesis. The AIC supports assuming 3 or even 4 latent classes, which would

Table 22.5 Information criteria for the 'easy' sensor problem data fitted using IRT, LCA and DCM models, all estimated in the software mdltm (von Davier, 2008) as special cases of the general diagnostic model (GDM)

Model	AIC	BIC	CAIC	Likelihood	Parameters
IRT	2835.90	2916.71	2936.71	−1397.95	20
LCA 2 classes	2879.94	2960.74	2980.74	−1419.97	20
LCA 3 classes	2589.22	2714.46	2745.46	−1263.61	31
LCA 4 classes	**2553.01**	2722.70	2764.70	−1234.50	42
GDM	2591.59	**2700.68**	**2727.68**	−1268.79	27

Table 22.6 Information criteria for the 'difficult' sensor problem data fitted using IRT, LCA and DCM models, all estimated in the software mdltm (von Davier, 2008) as special cases of the general diagnostic model (GDM)

Model	AIC	BIC	CAIC	Likelihood	Parameters
IRT	2251.05	2331.76	2351.76	−1105.52	20
LCA 2 classes	2152.97	2233.67	2253.67	−1056.48	20
LCA 3 classes	2084.48	2209.58	2240.58	−1011.24	31
LCA 4 classes	2082.09	2251.58	2293.58	−999.04	42
GDM	**2078.92**	**2187.87**	**2214.87**	−1012.46	27

Table 22.7 Information criteria for the 'difficult' lighting problem data fitted using IRT, LCA and DCM models, all estimated in the software mdltm (von Davier, 2008) as special cases of the general diagnostic model (GDM)

Model	AIC	BIC	CAIC	Likelihood	Parameters
IRT	2244.19	**2308.84**	**2328.84**	−1106.09	20
LCA 2 classes	2309.00	2381.72	2401.72	−1136.50	20
LCA 3 classes	**2214.35**	2327.48	2358.48	−1079.17	31
LCA 4 classes	**2214.19**	2367.72	2409.72	−1069.10	42
GDM	2304.44	2397.37	2424.37	−1129.22	27

Table 22.8 Information criteria for the 'easy' lighting problem data fitted using IRT, LCA and DCM models, all estimated in the software mdltm (von Davier, 2008) as special cases of the general diagnostic model (GDM)

Model	AIC	BIC	CAIC	Likelihood	Parameters
IRT	1997.92	**2062.49**	**2082.49**	−982.96	20
LCA 2 classes	2016.93	2089.57	2109.57	−990.46	20
LCA 3 classes	**1946.90**	**2059.89**	2090.89	−945.45	31
LCA 4 classes	**1948.24**	2101.58	2143.58	−936.12	42
GDM	2013.41	2106.23	2133.23	−983.70	27

provide some support for the hypothesis, the BIC and CAIC favor a unidimensional IRT model. As these results only partially support the expected strategy types, we will focus on the sensor problems for which a diagnostic model with 3 distinct strategies is supported by the data analyses in the following sections. An in-depth analyses of the sensor problems and lighting problems using the LCA modeling approach is provided in Abele and von Davier (2018).

22.7.3 Distribution of Strategy Types

The distribution of strategy types is represented in what diagnostic classification models call the skill attribute distribution. For three binary skill variables representing the application of the case-based, computer-based and model-based strategies, respectively, we obtain a discrete distribution with $2^3 = 8$ potential outcomes.

It is important to note that strategy variables were not assumed to be mutually exclusive, respondents may have used more than one strategy, or they may have produced observed behavior that was compatible with more than one strategy.

(a) **Easy sensor problem**

For the easy sensor problem, the distribution of strategy types is given in Table 22.9. It can be observed that a certain percentage of respondents were identified as not having applied any strategy, these respondents were not using any of the diagnostic actions given in Table 22.1.

Table 22.9 Distribution of strategy types for the easy sensor problem. Multiple strategies or no strategy could be indicated, depending on the observed behavior of diagnosticians

Computer	Case	Model	%
−	−	−	22.4
+	−	−	14.1
−	+	−	37.0
+	+	−	23.3
−	−	+	0.7
+	−	+	0.4
−	+	+	1.2
+	+	+	0.7

Table 22.10 Distribution of strategy types for the difficult sensor problem. Multiple strategies or no strategy could be indicated, depending on the observed behavior of diagnosticians

Computer	Case	Model	%
−	−	−	22.7
+	−	−	16.8
−	+	−	25.9
+	+	−	19.2
−	−	+	4.1
+	−	+	3.0
−	+	+	4.6
+	+	+	3.4

There is an obvious result to be gleaned from this table: The model-based strategy appears to be not applied by most respondents when working on the easy sensor problem. Only 3% appear to select the model-based strategy in this case. The computer-based strategy (consulting the expert system) was applied by $(14.1 + 23.3 + 0.4 + 0.7) = 38.5\%$ of respondents, while the case-based strategy was applied by $(37 + 23.3 + 1.2 + 0.7) = 62.2\%$ of cases. Note that 22.4% of the sample did not get assigned any strategy use, as these respondents did not choose any of the behavioral indicators that are associated with the three strategies.

(b) **Difficult sensor problem**

The difficult sensor problem shows a higher proportion of respondents who are classified as having applied the model-based strategy. Based on the estimates provided in Table 22.10, about 15.1% of respondents appear to have selected a model-based strategy. Again, about 22.7% (compared to 22.4% in the easy case) of respondents did not show any of the behaviors associated with the strategies described in this chapter. The computer-based strategy was applied by $(16.8 + 19.2 + 3.0 + 3.4) = 42.4\%$ and the case-based strategy was applied by $(25.9 + 19.2 + 4.6 + 3.4) = 53.1\%$ of respondents.

Note that those who exclusively used the case-based strategy were about 37.0% in the easy sensor case, but only 25.9% selected exclusively a case-based strategy in the difficult sensor case.

(c) Maximum a Posteriori Strategy Classification and Changes from Easy to Difficult Sensor Problem

While the previous two subsections examined the distribution of strategy application based on the model parameters, this section looks at the distribution of strategy usage if the task is to provide the best guess for each respondent based on their observed data. For that purpose, maximum a-posteriori (MAP) classifications were generated using the mdltm software (von Davier, 2008), and imported to a spreadsheet software for further processing.

Table 22.11 shows the frequencies of the three strategies based on MAP classifications separately for each of the sensor problems. Note that some respondents may not have employed any strategy, while others may have shown response behavior that is indicative of more than one strategy. This is the reason why the total will not add up to the sample size and is therefore not a meaningful summary. The sample size differed by 2 (0.6%) between difficult and easy sensor cases in the data available for this analysis, and 351 respondents could be matched between the two experiments.

Ignoring for now the minimal difference of 2 cases in sample size between experiments, it is evident that more respondents appear to use the model-based strategy while working on the difficult sensor problem than when working on the easy problem. The same holds for the computer-based strategy, while it appears that the change is more moderate. At the same time, fewer respondents appear to employ the case-based strategy when faced with the difficult sensor problem, which also agrees with expectations in that fewer respondents may have a readily available 'defect script' when working on a more difficult, or a less common, problem.

Table 22.12 looks at the 351 cases that are common between samples that took the two sensor problems and provides a cross-table of the joint frequency distribution for the separately computed MAP classifications.

Visual examination of the table makes it apparent that the movement from non-model-based strategies (only case-based or computer-based) to a strategy pattern that also involves model-based aspects is seen in the hypothesized way. Respondents who were classified in the groups (−,−,−), (+,−,−), (−,+,−) or (+,+,−) are representing the vast majority of 341 out of 351 respondents for the easy sensor problem, and 42 of these are moving to a strategy that involves model-based behaviors (−,−,+) (+,−,+) (−,+,+) or (+,+,+) when working on the difficult problem, while only 2 respondents who were classified as having applied the model-based strategy for the easy problem appear not to have done so in the difficult problem.

Table 22.11 Frequency distribution of the case-based, computer-based and model-based strategy types using maximum a-posteriori classifications

	Computer	Case	Model	Sample size
Easy sensor problem	144	226	10	351
Difficult sensor problem	160	157	50	353

Table 22.12 Distribution of strategy types for the difficult sensor problem

Strategy usage	(Computer, Case, Model) difficult sensor problem								
Easy sensor problem	(−,−,−)	(+,−,−)	(−,+,−)	(+,+,−)	(−,−,+)	(+,−,+)	(−,+,+)	(+,+,+)	Grand total
(−,−,−)	57	14	11	3	*1*		*2*		88
(+,−,−)	9	15	5	6					35
(−,+,−)	21	15	34	21	*2*	*2*	*13*	*4*	112
(+,+,−)	13	34	8	33	*2*	*8*	*5*	*3*	106
(−,−,+)									
(+,−,+)	*1*							**1**	2
(−,+,+)			*1*				**5**	**1**	7
(+,+,+)							**1**		1
Grand total	101	78	59	63	5	10	26	9	351

Multiple strategies or no strategy could be indicated, depending on the observed behavior of diagnosticians. Frequencies that indicate a move from non-model-based to model-based are in the off-diagonal elements and are printed in italics; frequencies of respondents who were classified into model-based strategy types in both cases are printed boldface

Only eight respondents (frequencies printed in boldface) appear to have used the model-based strategy for both the easy and the difficult sensor problem, while the other 343 respondents relied at least once on only case-based or computer-based strategies, and 299 respondents appear not to have used the model-based strategy at all, neither in the easy sensor problem, nor when attempting to solve the difficult problem.

While 35 respondents only employed the computer-based strategy (characterized as (+,-,-) in terms of attribute patterns in the diagnostic model) when working on the easy problem, this number more than doubled to 78 respondents who utilized only the computer-based strategy in the difficult sensor problem.

22.8 Discussion

The results presented in this chapter show that applications of DCMs can be useful for the identification of problem-solving strategies used by respondents in a computer-based diagnosis task. Trouble-shooting, bug-fixing, diagnosis of faulty technical systems, or diagnostic work in health care all require that the agent who tries to solve one of these problems applies the best possible strategy or combination of strategies to maximize outcomes. Computer-based tests that authentically simulate diagnostic problems can be used to trace what respondents faced with these types of problems do to successfully complete real-world tasks.

This chapter showed how a theory developed around the problem-solving process can be used to derive what types of evidence may be relevant for the identification of strategy types. The log-files produced by a computer-based assessment system provide a rich basis for this type of analyses, often containing several hundred or thousands of log-entries per respondent. However, expert knowledge about the problem-solving process is needed to derive higher-level aggregates of these very fine-grained log entries to define construct relevant behavioral indicators that can be used as indicator variables in a statistical analysis aimed at identification of problem solving strategies.

The model selection strategies applied here follow best statistical practice by balancing model-data fit with model parsimony. In doing so, it appears that the sensor problems produced a clear picture and the model favored by information criteria was indeed the diagnostic model we set out to examine. However, for the lighting problems, no such result could be obtained, so further analyses are required in a future study to examine the behaviors of respondents based on data collected for this less complex problem type. It should be noted that only using a diagnostic model in isolation would not have resulted in a rejection of the diagnostic model for the lighting problems. It is therefore recommended to follow a procedure that tests diagnostic models against other, more parsimonious models to ensure that the interpretations made when selecting a model can be made with some confidence

that the model chosen holds up well in terms of model data fit against a range of alternative models.

The results for the difficult and easy sensor problems largely agree with the research hypotheses developed in the theory sections of this chapter: More respondents appear to use the model-based strategy when faced with a difficult problem, and fewer use the case-based strategy here, compared with the easy sensor problem. It appears that the strategy shift seen also confirms the expected move when looking at a cross tabulation of the movement between strategy groups when the problem type changes from easy to difficult. Although not stated as research hypothesis, the findings also suggest another difficulty-induced strategy shift from the case-based to the computer-based strategy: Comparing both strategies, the computer-based strategy is associated with higher costs (e.g., more time and activities) and therefore especially applied when the case-based strategy is not successful (i.e., when difficult problems are solved).

Future directions of research could involve a direct modeling of strategy shift, potentially involving covariates of strategy classifications that could be based on, for example, variables that represent relevant experience of respondents, curricula, or institutions in a multilevel diagnostic model (e.g., von Davier, 2007), or diagnostic models that incorporate the change or growth (von Davier, Xu, & Carstensen, 2011) over the course of multiple encounters with diagnostic problems.

References

Abele, S. (2014). *Modellierung und Entwicklung berufsfachlicher Kompetenz* [Modeling and development of vocational competence]. Stuttgart, Germany: Franz Steiner.

Abele, S. (2017). Diagnostic problem-solving process in professional contexts: Theory and empirical investigation in the context of car mechatronics using computer-generated log-files. *Vocations and Learning, 11*, 133–159.

Abele, S., & von Davier, M. (2018). *Applying cognitive diagnosis models and latent class analysis to computer-generated process data to identify diagnostic problem-solving strategies in car mechatronics*. Manuscript in preparation.

Abele, S., Walker, F., & Nickolaus, R. (2014). Zeitökonomische und reliable Diagnostik beruflicher Problemlösekompetenzen bei Auszubildenden zum Kfz-Mechatroniker. *Zeitschrift für Pädagogische Psychologie, 28*, 167–179.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Akademiai Kiado.

Baethge, M., & Arends, L. (2009). *Feasibility study VET-LSA: A comparative analysis of occupational profiles and VET programmes in 8 European countries—International report. Vocational training research* (Vol. 8). Bielefeld, Germany: Bertelsmann.

Billett, S. (2011). *Vocational education: Purposes, traditions and prospects*. Dordrecht, The Netherlands: Springer.

Boshuizen, H. P., & Schmidt, H. G. (2008). The development of clinical reasoning expertise. In J. Higgs, M. A. Jones, S. Loftus, & N. Christensen (Eds.), *Clinical reasoning in the health professions* (3rd ed., pp. 113–121). Oxford, UK: Elsevier Ltd.

Bozdogan, H. (1987). Model selection and Akaike's information cri- terion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education, 37*, 695–703.

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022–1028.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical problem solving: A ten-year retrospective. *Evaluation & the Health Professions, 13*(1), 5–36.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53.

Goldhammer, F., Kröhne, U., Keßel, Y., Senkbeil, M., & Ihme, J. M. (2014). Diagnostik von ICT-literacy: Multiple-choice- vs. simulationsbasierte Aufgaben. *Diagnostica, 60*(1), 10–21.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105.

Gschwendtner, T., Abele, S., & Nickolaus, R. (2009). Computersimulierte Arbeitsproben: Eine Validierungsstudie am Beispiel der Fehlerdiagnoseleistungen von Kfz-Mechatronikern Mechatronikern [Computer-simulated work samples: A statistical validation study using the example of trouble-shooting competency of car mechatronics]. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 105*, 557–578.

Hanushek, E. A., Schwerdt, G., Woessmann, L., & Zhang, L. (2017). General education, vocational education, and labor-market outcomes over the lifecycle. *Journal of Human Resources, 52*(1), 48–87.

He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with N-grams. In A. van der Ark, D. Bolt, S. Chow, J. Douglas, & W. Wang (Eds.), *Quantitative psychology research: Proceedings of the 79th annual meeting of the psychometric society* (pp. 173–190). New York, NY: Springer. https://doi.org/10.1007/978-3-319-19977-1_13

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749–776). Hershey, PA: Information Science Reference. https://doi.org/10.4018/978-1-4666-9441-5.ch029

Hoc, J.-M., & Amalberti, R. (1995). Diagnosis: Some theoretical questions raised by applied research. *Current Psychology of Cognition, 14*(1), 73–101.

Jonassen, D. H. (2011). *Learning to solve problems. A handbook for designing problem-solving learning environments*. New York, NY: Routledge.

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-based reasoning and instructional design using stories to support problem solving. *Educational Technology Research and Development, 50*, 65–77.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Kassirer, J., Wong, J., & Kopelman, R. (2010). *Learning clinical reasoning* (3rd ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Kluwe, R. H., & Haider, H. (1990). Modelle zur internen Repräsentation komplexer technischer Systeme. *Sprache & Kognition, 9*(4), 173–192.

Konradt, U. (1995). Strategies of failure diagnosis in computer-controlled manufacturing systems: Empirical analysis and implications for the design of adaptive decision support systems. *International Journal of Human-Computer Studies, 43*(4), 503–521.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin Company.

Lerman, R. I. (2016). Restoring opportunity by expanding apprenticeship. In I. Kirsch & H. Braun (Eds.), *The dynamics of opportunity in America*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-25991-8_10

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Moustaki, I., & Knott, M. (2000). Generalised latent trait models. *Psychometrika, 65*(3), 391–411.

Nickolaus, R., Abele, S., Gschwendtner, T., Nitzschke, A., & Greiff, S. (2012). Fachspezifische Problemlösefähigkeit in gewerblich-technischen Ausbildungsberufen—Modellierung, erreichte Niveaus und relevante Einflussfaktoren [Occupation-specific problem solving competency as an essential competency dimension of professional competency: Models, achieved levels and relevant predictors in technical education]. *Zeitschrift für Berufs- und Wirtschaftspädagogik, 108*, 243–272.

Norman, G., Young, M., & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education, 41*, 1140–1145.

Norman, G. R. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*, 418–427.

Perez, R. S. (2012). A view from troubleshooting. In M. U. Smith (Ed.), *Toward a unified theory of problem solving* (pp. 127–166). New York, NY: Routledge.

Rasmussen, J. (1981). Models of mental strategies in process plant diagnosis. In J. Rasmussen & W. Rouse (Eds.), *Human detection and diagnosis of system failures* (NATO conference series, Bd. 3, Bd. 15, S., pp. 241–258). New York, NY: Springer US.

Rasmussen, J. (1993). Diagnostic reasoning in action. *IEEE Transactions on Systems, Man and Cybernetics, 23*(4), 981–992. https://doi.org/10.1109/21.247883

Rausch, A., Seifried, J., Wuttke, E., Kögler, K., & Brandt, S. (2016). Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. *Empirical Research in Vocational Education and Training, 8*(1), 1–23.

Rouse, W. B. (1983). Models of human problem solving: Detection, diagnosis, and compensation for system failures. *Automatica, 19*(6), 613–625. https://doi.org/10.1016/0005-1098(83)90025-0

Schaper, N., Hochholdinger, S., & Sonntag, K. (2004). Förderung des Transfers von Diagnosestrategien durch computergestütztes Training mit kognitiver Modellierung [Improving transfer of troubleshooting skills by computer-based training with modeling]. *Zeitschrift für Personalpsychologie, 3*(2), 51–62.

Schwartz, A., & Elstein, A. S. (2008). Clinical reasoning in medicine. In J. Higgs, M. A. Jones, S. Loftus & N. Christensen (Eds.), *Clinical reasoning in the health professions* (3. Aufl., pp. 223–234). Oxford, UK: Elsevier Ltd.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Sembill, D., Rausch, A., & Kögler, K. (2013). Non-cognitive facets of competence. Theoretical foundations and implications for measurement. In O. Zlatkin-Troitschanskaia & K. Beck (Eds.), *From diagnostics to learning success—Proceedings in vocational education and training* (pp. S. 199–S. 212). Rotterdam, The Netherlands: Sense.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

van Merriënboer, J. G. (2013). Perspectives on problem solving and instruction. *Computers & Education, 64*, 153–160.

van der Linden, W. (2018). *Handbook of item response theory*, three volume set. Chapman & Hall/CRC statistics in the social and Behavioral sciences. ISBN: 148228247X, 9781482282474

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series, 2005*, i–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2007). *Hierarchical general diagnostic models*. Research Report, RR-07-19. Princeton, NJ: ETS. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.2007.tb02061.x

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2009, March). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement—Interdisciplinary Research and Perspectives, 7*(1), 67–74.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model—Two variants of a model equivalency. *BJMSP, 67*, 49–71.. http://onlinelibrary.wiley.com/doi/10.1111/bmsp.12003/abstract

von Davier, M. (2014), The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). ETS research report series. http://onlinelibrary.wiley.com/doi/10.1002/ets2.12043/abstract

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, 2nd ed., pp. 393–406). Boca Raton, FL: CRC Press. http://www.crcnetbase.com/doi/abs/10.1201/9781315374512-24

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis*. Paper presented at the 4th Spearman conference "Diagnostics for education: Theory, measurement, applications." ETS: The Inn at Penn, Philadelphia, PA.

von Davier, M., DiBello, L., & Yamamoto, K. (2008). Reporting test outcomes using models for cognitive diagnosis. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 151–176). Göttingen, Germany: Hogrefe & Huber Publishers.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336. https://doi.org/10.1007/s11336-011-9202-z

von Davier, M., Yamamoto, K., Shin, H.-J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). *Evaluating item response theory linking and model fit for data from PISA 2000–2012*. Assessment in Education: Principles, Policy & Practice.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). *Scaling PIAAC cognitive data. Technical report of the survey of adult skills (PIAAC)*. Paris: Organisation for Economic Co-operation and Development.

Chapter 23
Applying the General Diagnostic Model to Proficiency Data from a National Skills Survey

Xueli Xu and Matthias von Davier

Abstract Large-scale educational surveys (including NAEP, TIMSS, PISA) utilize item-response-theory (IRT) calibration together with a latent regression model to make inferences about subgroup ability distributions, including subgroup means, percentiles, as well as standard deviations. It has long been recognized that grouping variables not included in the latent regression model can produce secondary bias in estimates of group differences (Mislevy, RJ, Psychometrika 56:177–196, 1991). To accommodate the ever-increasing number of background variables collected and required for reporting purposes, a principal component analysis based on the background variables (von Davier M, Sinharay S, Oranje A, Beaton AE, The statistical procedures used in national assessment of educational progress: recent developments and future directions. In: Rao CR, Sinharay S (eds) Handbook of statistics: vol. 26. Psychometrics. Elsevier B.V, Amsterdam, pp 1039–1055, 2007; Moran R, Dresher A, Results from NAEP marginal estimation research on multivariate scales. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 2007; Oranje A, Li D, On the role of background variables in large scale survey assessments. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY, 2008) is utilized to keep the number of predictors in the latent regression models within a reasonable range. However, even this approach often results in the inclusion of several hundred variables, and it is unknown whether the principal component approach or similar approaches (such as latent-class approaches) are able to generate consistent estimates for individual subgroups (e.g., Wetzel E, Xu X, von Davier M, Educ Psychol Meas 75(5):1–25, 2014). The primary goal of the current study is to provide an exemplary application of diagnostic models for

X. Xu (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: xxu@ets.org

M. von Davier
National Board of Medical Examiners (NBME), Philadelphia, PA, USA
e-mail: mvondavier@nbme.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_23

large-scale-assessment data. Specifically, a latent-class structure is used for covariates while continuing to use IRT models for item responses in the analytic model. Previous applications focused on adult literacy data (von Davier M, Yamamoto K, A class of models for cognitive diagnosis. Paper presented at the 4th Spearman invitational conference, Philadelphia, PA, 2004), as well as large-scale English-language testing programs (von Davier M; A general diagnostic model applied to language testing data (Research report no. RR-05-16). Educational Testing Service, Princeton, 2005, von Davier M, The mixture general diagnostic model. In: Hancock GR, Samuelsen KM (eds) Advances in latent variable mixture models. Information Age Publishing, Charlotte, pp 255–274, 2008), while the current application uses diagnostic modeling approaches on data from NAEP.

23.1 Background and Research Questions

The National Assessment of Educational Progress (NAEP) is often called the "Nation's Report Card" and can be considered the standard of measuring academic progress across the United States for fourth-, eighth-, and 12th- grade students. It covers a wide range of subjects, including reading, mathematics, writing, science, and social science. Every 2 or 4 years, these assessments are administered to nationally representative samples in order to measure trends in academic progress over years. Depending on the assessment subject, the nationally representative samples can have sizes from about 12,000 to about 140,000.

Because NAEP aims to measure the academic progress in policy-relevant subgroups and is mandated not to provide measures at the individual level, a sparse matrix sampling design (Johnson, 1992) is employed to administer NAEP assessments so that individual students take only a portion of entire assessment. For example, for subscales defined in the mathematics framework, students take only about 10–30 items from a pool of 100–150 items, and each student is assigned one of a number of different test forms with a different set of items from the pool. The relatively small number of items within subscales does not provide good ability estimates for individual students, but the aggregation of individual ability distributions is suitable to provide precise estimates of subgroup ability distributions. The methodologies utilized to achieve this were described by von Davier et al. (2007) as well as von Davier and Sinharay (2014). For instance, in order to estimate ability distributions for boys and girls, the item responses as well as the self-reported gender variable needs be included in the model in order to obtain precise estimates for the ability distributions (Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009). Modeling approaches that include both grouping variables as well as response variables are commonly referred to as multiple group models. The multiple group model used in NAEP—and other assessments—takes IRT to link the item responses and latent abilities and a latent regression model, in which group differences are regressed on a potentially very large number of background variables.

To facilitate analyses and to enable researchers to answer this demand for inclusion of a large variety of grouping variables, the predictors used in this latent regression model (Mislevy, 1991) are preprocessed by extracting principal components from the bulk of background variables (von Davier et al., 2007; Dresher, 2006; Moran & Dresher, 2007; Oranje & Li, 2008). In the currently operational latent regression procedure used in NAEP, individual subgroup indicators are not used directly. An analysis is conducted to extract principal components (PCs) that explain 90% of the variance of the observed grouping variables. These principal components are then used as the predictors in the latent regression model. This preprocessing raises a question: Is this approach suitable for providing reasonably good estimates of subgroup ability distributions, or does the preprocessing remove some of the between-group differences since the set of predictors used in the model might incompletely reflect differences in grouping variables? It is known that the estimates for subgroups that are not included in the model are usually biased to a certain degree (Mislevy, 1991), but, to our knowledge, the extent to which estimates are biased for subgroups that are only partially represented by means of proxy variables in the form of PCs is unknown. Existing studies were not able to provide definitive answers. For example, Dresher (2006), via a simulation study, found that with a reasonable number of items per students, the latent regression model with PCs (that explains 90% of variance) outperformed the latent regression model with only the subgroup variable of interest in terms of bias and root mean square error. However, Oranje and Li (2008) did not find alarming differences between these two types of models using real data. It is noted that both research studies used NAEP operational software to obtain estimates, assuming that the students share a common covariance structure and their abilities follow a multivariate normal distribution. Their conclusions might change if we use estimation algorithms that allow the assumption of a normal distribution to be relaxed.

The current chapter aims to demonstrate how the estimates of subgroup ability distributions may change when using a different type of conditioning model. For example, we want to obtain estimates of ability distributions of subgroups defined by a background variable using relaxed assumptions. The data analysis uses IRT to calibrate item parameters and takes a conditioning model (either the operational model or some alternative) to estimate the ability distributions for subgroups of interest. To estimate the ability distributions for subgroups, three types of conditioning models were considered and compared (from simplest to most complicated): (a) a model with only the subgroup variable of interest involved, (b) a model with the subgroup variable of interest as well as another important background variable, and (c) a model with preprocessed predictors—in the form of latent class indicator variables. Unlike the approaches used in Wetzel, Xu, and von Davier (2014) where the probabilities of latent classes were used as predictors in the latent regression models, the latent class membership is used in this paper to conduct a multiple-group/multi-dimensional analysis. If the number of latent classes is large enough to sufficiently account for the variation among students, we assume that all three conditioning models will produce approximately the same estimates for

the ability distributions of the subgroup of interest. However, a full model with not-saturated latent classes might incorrectly reflect the group differences. These three conditioning models were analyzed under the general diagnostic model (GDM) framework and compared using NAEP data.

The rest of this chapter is organized as follows: Section 23.2 briefly introduces the GDM, Sect. 23.3 describes the data and procedures used in this study, and Sect. 23.4 shows the results obtained by using the GDM software *mdltm* (von Davier, 2005) for estimation. The last section discusses some of the results and provides further thoughts on the research question.

23.2 The General Diagnostic Model

The GDM (von Davier, 2005) is one of the general frameworks for cognitive diagnostic modeling. As the name suggests, the GDM, as the other cognitive diagnosis models, is mainly developed to diagnose skill levels on finer-grained skills for individual test takers. For example, in the analysis of the well-known fraction subtraction data (Tatsuoka, 1983), the rule space approach, which can be viewed as a deterministic cognitive diagnosis model, was used to make judgments on whether certain skills that are related to fraction calculations are mastered by individual students. Usually, for a test that requires multiple skills, a Q-matrix (Tatsuoka, 1983) is defined based on expert judgments and describes which items require which skills. Quite a few cognitive diagnosis models have been developed in the last two decades, and many of these are described in the first part of this book.

Practically all probabilistic models for cognitive diagnosis can be described as located latent class models (von Davier, 2009). This also applies to the GDM, which expresses the levels for each of the skills as locations on the real line. While this is straightforward for the GDM as it defines a dichotomous or polytomous latent variable for each skill (von Davier, 2005, 2008), even the mastery/non-mastery variable as used in the DINA model (see Chaps. 1 and 7 in this volume) can be defined by two real numbers. This gives a meaning to mastery levels. By using real-valued located latent class, the GDM can easily be extended to more than two ability levels on each of the skills variables. In addition, the GDM bridges the gap between diagnostic models and multidimensional IRT models, and it can be shown that this approach can fit data as well as MIRT models with a multivariate normal ability distribution (Haberman, von Davier, & Lee, 2009). Hence, the GDM can be used to estimate item parameters and latent ability distributions just as commercial IRT estimation programs such as Parscale/or software for MIRT estimation usually do by specifying the skill levels as quadrature points. However, the multivariate ability distribution used in the GDM implementation (Xu & von Davier, 2008a, 2008b) can be estimated freely and, therefore, is more flexible than a (multivariate) normal ability distribution.

23.3 Methodology

23.3.1 Data

Data from a NAEP Grade four reading assessment administered to a national sample was used in the analysis. The data includes 97 items in total and about 140,000 students, each of whom received about 30 items in a balanced incomplete block design (von Davier et al., 2007). The background information we considered for this study includes gender, race/ethnicity, individualized education plan (IEP), limited English proficiency (LEP), free school lunch, location, and computer access at home.

This NAEP instrument measures reading abilities in two content areas: (a) the literary subscale (47 items) and (b) the informational subscale (50 items). By design, the two subscales share no common items.

23.3.2 Analysis Procedure

For each subgroup under consideration (race/ethnicity subgroups, gender groups, as well as school lunch groups), a number of models were estimated: (a) M1: subgroup-based two-dimensional model and (b) M2:latent-class-based two-dimensional model. The models were defined as follows:

M1: Subgroup-based two-dimensional model

In this model, item parameters and subgroup ability distributions are obtained simultaneously by calibrating a two dimensional IRT model (literary and informational subscales as two dimensions) in multiple populations defined by the grouping variables, while restricting item parameter estimates to be the same across subgroups. Under model M1, we estimate a multiple-group model with known assignment of each student to the subgroup of interest. Note that only a single nominal or dichotomous grouping variable is used in these cases, and that ability estimates are based on a Bayesian approach. The *mdltm* software allows expected a-posteriori, or maximum a-posteriori, or imputations based on the posterior distribution. Therefore, for race/ethnicity group comparisons, a model that contains the race/ethnicity grouping variable is appropriate, while for the school-lunch group-based analyses, comparisons of means of estimates between these groups only are appropriate.[1] This implies that any multiple-group analysis is a

[1]This is relevant in cases where Bayesian estimates of ability are used, and the knowledge about grouping, including the differences in ability distributions across groups, is utilized in the estimation. In cases where maximum-likelihood (ML), or bias-corrected ML is used, a multiple group model with item parameter equality will not produce more than trivially different estimates when different grouping variables are used, unless the item parameter estimates are affected by the grouping variables used. Note however, that ML and bias-corrected ML do not reduce measurement

one-time deal: If the grouping variable in the analysis model is not the same as the variable of interest; the estimates obtained from the multiple group model cannot be used for group comparisons. Based on the results reported by Mislevy (1991) and other subsequent publications on the use of this methodology, the gender-based calibration will not be suitable for race/ethnicity group comparisons, as these will likely result in secondary-biased estimates since the analysis model had no information on the race/ethnicity variable. The tables below contain four variants of M1: The first one (M1.1) that includes only the variable of interest, the second (M1.2) includes a grouping variable that is crossed with a second grouping variable (e.g., gender by race/ethnicity with 6 groups fully crossed in the example below), the third (M1.3) includes a grouping variable that is crossed with two other grouping variables (e.g. race/ethnicity by gender and by school lunch eligibility, and a fourth (M1.4) that includes a mismatched grouping variable (e.g., gender variable when deriving race/ethnicity subgroup results).

M2: Latent-class-based two-dimensional model

This model is similar to M1 but differs with respect to how subgroups are defined or, identified. In model M1, the subgroups are defined by observed background variables such as self-reported gender and race/ethnicity. However, in model M2, the subgroups are not assumed to be known, but rather are defined as clusters derived from students' background information. Specifically, the predictors in model M2 are based on the following steps: (a) Fit latent class models to the background variables available in addition to the response data and treat the estimated memberships as if they are observed; (b) then fit a two-dimensional multiple group IRT model to the item responses with groups defined by the estimated class membership.

Under this model, we used 5-latent-classes as groupings for the two-dimensional calibration (M2.1), 10-latent-classes as groupings for a two-dimensional calibration (M2.2) and 50-latent-classes as grouping for a two-dimensional calibration (M2.3). The latent classes based on the background data (e.g., gender, ethnicity, IEP, LEP, etc.) were obtained using the *mdltm* software (von Davier, 2005). The following table lists the levels for each of the background variables included in the latent class approach. The number of potential combinations of the levels of background variables is 1152, which equals the product over the number of levels across these variables given in Table 23.1 below. Each class profile can be represented by 14 parameters, so that there are sufficient degrees of freedom to estimate 10 (140 + 9 parameters) as well as 50 (700 + 49 parameters) latent classes.

This approach is different from assuming a mixture IRT model (e.g., von Davier & Rost, 2007, 2016) which would assume that there are unobserved groups that establish the differences in ability distributions. Instead, the background variables are used in the process of defining populations, but instead of principal components, class membership variables based on a latent class analysis involving all grouping variables of interest are generated.

error due to information about covariates, which is the main reason why background variables are used in latent regression models together with Bayesian ability estimates.

Table 23.1 Background variables used to derive latent classes of the NAEP reading student population

Variable	Levels
Gender	2
Race	6
IEP	2
LEP	2
School lunch	3
Location	4
Computer access	2

23.4 Results

The two models M1 and M2 are compared in terms of subgroup mean and standard deviation (SD) estimates, taking the correct-subgroup two-dimensional calibration (M1.1) as the baseline for comparisons. For example, if the target of inference is for a gender group, we examine the estimates of gender group means and SDs from the multiple group model that contains the gender variable only (correct grouping variable used) to the estimates from the other multiple-group models. The comparisons on estimates for each subgroup have two tables. The first one lists the mean and SD estimates from the different models and the second one shows the difference ratio (other model/M1.1-1).

Tables 23.2, 23.3, 23.4, and 23.5 present the mean and standard deviation estimates and difference ratios for gender, race/ethnicity, and school-lunch-status subgroups. The subgroups comprise the following proportions of the total sample: White students 51%, Black students 14%, Hispanic students 25%, school-lunch-eligible students 52%, and school-lunch not-eligible students 42%.

The following patterns can be discerned:

- As expected, the incorrect-subgroup models produce estimates that are different from the base model M1.1 (e.g., the model M1.4 for race/ethnicity-group inferences and M1.4 for school-lunch-group inferences).
- The model with an interaction including the reporting subgroup of interest produces estimates that are very close to the base model M1.1, such as M1.2 and M1.3 for race/ethnicity-group inference.
- The model with latent classes returns either a reasonably good estimate or inconsistent estimates compared to the correct-subgroup models. For example, the latent class models provide good estimates for White or Hispanic student groups compared with the baseline model, but showed somewhat imprecise estimates for the Black student group. It is unclear why this happens. One potential explanation is that smaller subgroups may not be fully represented in the 10 latent classes that were obtained. The Black student group makes up about 14% of the total sample, which is smaller than the proportions for the other two ethnicity-based subgroups.

Table 23.2 Mean and SD estimates for race/ethnicity subgroups from different conditioning models

	White (51%)				Black (14%)				Hispanic (25%)			
	Literary		Information		Literary		Information		Literary		Information	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
M1.1: Race-group-two-dimensional calibration	0.28	0.90	0.29	0.86	−0.37	0.86	−0.37	0.79	−0.31	0.87	−0.28	0.82
M1.2: Race-gender-crossed-group-two-dimensional calibration	0.28	0.90	0.29	0.85	−0.37	0.86	−0.37	0.79	−0.31	0.87	−0.28	0.82
M1.3: Race-gender-slunch-crossed-group-two-dimensional calibration	0.28	0.89	0.29	0.85	−0.37	0.88	−0.37	0.81	−0.31	0.88	−0.28	0.83
M1.4: Gender-group-two-dimensional calibration	0.22	0.91	0.22	0.87	−0.33	0.90	−0.30	0.85	−0.24	0.91	−0.22	0.86
M2.1: 5-lca-two-dimensional calibration	0.27	0.89	0.28	0.85	−0.35	0.90	−0.33	0.84	−0.29	0.90	−0.27	0.84
M2.2: 10-lca-two-dimensional calibration	0.27	0.89	0.28	0.85	−0.35	0.89	−0.33	0.83	−0.31	0.88	−0.28	0.83
M2.3: 50-lca-two-dimensional calibration	0.28	0.88	0.28	0.84	−0.35	0.90	−0.34	0.84	−0.30	0.91	−0.27	0.86

Table 23.3 The difference ratios for mean and SD estimates when compared to the race/ethnicity-group-two-dimension-model

	White				Black				Hispanic			
	Literary		Information		Literary		Information		Literary		Information	
	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio
M1.1: Race-group-two-dimension calibration	–	–	–	–	–	–	–	–	–	–	–	–
M1.2: Race-gender-crossed-group-two-dimensional calibration	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
M1.3: Race-gender-slunch-crossed-group-two-dimensional calibration	0%	−1%	0%	0%	0%	2%	0%	2%	−1%	2%	0%	1%
M1.4: Gender-group-two-dimensional calibration	−21%	2%	−22%	2%	−11%	4%	−18%	7%	−21%	4%	−23%	4%
M2.1: 5-lca-two-dimensional calibration	−2%	0%	−2%	0%	−8%	5%	−10%	6%	−5%	3%	−5%	2%
M2.2: 10-lca-two-dimensional calibration	−3%	0%	−3%	0%	−7%	4%	−11%	5%	−1%	1%	−2%	1%
M2.3: 50-lca-two-dimensional calibration	0%	−2%	−1%	−1%	−6%	5%	−8%	6%	−2%	5%	−3%	4%

Note: This model is taken as the baseline model

Table 23.4 Comparisons for school lunch subgroup estimates

	School lunch—Yes (52%)				School lunch—No (42%)			
	Literary		Information		Literary		Information	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
M1.1: School lunch-group-two-dimensional calibration	−0.30	0.87	−0.28	0.82	0.40	0.88	0.41	0.83
M1.2: School lunch-LEP-two-dimensional calibration	−0.30	0.88	−0.28	0.82	0.40	0.87	0.41	0.83
M1.3: School lunch-LEP-IEP-two-dimensional calibration	−0.29	0.90	−0.28	0.85	0.40	0.85	0.40	0.81
M1.4: Race-gender-group two-dimensional calibration	−0.26	0.89	−0.24	0.84	0.36	0.89	0.36	0.85
M2.1: 5-lca-two-dimensional calibration								
M2.2: 10-lca-two-dimensional calibration	−0.29	0.88	−0.28	0.82	0.39	0.88	0.40	0.84
M2.3: 50-lca-two-dimensional calibration	−0.29	0.90	−0.27	0.85	0.39	0.85	0.40	0.81

Table 23.5 The differences in mean estimates and the ratios of SD estimates when compared to the school lunch-group-two-dimension-model

	School lunch—Yes				School lunch—No			
	Literary		Information		Literary		Information	
	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio	Mean diff.	SD ratio
M1.1: School lunch-group-two-dimension calibration	–	–	–	–	–	–	–	–
M1.2: School lunch-LEP-two-dimensional calibration	0%	1%	−1%	1%	0%	−1%	0%	−1%
M1.3: School lunch-LEP-IEP-two-dimensional calibration	−2%	3%	−2%	4%	−1%	−3%	−2%	−3%
M1.4: Race-gender-group two-dimensional calibration	−13%	2%	−16%	3%	−11%	2%	−12%	2%
M2.1: 5-lca-two-dimensional calibration	−4%	1%	−5%	2%	−3%	0%	−4%	0%
M2.2: 10-lca-two-dimensional calibration	−2%	0%	−3%	0%	−2%	0%	−2%	1%
M2.3: 50-lca-two-dimensional calibration	−3%	3%	−5%	4%	−2%	−3%	−3%	−3%

Note: This model is taken as the baseline model

23.5 Summary

With the increasing scope of policy questions being raised in the context of NAEP, the number of background variables collected to obtain information for reporting purposes increased steadily over past assessment cycles. Educational large-scale survey assessments rely more and more on assumptions made in the latent regression in order to include all available background data. These models may use principal components as done in most operational programs (von Davier & Sinharay, 2014) or latent classes, as proposed by Wetzel et al. (2014) as predictors. Both approaches do not fully reflect the variability in the background data, but rather provide statistical summaries of the associations between the background variables collected in the assessment. The individual subgroup identification is replaced by such data summaries. The study presented in this chapter had the goal to investigate the possible effects of this data reduction. The findings reported above (a) confirm that the estimates for subgroups not included in the analysis models are biased, (b) confirm that the estimates for subgroups that are included in the form of fully crossed interaction models are consistent, and (c) raise concern regarding the use of data summaries (either latent classes or principal components) instead of observed background data. It appears that somewhat inconsistent estimates can result, in particular, if the subgroup information is only incompletely reflected in the statistical summaries that were used as predictors in the latent regression model. This implies that additional research may be needed to straighten this out. For example, the use of latent class analysis for auxiliary background data (Thomas, 2002) such as self-reports on out-of-school activities and educational resources at home together with a direct inclusion of the main reporting variables (gender, ethnicity, LEP, IEP, free school lunch) could be a promising way forward.

Note that the results presented here are limited by the number of background variables used to derive the latent classes. Only seven background variables were used. This approach is not comparable to the number of background variables used in the latent regression models applied in operational practice. A future study might expand to include all available background information to derive the latent classes (e.g., combined with the use of automatic variable selection methods).

References

Dresher, A. (2006, April). *Results from NAEP marginal estimation research*. Presented at the annual meeting of the national council on measurement in education, San Francisco, CA.

Haberman, S., von Davier, M., & Lee, Y.-H. (2009). *Comparison of multidimensional item response models: multivariate normal ability distributions versus multivariate polytomous ability distributions* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

Johnson, E. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement, 29*, 95–110.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristic from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–162.

Moran, R., & Dresher, A. (2007). *Results from NAEP marginal estimation research on multivariate scales.* Paper presented at the annual meeting of the national council on measurement in education, Chicago, IL.

Oranje, A., & Li, D. (2008, April). *On the role of background variables in large scale survey assessments.* Paper presented at the annual meeting of the national council on measurement in education, New York, NY.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika, 67*, 33–48. https://doi.org/10.1007/BF02294708

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 255–274). Charlotte, NC: Information Age Publishing.

von Davier, M. (2009, March). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement—Interdisciplinary Research and Perspectives, 7*(1), 67–74.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 9–36.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 643–661). Amsterdam, the Netherlands: Elsevier B.V.

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory. Vol. 1: Models* (pp. 393–406). Boca Raton, FL: Chapman and Hall/CRC.

von Davier, M., & Sinharay, S. (2014).. Analytics in international large-scale assessments: Item response theory and population models). In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). Boca Raton, FL: CRC Press.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. E. (2007). The statistical procedures used in national assessment of educational progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 1039–1055). Amsterdam, the Netherlands: Elsevier B.V.

von Davier, M., & Yamamoto, K. (2004, October). *A class of models for cognitive diagnosis.* Paper presented at the 4th Spearman invitational conference, Philadelphia, PA.

Wetzel, E., Xu, X., & von Davier, M. (2014). An alternative way to model population ability distributions in large-scale educational surveys. *Educational and Psychological Measurement, 75*(5), 1–25.

Xu, X., & von Davier, M. (2008a). *Fitting the structured general diagnostic model to NAEP data* (Research Report No. RR-08-27). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008b). *Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model* (Research Report No. RR-08-35). Princeton, NJ: Educational Testing Service.

Chapter 24
Reduced Reparameterized Unified Model Applied to Learning Spatial Rotation Skills

Susu Zhang, Jeff Douglas, Shiyu Wang, and Steven Andrew Culpepper

Abstract There has been a growing interest in measuring students in a learning context. Cognitive diagnosis models (CDMs) are traditionally used to measure students' skill mastery at a static time point, but recently, they have been combined with longitudinal models to track students' changes in skill acquisition over time. In this chapter, we propose a longitudinal learning model with CDMs. We consider different kinds of measurement models, including the reduced-reparameterized unified model (r-RUM) and the noisy input, deterministic-"and"-gate (NIDA) model. We also consider the incorporation of theories on skill hierarchies. Different models are fitted to a data set collected from a computer-based spatial rotation learning program (Wang S, Yang Y, Culpepper SA, Douglas JA, J Educ Behav Stat, 2016. https://doi.org/10.3102.1076998617719727) and we evaluate and compare these models using several goodness-of-fit indices.

24.1 Introduction

With the increasing use of online and computer-based instruction there is an opportunity for the development of psychometric models that can utilize this rich source of data and capture the dynamic nature of learning. Cognitive diagnosis

S. Zhang
Department of Statistics, Columbia University, New York, NY, USA
e-mail: sz2821@columbia.edu

J. Douglas (✉) · S. Andrew Culpepper
Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL, USA
e-mail: jeffdoug@illinois.edu; sculpepp@illinois.edu

S. Wang
Department of Educational Psychology, University of Georgia, Athens, GA, USA
e-mail: swang44@uga.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_24

models (CDMs) are naturally suited for this setting, because they allow for a fine breakdown of skills and abilities that can be monitored for discrete changes from nonmastery to mastery.

Computer-based learning also allows for the possibility of instruction to be administered together with assessment, which affords a chance to examine factors related to learning and assess when learning takes place. Utilizing CDMs in this setting requires that the latent attribute vector be allowed to change, which would typically be in a monotone fashion over a short duration, that is, within the course of learning, learners do not forget a mastered skill. There has been research on dynamic CDMs in the longitudinal setting, under which the respondents' attribute patterns are allowed to change over time (Kaya & Leite, 2016; Li, Cohen, Bottge, & Templin, 2015). These cases differ from our application in that much time could expire between assessments, and the attribute vector is seen as static within an assessment. Li et al. (2015) used the deterministic input, noisty-"and"-gate (DINA; e.g., Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model together with a transition model to measure the effects of an educational intervention. Through simulation studies Kaya and Leite (2016) studied such transition models for longitudinal applications using both the DINA and the DINO (Templin & Henson, 2006) models.

Our application to a spatial rotation skills intervention differs in that we consider learning taking place within a short time span. This could take place between items or between blocks of items, depending on how the intervention is administered. Using the same spatial rotation data as in this study, Wang et al. (2016) considered a hidden Markov model for attribute patterns to examine individual learners' attribute pattern trajectories. Both latent and observed covariates were incorporated to capture the effects of general learning ability, practice, and differences in how the intervention was administered. Chen, Culpepper, Wang, & Douglas (2017) also used the spatial rotation data to implement a learning model that considered a hidden Markov model for transition probabilities between all pairs of attribute patterns. Both Wang et al. and Chen et al. used the DINA model as the measurement model.

A similar approach to dynamic CDMs is the method of Knowledge Tracing (Corbett & Anderson, 1994), which has become popular in applications of intelligent tutoring systems. Knowledge Tracing has mostly focused on one attribute at a time. Studer (2012) proved that Knowledge Tracing is mathematically equivalent to an extension of the NIDA (Junker & Sijtsma, 2001; Maris, 1999) model for multiple time points, with a further strong restriction that each item can only depend on a single attribute. However, there have been several recent developments in Knowledge Tracing that allow for items to depend on several attributes at once, and have a wider variety of parameterizations (Xu & Mostow, 2012; González-Brenes & Mostow, 2013; González-Brenes, Huang, & Brusilovsky, 2014; Pardos & Heffernan, 2010).

24.1.1 Spatial Rotation Data

A computer-based assessment and training program was developed to conduct a study of learning spatial rotation skills (Wang et al., 2016). Subjects were students recruited from the paid subject pool of the Department of Psychology at the University of Illinois at Urbana-Champaign. Each subject was asked to complete a series of 50 items on a computer-based assessment of spatial rotation skills. The assessment items were comprised of an extended version of the Purdue Spatial Visualization Test (PSVT; Yoon, 2011). The assessment consisted of 5 test blocks each containing 10 questions. Following each test block, except the final one, was a learning intervention. Participants first answered the questions in a test block and then proceeded to a learning block in order to get feedback and instruction. In the learning block, they were able to revisit previously taken assessment items, and manipulate a figure with both x-axis and y-axis rotations to better visualize the operations needed to answer assessment items.

Each item in an assessment block (Fig. 24.1) featured a reference object (i.e., the one in the question stem) that had undergone a rotation. Subjects then considered a new object and attempted to determine which of five options corresponded to the same rotation as the reference object. All items included either an x-axis or y-axis rotation, or both, and varied in complexity. In the learning block, two types of learning interventions were developed. In the first type (Fig. 24.2), for each item in the learning block, a graphical box was provided that allowed the participant to use a left-to-right or an up-and-down bar to rotate the 3D object in the question along either the horizontal or vertical axis into the correct position. This opportunity

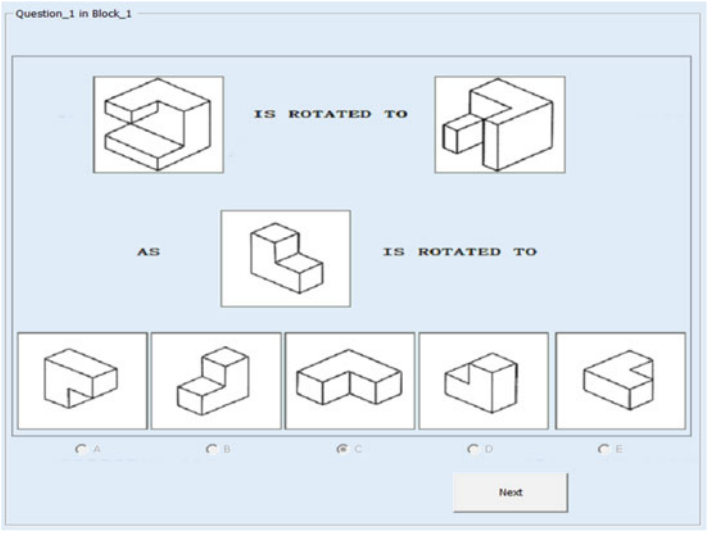


Fig. 24.1 Test Block

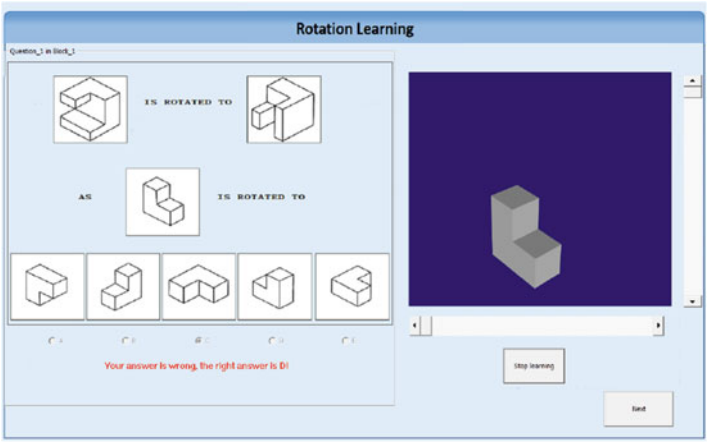


Fig. 24.2 Learning Block: Type 1

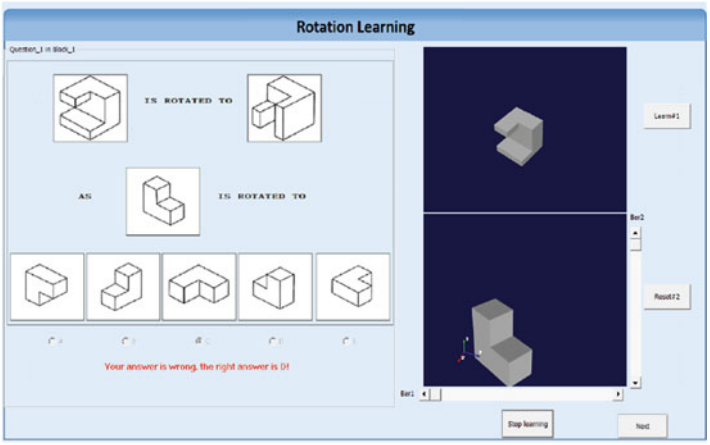


Fig. 24.3 Learning Block: Type 2

to interact and manipulate the objects was meant to promote learning the spatial rotation skills. In the second type (Fig. 24.3), an additional short clip was provided to show the participants how the reference items are rotated to the correct position. Four spatial rotation skills were identified: (1) 90° x-axis rotation, (2) 90° y-axis rotation, (3) 180° x-axis rotation and (4) 180° y-axis rotation. For the 90° rotations, clockwise and counter-clockwise rotations are treated indifferently.

An unresolved question in fitting dynamic CDMs is whether item parameter estimates can be affected by item position. This concern stems from the notion of attributes being mastered throughout the assessment, and if learning would then be confounded with the perceived difficulty of the item. In order to account for this possible source of nonidentifiability, five different versions of the assessment

were constructed and subjects were randomly assigned to them. This was done by rotating the position of each of the five blocks so that the position of each item in the assessment was balanced in the data set. Simulation studies (e.g., Wang et al., 2016; Chen et al., 2017) indicated that this ensured that item parameters could be identified and there was little bias in parameter estimates. A total of 351 University of Illinois students participated in this experiment, with 177 experiencing the first type of learning intervention and 174 receiving the second type of learning intervention.

24.2 Model Description

Our proposed model can be regarded as the combination of two parts, a learning model that describes the transition of attribute patterns over time, and a measurement model that assesses learners' attribute mastery at each time point. The section below provides an overview for these two components.

24.2.1 Learning Model

We denote the attribute pattern for subject $n \in \{1, \ldots, N\}$ at time $t \in \{0, \ldots, T\}$ by $\boldsymbol{\alpha}_{n,t} = (\alpha_{nt1}, \ldots, \alpha_{ntD})'$. Here, $t = 0$ represents the initial time point before any learning takes place, and $t = 1, \ldots, T$ represent each subsequent time point. Thus, the total number of time points is $T + 1$. Two transition models are considered below, a general monotonic Markov model and a restricted model with attribute hierarchies. We note that both transition models we consider here impose the Markovian assumption on the transitions of the attribute patterns over time, that is, given the attribute pattern at the current time point (t), the learner's attribute pattern at the next time point ($t+1$) is independent from the attribute patterns at all previous time points ($0, \ldots, t-1$).

Monotonic Markov model Given the attribute pattern of subject n at time t, $\boldsymbol{\alpha}_{n,t} = (\alpha_{n1}, \ldots, \alpha_{nD})'$, the probability that s/he masters attribute $d \in \{1, \ldots, D\}$ at time $t+1$ is given by

$$P(A_{n,t+1,d} = 1 \mid \boldsymbol{\alpha}_{n,t}) = \begin{cases} 1, & \alpha_{n,t,d} = 1; \\ \tau_d, & \alpha_{n,t,d} = 0, \end{cases} \tag{24.1}$$

where A is the random variable for attribute mastery with realization α, and τ_d is the probability of transitioning from non-master to master on attribute d at any given time point. In addition, we assume that the transitions on each attribute are independent conditioning on the attribute pattern at the current time point. If we denote the observed mastery status on attribute d at time $t+1$ by $\alpha_{n,t+1,d}$ we have

$$
\begin{aligned}
P(A_{n,t+1,d} &= \alpha_{n,t+1,d} \mid \boldsymbol{\alpha}_{n,t,d}, \tau_d) \\
&= P(A_{n,t+1,d} = 1 \mid \boldsymbol{\alpha}_{n,t}, \tau_d)^{\alpha_{n,t+1,d}} \\
&\quad \times [1 - P(A_{n,t+1,d} = 1 \mid \boldsymbol{\alpha}_{n,t}, \tau_d)]^{1-\alpha_{n,t+1,d}}. \qquad (24.2)
\end{aligned}
$$

The probability of attribute pattern $\boldsymbol{\alpha}_{n,t+1}$, given $\boldsymbol{\alpha}_{n,t}$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_D)'$, is

$$
P(\mathbf{A}_{n,t+1} = \boldsymbol{\alpha}_{n,t+1} \mid \boldsymbol{\alpha}_{n,t}, \boldsymbol{\tau}) = \prod_{d=1}^{D} P(A_{n,t+1,d} = \alpha_{n,t+1,d} \mid \alpha_{n,t,d}, \tau_d). \qquad (24.3)
$$

Restricted model with attribute hierarchies The monotonic Markov model above assumes that the probability of learning an attribute does not depend on whether another attribute is also learned. In practice, however, some attributes may be prerequisite to others, and as a result, we may have a restricted set of possible attribute patterns. This will limit the number of possible attribute patterns from all 2^D possibilities. Denote the set of prerequisites to attribute d as $\{\bar{d}\}$, then instead of Eq. (24.1), the probability of the transition is given by

$$
P(A_{n,t+1,d} = 1 \mid \boldsymbol{\alpha}_{n,t}, \tau_d) = \begin{cases} \prod_{d' \in \{\bar{d}\}} \alpha_{n,t+1,d'}, & \alpha_{n,t,d} = 1; \\ \tau_d \cdot \prod_{d' \in \{\bar{d}\}} \alpha_{n,t+1,d'}, & \alpha_{n,t,d} = 0. \end{cases} \qquad (24.4)
$$

Intuitively, the probability of transitioning from 0 to 1 on skill d is still τ, provided that the prerequisites to d are all mastered. If at time $t + 1$, any of the prerequisites to d are missing, the probability that $\alpha_{n,t+1,d} = 1$ will be 0. The hierarchical relationship between attributes can be captured by using a $D \times D$ reachability matrix, $\boldsymbol{R}$ (e.g., Leighton, Gierl, & Hunka, 2004), where $R_{dd'} = 1$ if attribute d requires attribute d' as a prerequisite, and $R_{dd'} = 0$ otherwise.

24.2.2 Measurement Models

At each time point, the learners' attribute mastery status ($\boldsymbol{\alpha}_{n,t}$) cannot be directly observed, and hence they need to be measured through assessment items. We therefore have a hidden Markov model, where the underlying transitions are modelled in a Markovian manner, but the states (i.e., attribute patterns) of the learners are latent, or hidden. To model the relationship between the observed item responses and the underlying attribute patterns at each time point, here we consider two possible response models, the noisy input, deterministic-"and"-gate model (NIDA; Junker & Sijtsma, 2001) and the reduced reparameterized unified model (rRUM; Hartz, 2002; Leighton & Gierl, 2007; DiBello, Stout, & Roussos, 1995). At a certain time point, we index the items by $k = 1, \ldots, K_t$, where K_t is

the number of items administered at time t. Under the NIDA model, the probability of a correct response is given by

$$P(X_{n,t,k} = 1 \mid \boldsymbol{\alpha}_{n,t}, \mathbf{s}, \mathbf{g}) = \prod_{d=1}^{D} [(1 - s_d)^{\alpha_{n,t,d}} g_d^{(1-\alpha_{n,t,d})}]^{q_{k,d}}, \tag{24.5}$$

where $\mathbf{s} = [s_1, \ldots, s_D]'$, $\mathbf{g} = [g_1, \ldots, g_D]'$ can be interpreted as the probabilities of incorrectly applying an acquired attribute (slipping) and probabilities of correctly applying an unacquired attribute (guessing), respectively, and $q_{k,d} = 1$ if attribute d is required by item k, $q_{k,d} = 0$ otherwise. Intuitively, on an item following the NIDA model, the learner has a probability of correctly/incorrectly applying each skill. Given that all required skills by the item are correctly applied, the probability of correct response is 1, 0 otherwise.

By relaxing the slipping and guessing parameters to be item-specific ($\mathbf{s}_k$, $\mathbf{g}_k$), we have the Generalized NIDA (GNIDA) model, given by

$$P(X_{n,t,k} = 1 \mid \boldsymbol{\alpha}_{n,t}, \mathbf{s}_k, \mathbf{g}_k) = \prod_{d=1}^{D} [(1 - s_{k,d})^{\alpha_{n,t,d}} g_{k,d}^{(1-\alpha_{n,t,d})}]^{q_{k,d}}. \tag{24.6}$$

Hartz (2002) reparameterized the GNIDA model through the following conversions,

$$\pi_k^* = \prod_{d=1}^{D} (1 - s_{k,d})^{q_{k,d}}, \tag{24.7}$$

$$r_{k,d}^* = \frac{g_{k,d}}{1 - s_{k,d}}, \tag{24.8}$$

and obtained the rRUM model, where the probability of correct response is given by

$$P(X_{n,t,k} = 1 \mid \boldsymbol{\alpha}_{n,t}, \mathbf{q}_k, \pi_k^*, \mathbf{r}_k) = \pi_k^* \prod_{d=1}^{D} r_{k,d}^{*(1-\alpha_{n,t,d})q_{k,d}}. \tag{24.9}$$

Intuitively, the probability of correct response for someone who has mastered all requisite skills to an item k, under the rRUM, is π_k^*. Missing each requisite skill d to item k results in a discount in the probability of correct response at $r_{k,d}^*$.

24.3 Parameter Estimation

A Bayesian formulation is adopted to estimate the learning model's parameters. Similar to Culpepper and Hudson (2017), a data augmentation approach is used for updating the generalized NIDA (NIDA and rRUM) model parameters, and similar

to Wang et al. (2016) and Chen et al. (2017), the forward-backward algorithm was used for sequentially updating the learning model and the attribute parameters under the hidden Markov model. The basic idea behind the forward-backward algorithm is that when we estimate the $\boldsymbol{\alpha}_n$s, we sequentially update the attribute pattern at each time point. For a time point $0 < t < T$, the distribution of $\boldsymbol{\alpha}_{n,t}$ depends on the previous time point ($\boldsymbol{\alpha}_{n,t-1}$) and the next time point ($\boldsymbol{\alpha}_{n,t+1}$).

24.3.1 Prior Distribution

We assume that the prior distribution for the initial population membership probabilities, $\boldsymbol{\Pi}$, is

$$\boldsymbol{\Pi} \sim Dirichlet(\boldsymbol{\delta}_0), \text{ where } \boldsymbol{\delta}_0 = (\delta_{01}, \ldots, \delta_{0C})', \text{ with } C = 2^D. \tag{24.10}$$

We further assume the prior distributions for the transition probabilities $\boldsymbol{T}$, are

$$p(\boldsymbol{T} = \boldsymbol{\tau}) \propto \prod_{d=1}^{D} \tau_d^{a-1}(1-\tau_d)^{b-1}. \tag{24.11}$$

In addition, for both the NIDA and the rRUM model, truncated Beta priors were used for $S_{k,d}$s and $G_{k,d}$s, the slipping and guessing parameters under the G-NIDA formulation (note that under the NIDA model, the $S_{k,d}$s and $G_{k,d}$s were constrained to be equal across items and hence could be simplified to s_d and g_d):

$$p(s_{k,d}, g_{k,d}) \propto s_{k,d}^{a_s-1}(1-s_{k,d})^{b_s-1} g_{k,d}^{a_g-1}(1-g_{k,d})^{b_g-1}\mathcal{I}(0 \le g_{k,d} < 1-s_{k,d} \le 1). \tag{24.12}$$

24.3.2 Full Conditional Distributions

Let $\mathbf{Z}_{n,t,k,\cdot} = (Z_{n,t,k,1}, \ldots, Z_{n,t,k,D})'$ denote the augmented latent responses to item k by subject n at time t, where $Z_{n,t,k,d} = 1$ if subject n has successfully applied attribute d on item k at time t, and $Z_{n,t,k,d} = 0$ otherwise. In addition, let $\mathbf{Z}_{n,t,k,(d)}$ denote the vector of the latent responses on item k by subject n at time t except on attribute d. With the assumed prior distributions of the parameters described above, the full conditional distributions for the parameters, given the observed responses $x_{n,t,k}$s, are described below.

- For $Z_{n,t,k,d}$s such that the corresponding $q_{k,d} = 1$:

$$Z_{n,t,k,d} \mid (X_{n,t,k} = x_{n,t,k}, \mathbf{Z}_{n,t,k,(d)}, \alpha_{n,t,d}, s_{k,d}, g_{k,d}) \sim \text{Bernoulli}(\tilde{\pi}_{n,t,k,d}), \tag{24.13}$$

where

$$\begin{aligned}
&\tilde{\pi}_{n,t,k,d} \\
&= \frac{P(x_{n,t,k} \mid \mathbf{Z}_{n,t,k,(d)}, Z_{n,t,k,d} = 1) P(Z_{n,t,k,d} = 1 \mid \alpha_{n,t,d}, s_{k,d}, g_{k,d})}{\sum_{z_{n,t,k,d}=0}^{1} P(x_{n,t,k} \mid \mathbf{Z}_{n,t,k,(d)}, Z_{n,t,k,d} = z_{n,t,k,d}) P(Z_{n,t,k,d} = z_{n,t,k,d} \mid \alpha_{n,t,d}, s_{k,d}, g_{k,d})} \\
&= \{(1 - \prod_{d' \neq d} z_{n,t,k,d'}^{q_{k,d'}})[(1 - s_{k,d})^{\alpha_{n,t,d}} g_{k,d}^{1-\alpha_{n,t,d}}]^{q_{k,d}}\}^{1-x_{n,t,k}}.
\end{aligned} \tag{24.14}$$

- For $A_{n,t,d}$'s: Let $\mathbf{s}_{\cdot d}, \mathbf{g}_{\cdot d}$ denote the vectors of slipping and guessing parameters associated with applying attribute d for all items administered to subject n at time t, and let $\boldsymbol{\alpha}^*$ denote the attribute vector of length D, whose dth entry is $\alpha_{n,t,d}$ and the other entries are equal to $\boldsymbol{\alpha}_{n,t,(d)}$. Then

$$\begin{aligned}
&p(A_{n,t,d} = \alpha_{n,t,d} \mid \mathbf{z}_{n,t,\cdot,d}, \boldsymbol{s}_{\cdot,d}, \boldsymbol{g}_{\cdot,d}, \boldsymbol{\alpha}_{n,t,(d)}) \\
&\qquad \propto p(\mathbf{z}_{n,t,\cdot,d} \mid \alpha_{n,t,d}, \boldsymbol{s}_{\cdot,d}, \boldsymbol{g}_{\cdot,d}) \tilde{\pi}_{n,t,d} \qquad (24.15) \\
&\qquad \propto [\prod_{k=1}^{K_t} P(z_{n,t,k,d} \mid \alpha_{n,t,d}, s_{k,d}, g_{k,d})] \tilde{\pi}_{n,t,d}, \qquad (24.16)
\end{aligned}$$

where

$$\tilde{\pi}_{n,t,d} = \begin{cases} P(\boldsymbol{A}_{n,t} = \boldsymbol{\alpha}^* \mid \boldsymbol{\pi}) P(\boldsymbol{\alpha}_{n,t+1} \mid \boldsymbol{A}_{n,t} = \boldsymbol{\alpha}^*, \boldsymbol{\tau}), & t = 0; \\ P(\boldsymbol{A}_{n,t} = \boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}_{n,t-1}, \boldsymbol{\tau}) P(\boldsymbol{\alpha}_{n,t+1} \mid \boldsymbol{A}_{n,t} = \boldsymbol{\alpha}^*, \boldsymbol{\tau}), & 1 \leq t < T; \\ P(\boldsymbol{A}_{n,t} = \boldsymbol{\alpha}^* \mid \boldsymbol{\alpha}_{n,t-1}, \boldsymbol{\tau}), & t = T, \text{ and} \end{cases} \tag{24.17}$$

$$\begin{aligned}
&P(z_{n,t,k,d} \mid \alpha_{n,t,d}, s_{k,d}, g_{k,d}) \\
&= [(1 - s_{k,d})^{\alpha_{n,t,d}} g_{k,d}^{(1-\alpha_{n,t,d})}]^{z_{n,t,k,d}} [s_{k,d}^{\alpha_{n,t,d}} (1 - g_{k,d})^{1-\alpha_{n,t,d}}]^{1-z_{n,t,k,d}}.
\end{aligned} \tag{24.18}$$

- For $\boldsymbol{\Pi}$: Denote the attribute patterns of all subjects at time $t = 0$ by $\boldsymbol{\alpha}_{\cdot,0}$, then

$$\begin{aligned}
&\boldsymbol{\Pi} \mid \boldsymbol{\alpha}_{\cdot,0} \sim Dirichlet(\boldsymbol{\delta}_0 + \tilde{\boldsymbol{\delta}}), \text{ where } \tilde{\boldsymbol{\delta}} \\
&= (\sum_{n=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_{n,0} = \boldsymbol{\alpha}_1), \ldots, \sum_{n=1}^{N} \mathcal{I}(\boldsymbol{\alpha}_{n,0} = \boldsymbol{\alpha}_C))'.
\end{aligned} \tag{24.19}$$

- For $\boldsymbol{S}_k, \boldsymbol{G}_k$'s: Given an item k, let t_n^* denote the time at which item k was administered to subject n, and let $\boldsymbol{\alpha}_{\cdot,\cdot}$ represent the attribute patterns for all

subjects across all time points. Then

$$P(s_{k,d}, g_{k,d} \mid \mathbf{z}_{\cdot,\cdot,\cdot,\cdot}, \boldsymbol{\alpha}_{\cdot,\cdot}) \propto s_{k,d}^{a_{s,k,d}-1}(1-s_{k,d})^{b_{s,k,d}-1} g_{k,d}^{a_{g,k,d}-1} \times (1-g_{k,d})^{b_{g,k,d}-1} \times \mathcal{I}(0 \leq g_{k,d} < 1 - s_{k,d} \leq 1), \tag{24.20}$$

where, under the rRUM model,

$$a_{s,k,d} = \sum_{n=1}^{N} \alpha_{n,t_n^*,d}(1 - z_{n,t_n^*,k,d})q_{k,d} + a_s; \tag{24.21}$$

$$b_{s,k,d} = \sum_{n=1}^{N} \alpha_{n,t_n^*,d} z_{n,t_n^*,k,d} q_{k,d} + b_s; \tag{24.22}$$

$$a_{g,k,d} = \sum_{n=1}^{N} (1 - \alpha_{n,t_n^*,d}) z_{n,t_n^*,k,d} q_{k,d} + a_g; \tag{24.23}$$

$$b_{g,k,d} = \sum_{n=1}^{N} (1 - \alpha_{n,t_n^*,d})(1 - z_{nt_n^*,k,d}) q_{k,d} + b_g. \tag{24.24}$$

Under the NIDA model, the $s_{k,d}$ and $g_{k,d}$s are the same across items, thus $a_{s,k,d}, b_{s,k,d}, a_{g,k,d}$, and $b_{g,k,d}$ can be simplified to $a_{s,d}, b_{s,d}, a_{g,d}$, and $b_{g,d}$, where

$$a_{s,d} = \sum_{n=1}^{N} \sum_{k=1}^{K_{t_n^*}} \alpha_{n,t_n^*,d}(1 - z_{n,t_n^*,k,d})q_{k,d} + a_s; \tag{24.25}$$

$$b_{s,d} = \sum_{n=1}^{N} \sum_{k=1}^{K_{t_n^*}} \alpha_{n,t_n^*,d} z_{n,t_n^*,d} q_{k,d} + b_s; \tag{24.26}$$

$$a_{g,d} = \sum_{n=1}^{N} \sum_{k=1}^{K_{t_n^*}} (1 - \alpha_{n,t_n^*,d}) z_{n,t_n^*,k,d} q_{k,d} + a_g; \tag{24.27}$$

$$b_{g,d} = \sum_{n=1}^{N} \sum_{k=1}^{K_{t_n^*}} (1 - \alpha_{n,t_n^*,d})(1 - z_{n,t_n^*,k,d}) q_{k,d} + b_g. \tag{24.28}$$

- For $\boldsymbol{T}$: Let $\boldsymbol{\alpha}_{\cdot,t}$ denote the attribute patterns for all subjects at time t, and let $\{\bar{d}\}$ denote the set of prerequisites to attribute d. Then

$$P(\boldsymbol{T} = \boldsymbol{\tau} \mid \boldsymbol{\alpha}_{\cdot,t}, \boldsymbol{\alpha}_{\cdot,t+1}) \propto \prod_{d=1}^{D} \tau_d^{a_{\tau_d}-1}(1-\tau_d)^{b_{\tau_d}-1}, \text{ with} \tag{24.29}$$

$$a_{\tau_d} = \sum_{t=0}^{T-1}\sum_{n=1}^{N}\left\{(1-\alpha_{n,t,d})\alpha_{n,t+1,d}\prod_{d'\in\{\bar{d}\}}\alpha_{n,t+1,d'}\right\} + a; \tag{24.30}$$

$$b_{\tau d} = \sum_{t=0}^{T-1}\sum_{n=1}^{N}\left\{(1-\alpha_{n,t,d})(1-\alpha_{n,t+1,d})\prod_{d'\in\{\bar{d}\}}\alpha_{n,t+1,d'}\right\} + b. \tag{24.31}$$

24.3.3 A Gibbs Sampling Algorithm

We developed a Markov Chain Monte Carlo (MCMC) algorithm to sample the parameters from the posterior distribution, by iteratively sampling each parameter from its corresponding conditional distribution given the other parameters. Because the conditional distributions of all parameters can be directly sampled from, we can use a Gibbs sampler to iteratively draw samples of the parameters from the full conditional distributions. More specifically, the parameters were updated following these steps:

1. Assign initial values to all parameters, namely $\boldsymbol{\pi}^{[0]}$, $\boldsymbol{\alpha}^{[0]}$, $\boldsymbol{s}^{[0]}$, $\boldsymbol{g}^{[0]}$, $\boldsymbol{\tau}^{[0]}$, and $\boldsymbol{z}^{[0]}$.
2. At each iteration r:
   (a) For each n, k, t, and d where $q_{k,d} = 1$, draw $z_{n,t,k,d}^{[r+1]}$ based on Eqs. (24.13), and (24.14), given $x_{n,t,k}$, $\mathbf{z}_{n,t,k,(d)}^{[r]}$, $\alpha_{n,t,d}^{[r]}$, $s_{k,d}^{[r]}$, and $g_{k,d}^{[r]}$;
   (b) For each n, t, and d, draw $\alpha_{n,t,d}^{[r+1]}$ based on Eqs. (24.15), (24.16), (24.17), and (24.18), given $\mathbf{z}_{n,t,\cdot,d}^{[r+1]}$, $\boldsymbol{s}_{\cdot,d}^{[r]}$, $\boldsymbol{g}_{\cdot,d}^{[r]}$, $\boldsymbol{\alpha}_{n,t-1}^{[r+1]}$, $\boldsymbol{\alpha}_{n,t+1}^{[r]}$, $\boldsymbol{\pi}^{[r]}$, and $\boldsymbol{\tau}^{[r]}$;
   (c) Draw $\boldsymbol{\pi}^{[r+1]}$ based on Eq. (24.19), given $\boldsymbol{\alpha}_{\cdot,0}^{[r+1]}$;
   (d) For the rRUM model, for each item k and attribute d, draw $g_{k,d}^{[r+1]}$ based on Eqs. (24.20), (24.21), (24.22), (24.23), and (24.24), given $\mathbf{z}_{\cdot,\cdot,k,d}^{[r+1]}$, $\boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}$ and $s_{k,d}^{[r]}$. Then, draw $s_{k,d}^{[r+1]}$ based on Eqs. (24.20), (24.21), (24.22), (24.23), and (24.24) given $\mathbf{z}_{\cdot,\cdot,k,d}^{[r+1]}$, $\boldsymbol{\alpha}_{\cdot\cdot}^{[r+1]}$ and $g_{jk}^{[r+1]}$. The corresponding $\pi_k^{*[r+1]}$ and $\boldsymbol{r}_k^{*[r+1]}$ can be obtained via algebraic transformations in Eqs. (24.7) and (24.8). For the NIDA model, for each attribute d, draw $g_d^{[r+1]}$ based on Eqs. (24.20), (24.25), (24.26), (24.27), and (24.28), given $\mathbf{z}_{\cdot,\cdot,\cdot,d}^{[r+1]}$, $\boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}$, and $s_d^{[r]}$, and draw $s_d^{[r+1]}$ based on Eqs. (24.20), (24.25), (24.26), (24.27), and (24.28) given $\mathbf{z}_{\cdot,\cdot,\cdot,d}^{[r+1]}$, $\boldsymbol{\alpha}_{\cdot,\cdot}^{[r+1]}$, and $g_d^{[r+1]}$.

(e) For each d, sample $\tau_d^{[r+1]}$ from the conditional distribution in Eqs. (24.29), (24.30), and (24.31), given $\boldsymbol{\alpha}_{i,\cdot}^{[r+1]}$, $[\tau_1, \ldots, \tau_{d-1}]'^{[r+1]}$, and $[\tau_{d+1}, \ldots, \tau_D]'^{[r+1]}$.

24.4 Application: A Spacial Reasoning Test with Learning Interventions

Six different models, with two types of measurement models (NIDA or rRUM) and three types of attribute relationships, were compared in terms of fit to the spatial rotation learning data set (Wang et al., 2016; Chen et al., 2017). The three different types of attribute relationships were captured by the corresponding reachability matrices. Specifically,

- Relationship 1: No attribute hierarchies exist, thus

$$R_1 = \begin{pmatrix} 0 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{pmatrix}. \tag{24.32}$$

- Relationship 2: 180° rotation along the y-axis requires 90° rotation along the y-axis as a prerequisite, and 180° rotation along the x-axis requires 90° rotation along the x-axis as a prerequisite. The reachability matrix is hence

$$R_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{24.33}$$

- Relationship 3: The 180° rotations (along x-axis and y-axis) has **both** the 90° rotation along x-axis and the 90° rotation along y-axis as prerequisites. The corresponding reachability matrix is

$$R_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}. \tag{24.34}$$

Note that when R_2 or R_3 are used to define the hierarchical relationships between skills, some of the attribute patterns will not be realizable under the hierarchical assumptions. For example, when R_2 is used, attribute pattern $(0, 0, 1, 0)$ would not be possible, because by assumption, one cannot master 180° $x-$axis rotation

without mastering 90° x-axis rotation. Therefore, at every time point (including $t = 0$), we need to restrict the parameter space for attribute patterns to the realizable ones based on the attribute hierarchies.

The MCMC algorithm described in the section above was applied to estimate the model parameters. Uninformative (i.e., uniform) priors were chosen for $\boldsymbol{\Pi}$, $\boldsymbol{S}$, $\boldsymbol{G}$, and $\boldsymbol{T}$. The initial value of $\boldsymbol{\Pi}$ was randomly sampled from Dirichlet($\mathbf{1}$), the initial values for each τ_d were sampled from the Uniform(0,1) distribution, and the initial values for the $s_{k,d}, g_{k,d}$'s were randomly sampled from $U(.1, .3)$. Using these random initial values, the initial values for $\boldsymbol{\alpha}$'s were simulated. Lastly, $z^{[0]}_{n,t,k,d}$ was set to 1 for all n, t, k and d.

24.4.1 MCMC Convergence

To evaluate the parameter convergence using the MCMC algorithm, five separate chains with different starting values were run with chain lengths of 50,000 iterations under the rRUM model with no attribute hierarchies. The Gelman-Rubin proportional scale reduction factor (PSRF), commonly known as $\hat{R}$, was calculated for each parameter at different chain lengths. The progression of the maximum $\hat{R}$ out of all estimated parameters is displayed in Fig. 24.4.

The Gelman-Rubin $\hat{R}$ compares the within-chain variance and the between-chain variance of the parameter samples. If the chains have mixed well, then $\hat{R}$, the ratio between the pooled (between and within) variance and the within-chain variance,

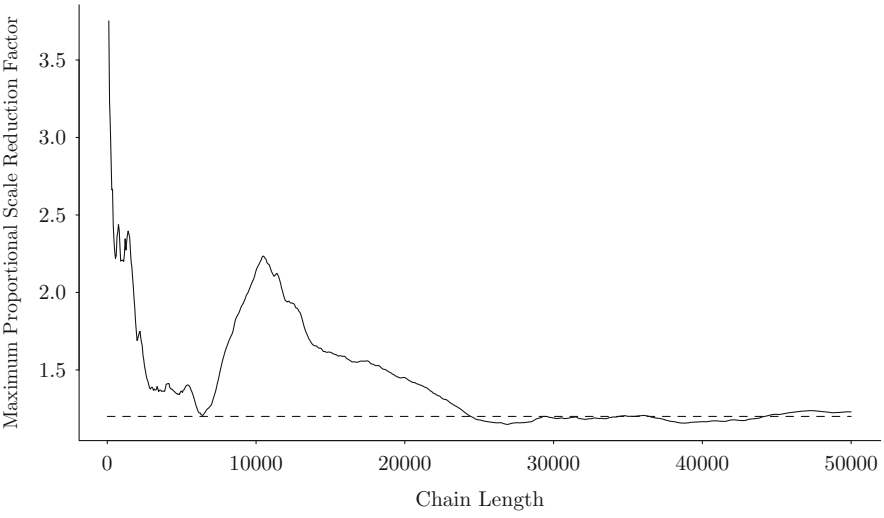


Fig. 24.4 Progression of maximum $\hat{R}$ as chain length increases. Dashed horizontal line indicates $\hat{R} = 1.2$

should be close to 1. An $\hat{R}$ value below 1.2 is commonly used to indicate the convergence of that parameter estimate. We can see that at around 25,000 iterations, the maximum $\hat{R}$ stabilizes to less than or slightly above (up to 0.02 above) 1.2, and the $\hat{R}$ of all the other estimated parameters stay below 1.2. Thus for subsequent analyses, a chain length of 40,000 was used for each of the 6 models, with 25,000 iterations as the burn-in.

24.4.2 Model Comparison

The fit of the six models were compared in terms of a few aspects, the Deviance Information Criterion (DIC), and posterior predictive model checks on the item means (first moments, M1), item pair-wise odds ratios (second moments, M2), and on the subjects' total scores across time points. The procedures for computing each are detailed below.

- DIC: DIC is commonly used to assess the relative global fit of the model, and intuitively, it is related to the likelihood of the observed data given the estimated model parameters. As described in Spiegelhalter, Best, Carlin, and van der Linde (2002), if we denote the set of unknown model parameters by θ, then the DIC can be calculated as

$$DIC = p_D + \bar{D}(\theta), \tag{24.35}$$

where $p_D = \bar{D}(\theta) - D(\bar{\theta})$, and

$$D(\theta) = -2\log[P(\mathbf{x} \mid \theta)] + C \tag{24.36}$$

$$= -2\log[P(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{s}, \boldsymbol{g})] + C \tag{24.37}$$

$$= -2\log\left\{\prod_{n=1}^{N}\Big[\sum_{\forall \boldsymbol{\alpha}_l \in \mathscr{A}^{T+1}} P(\boldsymbol{\alpha}_l)\prod_{t=0}^{T} P(\boldsymbol{x}_{n,t} \mid \boldsymbol{A}_{n,t} = \boldsymbol{\alpha}_{l,t}, \boldsymbol{s}_t, \boldsymbol{g}_t)\Big]\right\} + C. \tag{24.38}$$

Here, $\boldsymbol{\alpha}_l = (\boldsymbol{\alpha}_{l,0}, \boldsymbol{\alpha}_{l,1}, \ldots, \boldsymbol{\alpha}_{l,T})$ is any learning trajectory from time $t = 0$ to T, and $P(\boldsymbol{\alpha}_l)$ can be computed as

$$P(\boldsymbol{\alpha}_l) = P(\boldsymbol{A}_{n,0} = \boldsymbol{\alpha}_l \mid \boldsymbol{\pi})\prod_{t=1}^{T} P(\boldsymbol{A}_{n,t} = \boldsymbol{\alpha}_{l,t} \mid \boldsymbol{A}_{n,t-1} = \boldsymbol{\alpha}_{l,t-1}, \boldsymbol{\tau}). \tag{24.39}$$

To calculate $\bar{D}(\theta)$, at each post-burnin iteration t of the MCMC, we compute the $D(\theta^{[t]})$ based on the parameter samples in the tth iteration, namely $\theta^{[t]}$. The average of $D(\theta^{[t]})$s across all post-burnin iterations is computed to obtain $\bar{D}(\theta)$.

- Posterior predictive check for the item means (M1): Posterior predictive model checking (PPMC) is commonly used in Bayesian modeling to assess the local (e.g., item-level or person-level) fit of the model to the observed data. Intuitively, it assesses the position of the observed data in the posterior predictive distribution of the model. After burnin, at each iteration of the MCMC, the model parameter samples were used to simulate responses of the subjects. For each item, the item means (M1), that is the proportion of people who answered correctly was calculated based on the simulated responses as well as the observed data. Then the posterior predictive probability (PPP) of each item's mean is given by the proportion of simulated item means that lie below the observed item mean.
- Posterior predictive check for the item pairwise odds ratios (M2): For any given item pair, Sinharay, Johnson, and Stern (2006) suggested using $OR = (N_{11}N_{00})/(N_{01}N_{10})$, where N_{11} is number of respondents responding to both items correctly, N_{01} is number of respondents who answered item 1 wrong and item 2 correctly, etc., as a measure of item-pairwise associations. Similar to that of the item means, the item pair-wise ORs based on the simulated responses from sampled model parameters and those from the observed responses are obtained, and the PPP of each item pair's odds ratio is given by the proportion of simulated odds ratios for the item pair that lie below the observed.
- Posterior predictive model check for the subjects' total scores at each time point: Like above, simulated and observed responses were used to obtain the number of correct responses (total score) by each subject at each time point. Then, for each subject and each time point, the PPP for total score is given by the proportion of simulated total scores below the observed.

Table 24.1 summarizes the DIC statistics and the proportions of posterior predictive probabilities below 0.05 or above 0.95 (which indicates misfit, or in the extreme range) for item means (M1), item ORs (M2), and subject total scores for the six models. A smaller DIC value and a smaller proportion of PPPs outside the 90% interval would indicate better fit.

Table 24.1 suggests that out of the six models, the one assuming a rRUM measurement model and no attribute hierarchies achieved the best fit, indicated by the lowest DIC, the lowest proportion of extreme PPPs on item means and pair-wise odds ratios, and comparable proportion of extreme PPPs as the other models. We also see that compared to models with NIDA as the measurement model,

Table 24.1 Summary of fit statistics of the six different models

Model	DIC	% M1 misfit	% M2 misfit	% total misfit
NIDA R_1	16129.61	74.0	46.4	24.1
NIDA R_2	16154.09	70.0	47.1	23.1
NIDA R_3	16233.26	72.0	48.2	23.1
rRUM R_1	14860.91	0.0	25.1	23.5
rRUM R_2	15099.09	0.0	26.4	23.6
rRUM R_3	15188.04	0.0	27.3	23.8

models using rRUM as the measurement model performed much better in terms of item level fit. However, there was not an obvious difference between the NIDA-based and rRUM-based learning models in terms of total score posterior predictive probabilities.

Figure 24.5 presents the posterior predictive probabilities of each item's mean under the rRUM model without attribute hierarchies. The shaded area in each circle represents the proportion of simulated item means below the observed item mean. None of the observed item means were within the extreme range. There is a consistent tendency for the model to slightly underestimate the item means, as indicated by the PPPs above 50% on all items.

Figure 24.6 presents the density of the posterior predictive probability of the item pair-wise odds ratios. We observe that the distribution of the PPPs is skewed to the left, indicating a tendency for the model to underestimate the ratio $(N_{00}N_{11})/(N_{10}N_{11})$.

Figure 24.7 presents the density curves of the posterior predictive probabilities for total scores at different time points. For all time points, we observe a tendency

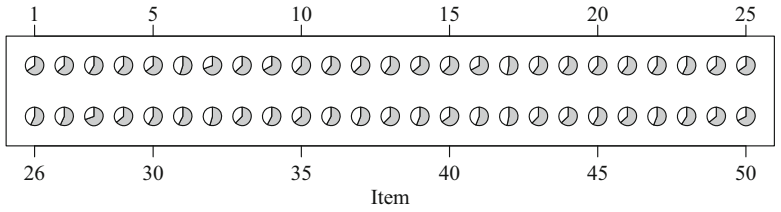


Fig. 24.5 Posterior Predictive Probabilities (PPPs) of the item means (i.e., proportion correct)

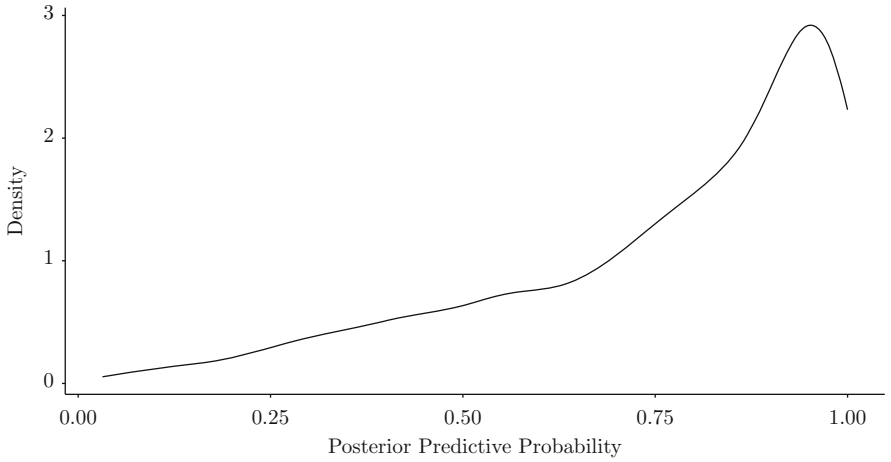


Fig. 24.6 Density of the posterior predictive probability for item pair-wise odds ratios

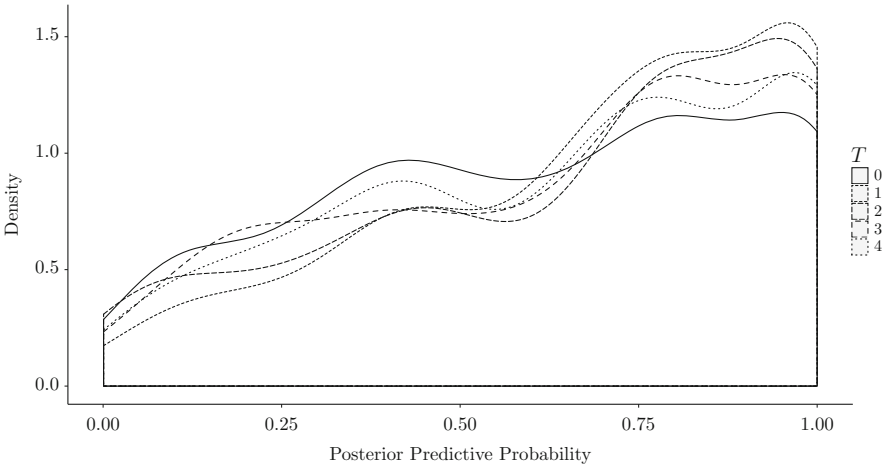


Fig. 24.7 Density of posterior predictive probability for total scores at different time points

for the model to underestimate the total scores of the subjects, as suggested by the higher densities at higher PPPs. This pattern seems to be most salient after the learning begins (i.e., for $T = 1, \ldots, 4$) than for the initial time point, $T = 0$.

The plots for PPPs of total scores over observed total score at each time point are shown in Fig. 24.8. Across all time points, we observe a consistent trend for the model to overestimate total scores for subjects with low observed scores and underestimate for those with high observed scores. The reason for this is not clear. One might suspect that it is either due to the measurement model or due to the learning transition model. We have investigated this by fitting the data separately at each timepoint with the more general GDINA model (see Chap. 7 in this volume), but the same phenomenon was observed. This suggested that perhaps it was due to the rather simple transition model that assumes independent transitions. However, this underestimation was also seen when fitting a Markov model for patternwise transitions (Chen et al., 2017). So we remain uncertain of the source of this bias. Perhaps it lies in the more elusive problem of Q-matrix misspecification, or reflects some degree of individual person misfit, such as those who were rapidly guessing. This behavior was seen in the response times of some subjects on some test blocks.

24.4.3 Observed Progression of Learning

Based on the estimated attribute patterns under the rRUM learning model without attribute hierarchies, we looked at the progression of attribute mastery rate over time, as well as the frequency for the number of mastered attributes at each time point. Table 24.2 summarizes the distribution of the number of mastered attributes

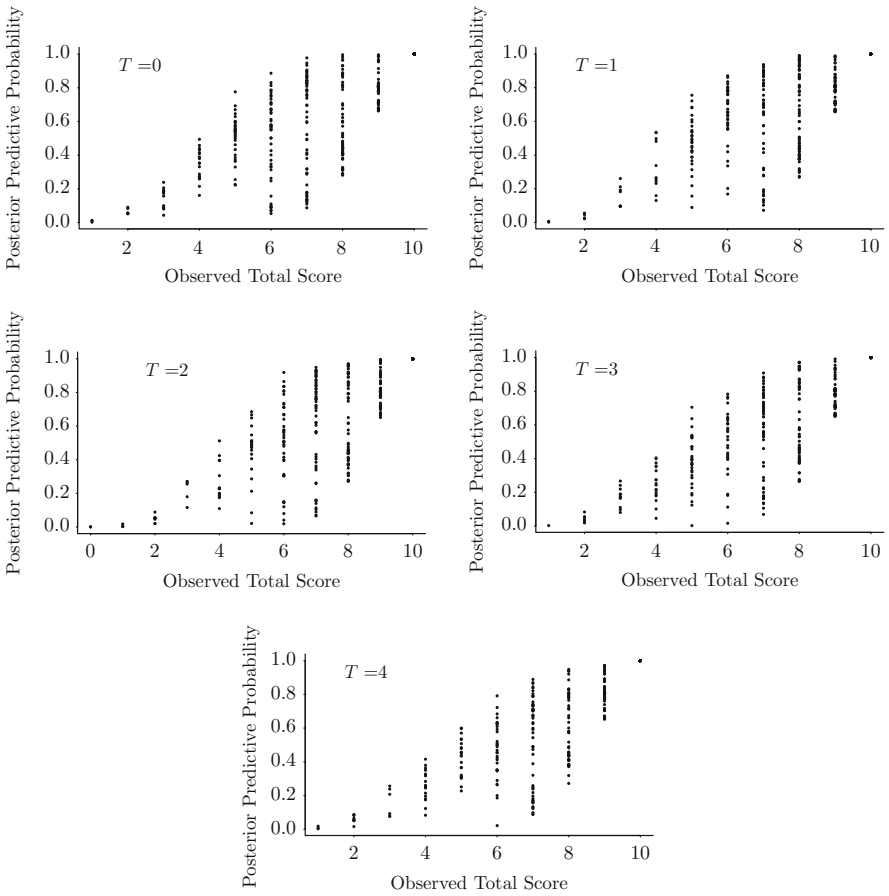


Fig. 24.8 Relationship between posterior predictive probability of total score and observed total score at each time point

at each time point, and Fig. 24.9 shows the progression of mastery rate of each attribute across time, for learners who received learning block 1 (dashed lines) and learners who received learning block type 2 (solid lines). For both types of learning interventions, as the learning time increases, the percentage of students mastering each attribute also increases, and a shift towards mastering more attributes over time is observed. Compared to individuals who received the first type of learning block, those who received the second kind consistently had slightly higher mastery rate on each skill across time points.

Table 24.2 Frequency distribution (and percentage) of number of skills mastered at each time point

Number of skills mastered	t =0	t =1	t =2	t =3	t =4
0	53(15.1%)	40(11.4%)	35(9.97%)	31(8.83%)	27(7.69%)
1	53(15.1%)	57(16.24%)	56(15.95%)	56(15.95%)	56(15.95%)
2	77(21.94%)	76(21.65%)	70(19.94%)	58(16.52%)	45(12.82%)
3	0(0%)	6(1.71%)	13(3.7%)	28(7.98%)	38(10.83%)
4	168(47.86%)	172(49%)	177(50.43%)	178(50.71%)	185(52.71%)

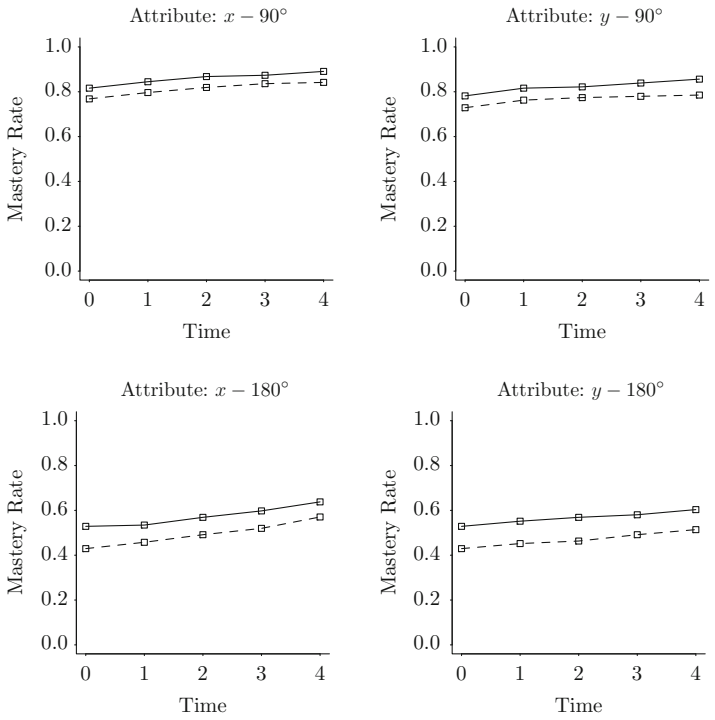


Fig. 24.9 Progression of mastery rate of each attribute across time for learners receiving the two types of treatments. Dashed line represents the mastery rate progression of learners who received the first type of learning block. Solid line represents that of the learners who received the second type of learning block

24.5 Discussion

CDMs for learning allow examining the rate at which students learn, and factors associated with learning, when using models with covariates. Learning models are also more realistic than static models for testing data collected with intermittent educational interventions, because the entire purpose of the interventions promote

learning. Even merely completing items can be a form of learning, which is the essence and purpose of the ages old practice of homework. Extensions to CDMs for learning may even include item parameters to describe the educational quality of items and how they can promote learning, for example by incorporating item-specific characteristics to the modeling of the transition probability from nonmastery to mastery.

In this chapter, we presented a model that combines the rRUM or NIDA measurement model with a simple Markov model for learning that treats the separate attributes independently. A MCMC algorithm for parameter estimation was described, and methods of assessing model fit using deviance information criteria and posterior predictive model checking were discussed. In practice, researchers can use the DIC and posterior predictive model checking to examine the fit of a proposed learning model to a data set, or to compare different models and select the one providing the best fit. For example, a researcher might be interested in which measurement model, rRUM or DINA, would be the most appropriate for the longitudinal data set at hand, and DICs and PPMC could be used to select the best fitting one. As an example of the use of posterior predictive model checking for learning model with the DINA measurement model, readers can refer to Wang et al. (2016), where the posterior predictive probabilities of item means were calculated under the high-order hidden Markov CDM.

In the application to spatial rotation data, a reasonably good but imperfect fit was seen. In particular, posterior predictive checks found the predicted total scores over time did not advance like observed scores. After some analysis, we do not believe this misfit was due to the measurement model or the learning model, but may be due to Q-matrix misspecification or person misfit. Analysis of response times indicates aberrant behavior of some subjects, which could manifest in this observed bias. The application also considered the notion of attribute hierarchies, which when present could greatly simplify the learning model. However, in this application goodness of fit measures indicated the superior fit of the unrestricted model without attribute hierarchies.

Modeling learning requires additional parameters, but this application shows that somewhat complex models can be fitted with as few as 350 subjects. Computer administered assessments are becoming more and more prevalent and will provide ample subjects for fitting learning models and even assessing item quality. Such models can be informative for item selection when the goal is promoting and verifying learning.

References

Chen, Y., Culpepper, S., Wang, S., & Douglas, J. (2017). A hidden Markov model for learning trajectories with application to spatial rotation skills. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621617721250

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Culpepper, S. A., & Hudson, A. (2017). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. *Applied Psychological Measurement*. https://doi.org/10.1177.0146621617707511

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). New York: Routledge.

González-Brenes, J., Huang, Y., & Brusilovsky, P. (2014). General features in knowledge tracing to model multiple subskills, temporal item response theory, and expert knowledge. In *The 7th International Conference on Educational Data Mining* (pp. 84–91).

González-Brenes, J. P., & Mostow, J. (2013). What and when do students learn? Fully data-driven joint estimation of cognitive and student models. In *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 236–240).

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301–321.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kaya, Y., & Leite, W. L. (2016). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling an evaluation of model performance. *Educational and Psychological Measurement*. https://doi.org/10.1177/0146621617697959

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2015). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*(2), 181–204.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics, 2*(2), 99–120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 255–266).

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298–321.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639.

Studer, C. (2012). *Incorporating learning over time into the cognitive assessment framework* (Unpublished doctoral dissertation). Carnegie Mellon University.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2016). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102.1076998617719727

Xu, Y., & Mostow, J. (2012). Comparison of methods to trace multiple subskills: Is LR-DBN best? In *Proceedings of the 5th International Conference on Educational Data Mining* (pp. 41–48).

Yoon, S. Y. (2011). *Psychometric properties of the revised Purdue Spatial Visualization Tests: Visualization of Rotations (The Revised PSVT-R)*. (Unpublished doctoral dissertation). Purdue University.

Chapter 25
How to Conduct a Study with Diagnostic Models

Young-Sun Lee and Diego A. Luna-Bazaldua

Abstract In recent years there has been a wave of new assessment designs, measurement methods, and frameworks to connect psychometrics with cognitive science due to the need to enhance traditional and new assessments in order to provide more information about the examinees and the quality of the assessment tools. The purpose of this chapter is to explore the use of a set of guidelines developed for CDM retrofitting using data from the 2007 TIMSS test administration as an example. Three research questions for the study are: Is it feasible to use a retrofitting approach using TIMSS data? Does relative model fit improve when using CDMs compared to IRT models? What additional information regarding the examinees' skills and items are gained from using CDM retrofitting?

25.1 Introduction

Traditional educational testing practice has tended to follow a normative assessment approach. This trend has been supported by the development of psychometric frameworks that inform researchers and test developers about examinees' ability and tests psychometric characteristics under a unidimensional model (Embretson & Reise, 2000). In recent years there has been a wave of new assessment designs, measurement methods, and frameworks to connect psychometrics with cognitive science due to the need to enhance traditional and new assessments in order to provide more information about the examinees and the quality of the assessment tools (Embretson & Gorin, 2001; Mislevy et al., 2014; Park & Lee, 2014; von Davier, 2009; Yan, Mislevy, & Almond, 2003). Recently, there has been growth

Y.-S. Lee (✉)
Teachers College, Columbia University, New York, NY, USA
e-mail: sly2003@columbia.edu

D. A. Luna-Bazaldua
School of Psychology, National Autonomous University of Mexico, Mexico City, Mexico
e-mail: diegobazaldua@comunidad.unam.mx

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_25

in the number of new diagnostic psychometric methods, which either expand the Classical Test Theory (CTT) or the Item Response Theory (IRT) frameworks or propose new latent variable models (Embretson & Daniel, 2008; Embretson & Yang, 2013; Magidson & Vermunt, 2001; Mislevy & Verhelst, 1990; Rupp, Templin, & Henson, 2010; Wilson, 2008; Yamamoto, 1989).

Among these methods, the different models for cognitive diagnosis stand out because of their integration of fine-grained information on skills measured by the test within a psychometric framework (Geisinger, 2012; Rupp, 2007; Rupp & Templin, 2008). From a statistical perspective, Cognitive diagnostic models (CDMs) are characterized as a confirmatory psychometric extension of the latent class model (von Davier, 2005). From an assessment perspective, CDMs are useful to provide feedback to the examinees on their strengths and weaknesses based on the mastered and non-mastered skills, rather than reporting only a single score with respect to a reference group such as those commonly obtained in the traditional CTT and IRT frameworks in psychological measurement (Crocker & Algina, 1986; Geisinger, 2012). Moreover, CDMs are intrinsically multidimensional models given their capacity to analyze fine-grained skills that interact with each other to produce a correct answer to the items in a test (Rupp & Templin, 2008). In contrast, the majority of the models developed in the CTT and IRT frameworks are unidimensional, producing a single total score that reflects an overall ability in the general domain assessed by the test (Crocker & Algina, 1986).

In some instances a researcher, instructor, or examinee may want to obtain supplementary performance feedback on specific skills measured by an already developed test but not fully reflected by the test score, as well as additional information about how the skills are related to the general domain ability measured by the test (Park & Lee, 2014). However, the differences among the traditional and new psychometric frameworks produce several challenges when trying to analyze data from standardized tests originally developed under the CTT or IRT frameworks, using new models developed under the CDM framework (Liu, Huggins-Manley, & Bulut, 2017). For instance, conventional practices in test construction are usually focused on increasing the total test score reliability by selecting items with psychometric features that are in harmony with a unidimensional construct (Liu et al., 2017). Similarly, the IRT framework has generated robust research on areas such as scale linking and equating and the development of item banks, but such processes intrinsically require unidimensionality of the construct measured by the items and tests (Embretson & Reise, 2000).

Because of the underlying differences between the traditional and new psychometric frameworks and the difficulties in analyzing an already developed test using CDM approaches, both Gierl and Cui (2008) and Liu et al. (2017) have discussed the use of *retrofitting*. In this context, retrofitting refers to the secondary data analysis process of fitting a model for cognitive diagnosis to test data from assessments originally designed under a different measurement (e.g., CTT or unidimensional IRT) framework. It is important to note that while retrofitting is presented here as an approach to analyze assessment data not originally generated under a cognitive diagnostic framework, some authors consider that retrofitting is not to a suitable

approach in psychometric data analysis (see Haberman & Davier, 2006; von Davier & Haberman, 2014).

An optimal approach to retrofitting starts by gathering comprehensive information about the test to identify subdomains or skills measured by each item, as well as any additional information about the test psychometric characteristics, the examinees, and the general purpose of the assessment. In this sense, the researcher seeks to recover multidimensional information from a test narrowed to fit unidimensional models. Once the skills have been identified, the researcher proceeds to work with test content experts in the development of a Q-matrix, which is defined as an item-by-skill matrix that specifies the skills that are required to correctly answer each item in a test (Tatsuoka, 1990). If there is disagreement or uncertainty among the experts about the definition of specific cell entries in the Q-matrix, then the researcher explores additional analyses proposed for Q-matrix validation (Chiu, 2013; de la Torre, 2008; de la Torre & Chiu, 2016; DeCarlo, 2012; Liu, Xu, & Ying, 2012). Ideally, the Q-matrix presents conditions that guarantee the identifiability or local identifiability of the conjunctive CDMs, such as having items requiring only a single skill, having skills measured by at least two or three items, and having two items with identical skill requirements for every skill defined in the Q-matrix (Xu & Zhang, 2016).

The subsequent step in the retrofitting process is focused on the CDM modeling process. First, one or more CDMs are fit to educational and psychological measurement data. The model selection process may be based on theory regarding the relationships and hierarchy among the latent skills or on information produced by the fit statistics (de la Torre & Douglas, 2004; Leighton, Gierl, & Hunka, 2004; Liu et al., 2017; Templin & Bradshaw, 2014). Some general models for cognitive diagnosis have been proposed in the literature, such as the general diagnostic model (GDM; von Davier, 2005), the log-linear cognitive diagnostic model (LCDM; Henson, Templin, & Willse, 2009), and the generalized DINA model (G-DINA; de la Torre, 2011). Specific constrained models nested within these general models correspond to models formerly proposed in the literature such the reduced reparametrized unified model (R-RUM; DiBello, Stout, & Roussos, 1995), the deterministic inputs, noisy "and" gate model (DINA; Junker & Sijtsma, 2001), and the deterministic inputs, noisy "or" gate model (NIDO; Templin & Henson, 2006). General CDMs tend to provide better model fit to data when compared to the specific models, but more parsimonious and straightforward interpretations can results from the use of specific models. Moreover, if the adequate specific model is used to fit the data, then a higher rate of correct skill patterns is estimated (de la Torre & Lee, 2013; Rojas, de la Torre, & Olea, 2012). In this sense, a retrofitting analysis could start by fitting a general model followed by its nested models, which usually differ from each other in the way the attributes interact to produce a correct answer to the items. Then, a specific model is selected depending on its fit statistics and classification of the latent attributes.

Assessing model fit and selecting an optimal model is a relevant topic given the wide variety of models that exist within the CDM framework. Liu et al. (2017) present an exhaustive review of overall model fit, item fit, and person fit indices

formerly proposed for CDM model selection. Model fit, also referred as test-level fit, is used to analyze if the selected model fits the data in their entirety (de la Torre & Lee, 2013). Model level fit measures can be further separated into absolute fit measures and relative fit measures (Chen, de la Torre, & Zhang, 2013). Absolute fit measures indicate the capacity of the model to reproduce the data. Relative fit measures compare fit of two or more models, so the model with the best relative fit statistics is chosen. For most measures, smaller values for both absolute and relative fit indices indicate better fit of the model to the data. The absolute fit indices mentioned in Chen et al. (2013) and in Liu et al. (2017) are the limited information fit statistics (Hansen, Cai, Monroe, & Li, 2016), the residuals between the observed and predicted correlations and log-odds ratios of item pairs and the residuals between the observed and predicted proportion correct of individual items (de la Torre & Douglas, 2008), the maximum of all pairwise χ^2 statistics (Chen et al., 2013; Rupp et al., 2010), and the standardized root mean square root of squared residuals (Maydeu-Olivares & Joe, 2014). Relative fit measures commonly reported in psychometric literature on CDMs are the -2 log likelihood (LL; Neyman & Pearson, 1992), Akaike Information Criterion (AIC; Akaike, 1987), the Bayesian Information Criterion (BIC; Schwarz, 1976); the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002), and the Bayes factor (Kass & Raftery, 1995).

Item level fit measures indicate if the model fits to individual items (de la Torre & Lee, 2013). Measures of item fit include the Wald test statistic (de la Torre & Lee, 2013), the mean absolute deviations between the observed and predicted item conditional probabilities of success (MAD; Henson et al., 2009), and the root mean square error of approximation between the observed and predicted item conditional probabilities of success (RMSEA; Kunina-Habenicht, Rupp, & Wilhelm, 2012; von Davier, 2005). The Wald test, developed for the G-DINA approach (de la Torre, 2011), compares item fit of a general model against a reduced nested model. The MAD and RMSEA item fit measures compare predicted and observed probabilities for different latent classes; the two indices differ in the unweighted or weighted manner in which they compare these probabilities for each latent class (Kunina-Habenicht et al., 2012). RMSEA and MAD can be averaged across items to create a test-level fit index (Lei & Li, 2016).

In the context of CDMs, person fit refers to the correspondence between an examinee's observed response pattern and her expected response pattern given her estimated skill profile. Liu, Douglas, and Henson (2009) proposed a likelihood ratio test that identifies how well the estimated skill profile describes the observed response pattern for each examinee. Lack of person fit for a considerable proportion of examinees may be due to model selection. However, person misfit could also reflect strategies used by the examinees that are not correctly explained by the model (Liu et al., 2009). Cui and Leighton (2009) have introduced the hierarchy consistency index (HCI), which ranges from -1.0 to 1.0 and identifies examinee misfit, by comparing observed and expected response patterns.

Once a model has shown good fit to the data, the last step in the retrofitting process involves the interpretation of the results (Liu et al., 2017). This final process

may involve the examination of attribute distribution in the sample of examinees, the examination of relationships between general latent traits measured by the test and the attributes estimated by the CDM, and the examination of any additional analyses (e.g., attribute correlations and hierarchies) that further inform about the constructs measured by the test.

The use of CDMs in standardized assessment remains rare compared to traditional psychometric models, despite being theoretically appealing and informative. In addition, most cases of CDM retrofitting reported in the literature have not followed guidelines as explicit as those presented in Liu et al. (2017). Nevertheless, previous research has reported cases in which CDMs have been fitted to data from tests originally constructed and fitted using other psychometric models. For instance, the Fraction Subtraction data set has been widely analyzed in CDM research (Chen & de la Torre, 2013; Chiu, 2013; Chiu & Douglas, 2013; de la Torre, 2008, 2009, 2011; de la Torre & Lee, 2010; DeCarlo, 2011), the Revised Purdue Spatial Visualization Test-Visualization of Rotations (PSVT-R; Culpepper, 2015), the Force Concept Inventory (FCI; Bradshaw & Templin, 2013), the National Assessment of Educational Progress (NAEP; Xu & von Davier, 2008, 2003) Trends in International Mathematics and Science Study mathematics test (TIMSS; Skaggs, Wilkins, & Hein, 2016), the Examination for the Certificate of Proficiency in English test (ECPE; Chiu, Douglas, & Li, 2009; Templin & Bradshaw, 2014; Templin & Hoffman, 2013, von Davier, 2014), and the Test of English as Foreign Language Internet-based testing (TOEFL-IBT; von Davier, 2005), as well as TOEFL preparation tests (Liu et al., 2017).

Moreover, despite the additional item and examinee information and gains from retrofitting test data, authors such as Gierl and Cui (2008) and Haberman and Davier (2006) warn about limitations in the use of retrofitting. First, since the original test was not constructed with the objective of diagnostic feedback, it is likely that CDM retrofitting may not produce a better fit to the data compared to other psychometric models. Moreover, it is unlikely that an appropriate CDM will be optimal to fit the data, but researchers are encouraged to test goodness of fit using different models that differ in complexity and underlying skill condensation rules. Finally, it may be the case that the test does not include an acceptable number of items to measure the latent skills, which in an extreme case may become a limitation to model identifiability (Xu & Zhang, 2016).

Considering this background, the purpose of this chapter is to explore the use of these explicit guidelines for CDM retrofitting using data from the 2007 TIMSS test administration as an example. Three research questions for the study are: Is it feasible to use a retrofitting approach using TIMSS data? Does relative model fit improve when using CDMs compared to IRT models? What additional information regarding the examinees' skills and items are gained from using CDM retrofitting?

25.2 Methods

25.2.1 Data

The models were fitted using the 25 mathematics items included in the *Booklet 4* from the 4th grade TIMSS 2007 test (Foy & Olson, 2009). The data set is composed of 960 examinees from Germany (N = 362), Iran (N = 278), and Japan (N = 320). These three countries were selected because, on average, their students reached different levels of achievement. Japan was among those countries with the highest proportion of students in the top benchmark and an average scale score of 568 points on the mathematics test. A fair proportion of students in Germany reached a high benchmark in the test with an average scale score of 525 points, while examinees from Iran tended to score on the lowest achievement benchmark with an average scale score of 402 points (Mullis et al., 2007, 2009). Each examinee answered a total of 25 mathematics items.

Following the first step for CDM retrofitting, a review of the test design and its psychometric properties indicates that TIMSS 2007 shows a median reliability across countries equal to 0.83 for the 4th grade mathematics test (Mullis et al., 2007, 2009). The TIMSS assessment scores are scaled using a three-parameter logistic IRT model (Mullis et al., 2007). The test specifications make reference to three domains of content assessed by the test: Numbers, Geometric Shapes and Measures, and Data Displays. Each domain measures specific understandings and skills as listed in Table 25.1.

Thus, the test specifications and the reliability evidence may suggest the presence of multidimensionality in the constructs measured in TIMSS. In this scenario, it seems feasible to use a retrofitting framework to gather additional diagnostic information about the examinees and psychometric information about the test.

A Q-matrix was constructed to reflect the measurement of the skills in the 25 items. Table 25.2 includes the resulting Q-matrix representing the measurement of 7 attributes. The item-by-skill structure in Table 25.2 was generated using information reported by the test developers (Mullis et al., 2007, 2009) and validated using the insight from researchers and practitioners in the field of Mathematics Education. An

Table 25.1 Domains, understandings and skills measured in the 4th grade TIMMS 2007 mathematics test

Domain	Understandings and skills
Number	Whole numbers
	Fractions and decimals
	Number sentences, patterns, and relationship
Geometric shapes and measures	Lines and angles
	Two- and three-dimensional shapes
	Location and movement
Data display	Reading, interpreting, organizing, and representing data

Table 25.2 Q-matrix of 4th grade TIMSS 2007 mathematics booklet 4

	Number			Geometric shapes and measures			Data display	
Item	α_1	α_2	α_3	α_4	α_5	α_6	α_7	Number of attributes per item
1	1	0	0	0	0	0	0	1
2	0	1	0	0	0	0	0	1
3	0	1	0	0	0	0	0	1
4	1	1	0	0	0	0	0	2
5	1	0	1	0	0	0	0	2
6	0	0	0	0	1	1	0	2
7	0	0	0	1	1	1	0	3
8	1	0	0	0	1	0	0	2
9	0	0	0	0	1	0	0	1
10	0	0	0	0	1	0	0	1
11	1	0	0	1	0	0	0	2
12	1	0	0	0	0	0	1	2
13	1	0	0	0	0	0	1	2
14	1	1	0	0	0	0	1	3
15	1	0	0	0	0	0	0	1
16	1	0	0	0	0	0	0	1
17	1	0	1	0	0	0	0	2
18	1	0	1	0	0	0	0	2
19	1	0	0	0	0	0	1	2
20	1	0	1	0	0	0	1	3
21	1	0	1	0	0	0	0	2
22	0	0	0	0	1	1	0	2
23	1	0	0	0	0	0	0	1
24	0	0	0	0	1	0	0	1
25	1	0	0	0	0	0	1	2
Total	17	4	5	2	7	3	6	46

Note: α_1 is linked to whole numbers; α_2 to fractions and decimals; α_3 to number sentences, patterns, and relationships; α_4 to lines and angles; α_5 to two- and three-dimensional shapes; α_6 to location and movement concepts; and α_7 to skills on reading, interpreting, organizing, and representing data

additional empirical validation using the approach suggested in DeCarlo (2012) was done to confirm the structure of the Q-matrix. It is relevant to highlight that while the Q-matrix was generated following the guidelines for retrofitting and with the support of experts in the domains of content, the Q-matrix lacks of some necessary conditions for full identifiability of conjunctive CDMs outlined by Xu and Zhang (2016). Specifically, although about a third of the items in the test measured only one skill, some skills were not uniquely measured by any item (e.g., the skills *number sentences, patterns, and relationships*, and *reading, interpreting, organizing, and representing data*), but they were always measured in conjunction with other skills.

25.2.2 Models

Several unidimensional and multidimensional IRT models and CDMs were fitted and compared with respect to their relative fit indices. The unidimensional logistic IRT models included the one-parameter (1-PL), two-parameter (2-PL), and three-parameter (3-PL) models (Baker & Kim, 2004). Two multidimensional IRT (MIRT; Reckase, 2009) models were also fitted, one model using item-by-latent trait structure analogous to the Q-matrix in Table 25.2, and a second MIRT model with an item-by-general domain (i.e., Number, Geometric Shapes and Measures, and Data Display) structure. For the item-by-general domain matrix, if any fine-grained skill within a domain is measured by a given item as indicated in Table 25.2, then it is assumed that the item is also measuring the general domain.

With respect to the CDMs, variations of the G-DINA model (de la Torre, 2011) were fitted. Among the general models for cognitive diagnosis, the G-DINA model is particularly flexible in terms of the link function included in the estimation of the model: identity, logit, or log link functions can be used (de la Torre, 2011). In addition, there has been extensive research on extensions of the G-DINA framework (Chen et al., 2013; de la Torre & Lee, 2013). The G-DINA model also makes it feasible to estimate simpler nested models such as the DINA model, the DINO model, and models with additive skill effects (A-CDM; de la Torre, 2011). In our analysis, a saturated G-DINA model with main effects and interactions among the latent skills was estimated, as well as nested models corresponding to the DINA model, the DINO model, and the A-CDM model. Additionally, a higher-order DINA model (HO-DINA; de la Torre & Douglas, 2004) was also fitted allowing for conditional independence among the latent skills given a higher-order continuous latent trait parameter. All models for cognitive diagnosis were estimated using a logit link function. The reader can review more information on the G-DINA model and its specific nested models in the corresponding chapter (de la Torre & Minchen, Chap. 7 of this volume) in this book.

Data analyses were done in R (R Core Team, 2015). The R packages 'CDM' (George, Robitzsch, Kiefer, Gross, & Uenlue, 2016), 'ltm' (Rizopoulos, 2006), 'mirt' (Chalmers, 2012), and 'TAM' (Robitzsch, Kiefer, & Wu, 2017) were used for the estimation of the psychometric models. The EM algorithm was used to estimate the psychometric models (McLachlan & Krishnan, 1996).

25.3 Results

25.3.1 Absolute Fit

The root mean square error approximation (RMSEA) and the standardized root mean square root (SRMSR) of squared residuals indices were estimated to determine absolute fit (see Table 25.3). Optimal fit is reached when RMSEA and/or

Table 25.3 Absolute fit indices

Model	RMSEA	SRMSR
1-PL IRT	0.062	0.084
2-PL IRT	0.054	0.047
3-PL IRT	0.047	0.045
MIRT 1	0.073	0.155
MIRT 2	0.060	0.114
G-DINA	0.061	0.050
DINA	0.063	0.066
DINO	0.061	0.061
A-CDM	0.052	0.051
HO-DINA	0.067	0.057

Note: RMSEA refers to the root mean square error approximation, SRMSR corresponds to the standardized root mean square root of squared residuals

Table 25.4 Relative fit indices

Model	Number of parameters	LL	AIC	BIC
1-PL IRT	25	−12978.93	26007.86	26129.54
2-PL IRT	50	−12732.13	25564.26	25807.61
3-PL IRT	75	−12675.91	25501.83	25866.85
MIRT 1	69	−14743.23	29624.46	29960.28
MIRT 2	50	−12893.05	25902.11	26184.39
G-DINA	123	−12705.30	25656.61	26255.24
DINA	79	−12887.92	25933.83	26318.32
DINO	79	−12960.97	26079.95	26464.44
A-CDM	98	−12724.16	25644.29	26121.29
HO-DINA	108	−12746.40	25708.79	26234.42

Note: LL corresponds to the log-likelihood

SRMSR are below 0.05 (Liu et al., 2017; Maydeu-Olivares, Cai, & Hernández, 2011). As shown in Table 25.3, the 2-PL and 3-PL IRT models showed the best absolute fit. The G-DINA model showed the best fit among the CDMs in terms of SRMSR index. While the absolute fit measures cannot be used to compare among models, the better absolute fit of the 3-PL model over any CDM may reflect the fact that the test was originally constructed under the principles of the IRT framework.

25.3.2 Model Relative Fit

The CDM and IRT models were compared using fit metrics of log-likelihood, AIC, and BIC. Results for the model relative fit indices are presented in Table 25.4. In terms of log-likelihood, the best fit was reached by the 3-PL IRT model, then the G-DINA and the A-CDM models. These same models also obtained the lowest AIC

values. However, the 2-PL IRT model yielded the lowest BIC among all estimated models, and the A-CDM model among the CDMs.

25.3.3 Item Fit

Item fit was determined using the item level RMSEA index. In Table 25.5, the two IRT models tended to show the best item level fit, each model having 18 items with RMSEA values below 0.05. Among the CDM model, the A-CDM model showed the best item level fit with 15 items below the fit threshold. Moreover, the A-CDM model reached lower RMSEA values in 9 items (i.e., items 3, 6, 7, 13, 17, 19, 20, 21, and 24) compared to the 3-PL IRT model. The DINA and DINO models showed the worst item level fit.

25.3.4 Person Fit

Person fit was assessed using the likelihood ratio test for aberrant behavior (Liu et al., 2009). Table 25.6 summarizes the proportion of examinees showing misfit due to spuriously high and low scores given their estimated skill profile. Once again, the G-DINA and the A-CDM presented the lowest proportion of examinees showing misfit, whereas the DINA and DINO models showed the highest. Consistently across models, a higher proportion of examinees showed misfit due to incongruence between their corresponding lower observed scores and their estimated skill profile.

The evidence from the absolute and relative fit indices favored the 3-PL IRT model followed by the 2-PL IRT model, and then the A-CDM and G-DINA models. The item and person fit results also confirmed a better fit to the data for the G-DINA and the A-CDM models. With the aim to put emphasis on the most parsimonious CDM that showed good fit to the data, the next results only focus on item and examinee estimates produced by the 3-PL IRT model and the A-CDM model.

25.3.5 Item Psychometric Characteristics

With respect to the item characteristics, Table 25.7 presents the 3-PL IRT and A-CDM item parameter estimates. Most items showed an average level of difficulty, with only item 6 being visibly easier than the rest. Item 6 also showed the lowest discrimination among all items in the test. Most items showed very low probability of a correct answer due to guessing, only items 7 and 24 showed guessing estimates above the probability of random chance considering the number of answer options for those items.

Table 25.5 RMSEA item fit indices

Item	2-PL	3-PL	G-DINA	DINA	DINO	A-CDM	HO-DINA
1	0.02	0.02	0.079	0.029	0.039	0.024	0.039
2	0.056	0.056	0.064	0.033	0.034	0.056	0.048
3	0.048	0.048	0.056	0.04	0.031	0.039	0.049
4	0.018	0.018	0.037	0.037	0.124	0.036	0.049
5	0.024	0.024	0.015	0.134	0.064	0.036	0.048
6	0.065	0.065	0.055	0.077	0.067	0.054	0.075
7	0.063	0.064	0.029	0.045	0.022	0.024	0.061
8	0.043	0.043	0.109	0.079	0.064	0.085	0.087
9	0.04	0.04	0.143	0.068	0.061	0.121	0.115
10	0.051	0.052	0.079	0.059	0.065	0.06	0.087
11	0.023	0.023	0.032	0.041	0.062	0.027	0.054
12	0.017	0.017	0.055	0.046	0.027	0.032	0.052
13	0.052	0.051	0.048	0.082	0.032	0.028	0.049
14	0.032	0.031	0.03	0.074	0.076	0.033	0.049
15	0.017	0.018	0.095	0.044	0.057	0.047	0.056
16	0.036	0.038	0.107	0.07	0.058	0.085	0.083
17	0.064	0.064	0.028	0.145	0.145	0.027	0.118
18	0.011	0.011	0.119	0.033	0.052	0.182	0.198
19	0.022	0.023	0.055	0.08	0.037	0.019	0.058
20	0.043	0.042	0.023	0.039	0.057	0.031	0.058
21	0.022	0.022	0.028	0.071	0.071	0.018	0.014
22	0.038	0.038	0.037	0.058	0.059	0.054	0.06
23	0.022	0.023	0.103	0.06	0.08	0.076	0.078
24	0.057	0.057	0.057	0.065	0.067	0.053	0.06
25	0.018	0.019	0.04	0.073	0.074	0.048	0.038
Minimum	0.011	0.011	0.015	0.029	0.022	0.018	0.014
Median	0.036	0.038	0.055	0.06	0.061	0.039	0.058
Maximum	0.065	0.065	0.143	0.145	0.145	0.182	0.198

Table 25.6 Proportion of examinees with aberrant scores

Model	Proportion of examinees with aberrant high scores	Proportion of examinees with aberrant low scores
G-DINA	0.095	0.125
DINA	0.108	0.145
DINO	0.117	0.144
A-CDM	0.096	0.130
HO-DINA	0.098	0.128

In terms of the A-CDM item parameter estimates, the intercept can be interpreted as the baseline probability of correctly answering an item if the examinee has not mastered the skills measured by the item yet. For the most part, these intercept coefficients were negative indicating a low probability for the correct answer if the

Table 25.7 Item psychometric characteristics in 3-PL IRT model and A-CDM model

	3-PL IRT			A-CDM	
Item	Difficulty (b_j)	Discrimination (a_j)	Guessing (c_j)	Intercept (λ_{j0})	Skill coefficients (λ_{jk})
1	−1.401 (0.711)	1.263 (0.288)	0.149 (0.359)	0.534 (0.111)	$\lambda_1 = 2.195$ (0.201)
2	0.935 (0.076)	2.572 (0.506)	0.169 (0.025)	−1.315 (0.089)	$\lambda_2 = 2.914$ (0.202)
3	1.867 (0.164)	2.494 (0.784)	0.082 (0.015)	−2.511 (0.139)	$\lambda_2 = 1.894$ (0.198)
4	0.588 (0.084)	1.862 (0.282)	0.071 (0.033)	−1.923 (0.160)	$\lambda_1 = 1.232$ (0.192), $\lambda_2 = 2.705$ (0.235)
5	−0.829 (0.076)	1.530 (0.130)	0.000 (0.000)	−0.431 (0.109)	$\lambda_1 = 2.295$ (0.187), $\lambda_3 = 0.703$ (0.293)
6	−3.139 (0.583)	0.467 (0.090)	0.000 (0.004)	0.772 (0.112)	$\lambda_5 = 0.596$ (0.182), $\lambda_6 = 1.194$ (0.187)
7	0.265 (0.125)	3.158 (0.875)	0.491 (0.042)	−2.585 (0.390)	$\lambda_4 = 2.510$ (0.381), $\lambda_5 = 3.114$ (0.342), $\lambda_6 = 3.222$ (0.392)
8	0.526 (0.117)	1.959 (0.401)	0.215 (0.048)	−1.020 (0.121)	$\lambda_1 = 0.486$ (0.176), $\lambda_5 = 1.835$ (0.180)
9	−0.285 (0.080)	1.008 (0.095)	0.000 (0.000)	−0.380 (0.083)	$\lambda_5 = 1.834$ (0.157)
10	0.373 (0.945)	0.776 (0.404)	0.078 (0.299)	−0.727 (0.087)	$\lambda_5 = 1.739$ (0.147)
11	1.093 (0.358)	0.739 (0.261)	0.049 (0.126)	−2.321 (0.202)	$\lambda_1 = 1.529$ (0.172), $\lambda_4 = 1.145$ (0.163)
12	−0.229 (0.104)	1.884 (0.229)	0.018 (0.050)	−1.920 (0.160)	$\lambda_1 = 1.562$ (0.187), $\lambda_7 = 2.839$ (0.187)
13	−0.386 (0.064)	1.497 (0.122)	0.000 (0.000)	−1.339 (0.134)	$\lambda_1 = 2.005$ (0.167), $\lambda_7 = 1.128$ (0.165)
14	1.001 (0.109)	1.488 (0.295)	0.112 (0.038)	−1.817 (0.153)	$\lambda_1 = 0.425$ (0.188), $\lambda_2 = 0.851$ (0.182), $\lambda_7 = 1.1$ (0.171)

15	−0.531 (0.181)	1.919 (0.300)	0.203 (0.089)	−0.385 (0.109)	$\lambda_1 = 2.680$ (0.177)
16	−0.367 (0.055)	2.055 (0.163)	0.000 (0.000)	−1.371 (0.133)	$\lambda_1 = 3.097$ (0.174)
17	1.193 (0.069)	3.444 (0.578)	0.074 (0.012)	−2.396 (0.193)	$\lambda_1 = -0.192$ (0.280), $\lambda_3 = 2.945$ (0.241)
18	−0.446 (0.056)	2.072 (0.168)	0.000 (0.003)	−0.904 (0.118)	$\lambda_1 = 2.012$ (0.168), $\lambda_3 = 2.264$ (0.380)
19	−0.106 (0.050)	2.344 (0.187)	0.000 (0.000)	−2.839 (0.208)	$\lambda_1 = 2.210$ (0.213), $\lambda_7 = 3.166$ (0.198)
20	−0.177 (0.137)	1.811 (0.267)	0.164 (0.063)	−1.044 (0.129)	$\lambda_1 = 1.146$ (0.173), $\lambda_3 = 0.335$ (0.300), $\lambda_7 = 1.952$ (0.199)
21	1.915 (0.154)	1.811 (0.257)	0.001 (0.003)	−5.825 (0.989)	$\lambda_1 = 2.517$ (1.028), $\lambda_3 = 2.336$ (0.315)
22	0.467 (0.161)	1.415 (0.280)	0.234 (0.058)	−0.997 (0.113)	$\lambda_5 = 1.474$ (0.153), $\lambda_6 = 1.341$ (0.146)
23	−0.264 (0.054)	2.028 (0.160)	0.000 (0.000)	−1.681 (0.147)	$\lambda_1 = 3.247$ (0.182)
24	0.720 (0.099)	2.288 (0.488)	0.301 (0.035)	−0.717 (0.087)	$\lambda_5 = 2.012$ (0.154)
25	−0.520 (0.182)	1.790 (0.273)	0.018 (0.101)	−1.161 (0.131)	$\lambda_1 = 2.511$ (0.178), $\lambda_7 = 0.950$ (0.182)

Note: the values in parenthesis correspond to the standard error of the estimated coefficient. λ_1 is the coefficient linked to whole numbers; λ_2 to fractions and decimals; λ_3 to number sentences, patterns, and relationships; λ_4 to lines and angles; λ_5 to two- and three-dimensional shapes; λ_6 to location and movement concepts; and λ_7 to skills on reading, interpreting, organizing, and representing data

examinee has not acquired the knowledge and skills measured by the test. When it comes to the skill coefficients, most items showed large coefficients indicating that the mastery of each skill would increase the probability of correctly answering the item. Only item 17 presented a negative coefficient for the *whole numbers* skill, but even the negative impact of the mastery of such skill was compensated by the mastery of the *number sentences, patterns, and relationships* skill also measured by the item.

The skill coefficients also help to determine what skills are potentially more relevant in order to boost the performance in the test. For instance, item 7 measures the three skills linked to the Geometric Shapes and Measures domain, but it is more relevant to master the skill linked to concepts of location and movement in order to correctly answer this item.

25.3.6 Examinee Skill Profile and Ability Estimates

As depicted in Table 25.8, the estimated skill distributions produced by the A-CDM model showed that some skills have been mastered by more than half of the examinees in the sample (e.g., *lines and angles*, or *whole numbers*), whereas others have been acquired only by a small number of examinees (e.g., *fractions and decimals*, or *number sentences, patterns and relationships*).

The A-CDM model estimated 105 different skill profiles out of 128 possible skill arrangements. Because of the large number of skill profiles, Table 25.9 reports only the estimated skill profiles that account for at least 1% of the sample of examinees. Average 3-PL IRT abilities θ were calculated for the group of examinees classified within each one of these skill profiles. As shown in Table 25.9, examinees classified as not being proficient in any of the skills measured by the test (i.e., those having the skill profile 0000000) presented the lowest average IRT ability among all skill profile groups. On the other side, those examinees that have mastered the seven skills reached the highest average IRT ability. The most frequent skill profile corresponded to those examinees that are considered to be proficient in some skills (α_4 *Lines and angles* and α_6 *Location and movement concepts*) corresponding to the Geometric

Table 25.8 Skill mastery distribution

Skill	Not mastered	Mastered
α_1 Whole numbers	0.362	0.637
α_2 Fractions and decimals	0.773	0.226
α_3 Number sentences, patterns, and relationships	0.750	0.249
α_4 Lines and angles	0.382	0.617
α_5 Two- and three-dimensional shapes	0.621	0.378
α_6 Location and movement concepts	0.551	0.448
α_7 Reading, interpreting, organizing, and representing data	0.482	0.517

Table 25.9 A-CDM skill profiles and 3-PL IRT average ability

Skill profile $\alpha_1\alpha_2\alpha_3\alpha_4\alpha_5\alpha_6\alpha_7$	$\overline{\theta}$	SD	# of examinees
0000000	−1.429	0.398	58
0000010	−1.275	0.361	11
0001000	−1.153	0.324	30
0001010	−1.098	0.335	60
0010000	−1.011	0.506	10
0001100	−0.876	0.310	10
0000001	−0.847	0.243	11
0000110	−0.838	0.318	14
0001001	−0.633	0.332	17
0001011	−0.456	0.300	22
1000000	−0.232	0.188	19
1001000	−0.119	0.243	17
1000100	−0.016	0.183	20
1001010	0.008	0.227	19
1000110	0.073	0.236	22
1000001	0.123	0.176	12
1001001	0.240	0.187	16
1000011	0.279	0.155	32
1000101	0.465	0.203	17
1001011	0.480	0.181	27
1000111	0.576	0.152	30
1001111	0.692	0.167	26
1011011	0.718	0.263	10
1011101	0.813	0.173	12
1100111	0.875	0.225	21
1010111	1.044	0.244	10
1101111	1.098	0.199	19
1011111	1.128	0.118	15
1110101	1.279	0.362	17
1111101	1.401	0.330	14
1110111	1.488	0.318	23
1111111	1.649	0.334	32

Note: $\overline{\theta}$ corresponds to the average ability for each skill profile group, SD to the standard deviation of the ability estimates for each skill profile group

Shapes and Measures domain. In general, the higher number of mastered skills, the higher average ability by skill profile group.

Tetrachoric correlations between every pair of latent skills α_k, and biserial correlations between the skills α_k and the estimated ability θ from the 3-PL model were computed. As shown in Table 25.10, most latent skills showed positive correlations among them, being the correlation between skill α_1 the capacity to perform operations using whole numbers and α_5 the understanding and properties of

Table 25.10 Tetrachoric correlations among latent skills and biserial correlations with IRT ability

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	θ
α_1	1							
α_2	0.456	1						
α_3	0.183	0.484	1					
α_4	−0.210	0.020	0.109	1				
α_5	0.638	0.513	0.421	−0.161	1			
α_6	0.316	0.191	0.019	0.107	0.098	1		
α_7	0.502	0.372	0.352	0.102	0.393	0.187	1	
θ	0.788	0.664	0.549	0.028	0.754	0.320	0.752	1

Note: α_1 is linked to whole numbers; α_2 to fractions and decimals; α_3 to number sentences, patterns, and relationships; α_4 to lines and angles; α_5 to two- and three-dimensional shapes; α_6 to location and movement concepts; α_7 to skills on reading, interpreting, organizing, and representing data, and θ is the estimated 3-PL IRT ability

Table 25.11 Mastered skill distribution and average IRT ability by country

Latent variable	Germany	Iran	Japan
α_1 Whole numbers	0.756	0.258	0.762
α_2 Fractions and decimals	0.146	0.169	0.496
α_3 Number sentences, patterns, and relationships	0.223	0.161	0.443
α_4 Lines and angles	0.469	0.510	0.525
α_5 Two- and three-dimensional shapes	0.408	0.276	0.631
α_6 Location and movement concepts	0.624	0.561	0.537
α_7 Reading, interpreting, organizing, and representing data	0.624	0.068	0.831
$\overline{\theta}(SD)$	0.159 (0.651)	−0.769 (0.681)	0.542 (0.845)

two- and three-dimensional shapes the largest. Skill α_4 corresponding to knowledge about lines and angles in geometric figures showed an inconsistent pattern of correlations with the rest of the skills, including negative correlations with other skills linked to the geometry domain. The estimated IRT ability θ showed positive correlations with the majority of the latent skills, being α_4 the only skill with a correlation close to zero.

Additional analyses of the skills and ability estimates by country showed that, on average, test takers from Japan presented a higher proportion of mastered skills and an average ability θ higher than the other two countries. In contrast, examinees from Iran tended to have a lower proportion of mastered skills; the average ability of students in Iran was also lower than that of the other two countries. Interestingly, more than half of the students in Iran were proficient in skills linked to the Geometric Shapes and Measures domain (e.g., *lines and angles*, and *location and movement concepts*) (Table 25.11).

25.4 Discussion

This paper showed an example of how to conduct CDM research retrofitting data from a test originally developed under a different psychometric framework (Gierl & Cui, 2008; Liu et al., 2017). The retrofitting was facilitated by the systematic compiled documentation about the test and items produced by the TIMSS and PIRLS International Study Center (Mullis et al., 2007, 2009; Olson, Martin, & Mullis, 2008). This information allowed us to identify the mathematics domains measured in TIMSS 2007, as well as fine-grained understandings and skills measured by each one of the items. Having this contextual information and the feedback from experts permitted the construction of a Q-matrix that reflected the multidimensionality of the test items.

The results showed that the best absolute- and relative-model fit was produced by the 3-PL and 2-PL IRT models. This result reflects the fact that, after the test administration and to create an item bank, TIMSS 2007 items were analyzed using the IRT framework (Olson et al., 2008). Thus, items may be potentially removed from the test if they do not satisfy strict psychometric criteria. Among the CDMs estimated, the G-DINA and A-CDM models showed the best model fit when compared to the unidimensional 1-PL IRT model, MIRT models, and other CDMs. These two models for cognitive diagnosis also showed better item- and person-fit compared to other CDMs.

Item and skill estimates of the A-CDM model also revealed that the mastery of each skill increases the probability of a correct answer on most of the mathematics items in TIMSS 2007. Similarly, the increase of mastered skills in the estimated skill profiles was associated with a higher average IRT ability estimate. Biserial correlations between each dichotomous skill and the IRT ability estimate in most instances also showed positive correlations. These results showed the convergence of examinee information from the IRT and CDM frameworks; however, whereas the IRT models produce an overall continuous ability estimate, the CDM framework permits a more fine-grained multidimensional analysis of dichotomous skills measured by the items (Rupp & Templin, 2008).

The aggregated analysis by country produced new diagnostic information for the examinees from the three countries. Overall, 4th grade students from Japan have mastered a higher proportion of the skills measured by the mathematics test, followed by their peers from Germany, and then by examinees from Iran. While showing a good performance in the test, German examinees may benefit from more in-depth instruction and practice to promote their understanding on topics such as *fractions and decimals* or *number sentences, patterns, and relationships*. Iranian students require even more opportunities to expand their knowledge in most of the domains and skills assessed in the TIMSS 2007 mathematics test. These results for each country are consistent with the academic achievement trends reported by the test developers (Mullis et al., 2007).

This study presents some limitations that have been previously documented in the context of CDM research and retrofitting (Gierl & Cui, 2008; Haberman &

Davier, 2006; Liu et al., 2017; Xu & Zhang, 2016; von Davier & Haberman, 2014). First, the Q-matrix lacks of some necessary characteristics to support the model identifiability for conjunctive models. This situation will probably be encountered in most retrofitting contexts where the researcher generates a Q-matrix using a test not originally designed to inform about fine-grained attributes linked to the items. Future studies on CDM retrofitting should pay attention to the minimum conditions that must be present in the Q-matrix to achieve model identifiability.

A second limitation comes from the evidence produced by the tetrachoric correlations. The correlation patterns showed that the skill corresponding to *lines and angles* in Geometric Shapes is not internally consistent with the rest of the skills measured by the test. This skill also presented a low biserial correlation close to zero with the IRT ability estimate. The Q-matrix revealed that this skill was only measured twice, so probably these results may suggest that more items measuring this skill should be included in order to reach more consistent results.

Acknowledgements Dr. Luna Bazaldua thanks UNAM for the PAPIIT research grant IA303018.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Baker, F., & Kim, S.-H. (2004). *Item response theory*. New York, NY: Marcel Dekker.

Bradshaw, L., & Templin, J. (2013). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*(3), 403–425.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Chiu, C.-Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*(2), 225–250.

Chiu, C.-Y., Douglas, J., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633–665.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart, and Winston.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429–449.

Culpepper, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics, 40*(5), 454–476.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253–273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595–624.

de la Torre, J., & Lee, Y.-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement, 47*(1), 115–127.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355–373.

de la Torre, J., & Minchen, N. D. (this volume). The G-DINA model framework. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*(1), 8–26.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

DiBello, L. V., Stout, W. F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Erlbaum.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50*, 328–344.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343–368.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Psychology Press.

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*(1), 14–36.

Foy, P., & Olson, J. F. (2009). *TIMSS 2007 user guide for the international database*. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement.

Geisinger, K. F. (2012). Norm- and criterion-referenced testing. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindksopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 371–393). Washington, DC: American Psychological Association.

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Uenlue, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24.

Gierl, M. J., & Cui, Y. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement, 6*, 263–275.

Haberman, S. J., & von Davier, M. (2006). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1031–1038). Amsterdam, The Netherlands: Elsevier.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology, 69*(3), 225–252.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49*(1), 59–81.

Lei, P. W., & Li, H. (2016). Performance of fit indices in choosing correct cognitive diagnostic models and Q-matrices. *Applied Psychological Measurement, 40*(6), 405–417.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*, 548–564.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*, Advance online publication. https://doi.org/10.1177/0013164416685599.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*(8), 579–598.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, Bi-plots, and related graphical displays. *Sociological Methodology, 31*(1), 223–264.

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal, 18*(3), 333–356.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328.

McLachlan, G. J., & Krishnan, T. (1996). *The EM algorithm and extensions*. New York, NY: Wiley.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometric considerations in game-based assessment*. Redwood, CA: GlassLab.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195–215.

Mullis, I. V. S., Martin, M. O., Foy, P., Olson, J. F., Preuschoff, C., Erberber, E., … Galia, J. (2009). *TIMSS 2007 international mathematics report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2007). *TIMSS 2007 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Neyman, J., & Pearson, E. S. (1992). On the problem of the most efficient tests of statistical hypotheses. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 73–108). New York, NY: Springer.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38*(5), 376–390.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Reckase, M. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Robitzsch, A., Kiefer, T., & Wu, M. (2017). *TAM: Test analysis modules* (R package version 2.6-2). Retrieved from https://CRAN.R-project.org/package=TAM

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small.* Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada.

Rupp, A. A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7*, 95–125.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219–262.

Schwarz, G. (1976). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Skaggs, G., Wilkins, J. L. M., & Hein, S. F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic model. *International Journal of Testing, 16*(4), 310–330.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, 64*, 583–639.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317–339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: ETS.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement, 7*(1), 67–74.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)*. ETS research report series. http://onlinelibrary.wiley.com/doi/10.1002/ets2.12043/abstract

von Davier, M., & Haberman, S. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models—A commentary. *Psychometrika, 79*(2), 340–346.

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift für Psychologie/Journal of Psychology, 216*(2), 74–88.

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*(3), 625–649.

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (ETS Research Report No. RR-08-27). Princeton, NJ: ETS.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (ETS Research Report No. RR-89-41). Princeton, NJ: Educational Testing Service.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (Research Report No. RR-03–32). Princeton, NJ: Educational Testing Service.

Part IV
Software, Data, and Tools

Chapter 26
The R Package CDM for Diagnostic Modeling

Alexander Robitzsch and Ann Cathrice George

Abstract In this chapter, the R (R Core Team, R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, 2017) package CDM (Robitzsch A, Kiefer T, George AC, Uenlue A, CDM: cognitive diagnosis modeling. R package version 6.0-101. https://CRAN.R-project.org/package=CDM, 2017; George AC, Robitzsch A, Kiefer T, Groß J, Ünlü A, J Stat Softw 74(2):1–24. 10.18637/jss.v074.i02, 2016) for estimating diagnostic classification models is introduced. First, the model classes that can be estimated with the CDM package are introduced. Second, the CDM package structure and some of its features are discussed. Third, the usage of the CDM package is demonstrated in a data application. Finally, potential future developments of the CDM package are discussed.

In this chapter, the R (R Core Team, 2017) package CDM (George et al., 2016; Robitzsch et al., 2017) for estimating diagnostic classification models is introduced. Diagnostic classification models (DCMs; often also labeled as cognitive diagnostic models, CDMs) are latent variable models in which the multidimensional latent

Correspondence concerning this article should be sent to Alexander Robitzsch, Leibniz Institute for Science and Mathematics Education (IPN), Olshausenstr. 62, 24118 Kiel, Germany. Email: robitzsch@ipn.uni-kiel.de

A. Robitzsch (✉)
Department of Educational Measurement, IPN Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Centre for International Student Assessment, Munich, Germany
e-mail: robitzsch@ipn.uni-kiel.de

A. C. George
Federal Institute for Educational Research, Innovation and Development of the Austrian School System, Salzburg, Austria
e-mail: a.george@bifie.at

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_26

variables are mostly assumed to be discrete (Rupp & Templin, 2008). DCMs can be considered to be restricted latent class models (von Davier, 2009; von Davier & Lee, Chap. 1, this volume). DCMs are often applied to estimation and reporting of individual profiles (e.g., Jang, 2009; see also Rupp & Templin, 2008).

In the following, the model classes that can be estimated with the CDM package are introduced. Second, the CDM package structure and some of its features are discussed. Third, the usage of the CDM package is demonstrated in a data application. Finally, potential future developments of the CDM package are discussed. Readers who are interested in estimation details should additionally consult George et al. (2016). Readers with a particular interest in a step-by-step tutorial of how to use the CDM package are referred to George and Robitzsch (2015) or Ravand and Robitzsch (2015).

26.1 Model Classes

In this section, we introduce the main model classes that can be estimated with the CDM package. We discuss the estimation of DCMs within the frameworks of the generalized DINA model, the generalized diagnostic model, structured latent class analysis, and regularized latent class analysis.

26.1.1 Generalized Deterministic Inputs, Noisy "and" Gate Model (G-DINA)

The generalized deterministic inputs, noisy "and" gate model (G-DINA; de la Torre & Minchen, Chap. 7, this volume; de la Torre, 2011) is a general class of DCMs and is implemented as the `CDM::gdina()` function in the CDM package. In the G-DINA model, the skill vector $\mathbf{A} = (A_1, \ldots, A_D)$ containing dichotomous skills creates $L = 2^D$ latent classes. For a moderate to large number of skills, estimating all $L - 1$ skill probabilities is often not necessary because a model with some parametric reduction often leads to a comparable fit. In more detail, the vector of skill class sizes $\boldsymbol{\pi}$ can be represented as $\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\delta})$ with a specified function $\mathbf{f}$ and an unknown vector of distribution parameters $\boldsymbol{\delta}$. Examples of these parametric reductions are (1) a higher-order model that approximates the D-dimensional skill distribution by a continuous latent variable (de la Torre & Douglas, 2004), (2) an approximation by the representation of a tetrachoric correlation matrix (Templin & Henson, 2006), (3) a log-linear smoothing of skill class sizes (Xu & von Davier, 2008b), or (4) fixing some skill class sizes to zero (Leighton, Gierl, & Hunka, 2004). All of these possibilities are available in the `CDM::gdina()` function (see the arguments `HOGDINA`, `reduced.skillspace` and `zeroprob.skillclasses`).The linking of items to skill classes is specified in the Q-matrix (see Rupp & Templin, 2008). Let $\mathbf{q}_i$ denote the ith row vector of the Q-matrix corresponding to item i.

The nonzero elements in $\mathbf{q}_i$ indicate the skills that are necessary for solving item i. The G-DINA model specifies the probabilities $p_{ic} = P(X_i = 1 \mid \mathbf{a}_i)$, where the latent class c ($c = 1, \ldots, 2^D$) is associated with the skill vector $\mathbf{a}_i$. A link function g (identity, log, or logit link) and the item-specific design matrices $\mathbf{M}_i$ are chosen such that $g(p_{ic}) = \mathbf{m}_{ic}{}^{\mathrm{T}} \boldsymbol{\gamma}_i$, where $\boldsymbol{\gamma}_i$ contains all parameters belonging to item i and the vector $\mathbf{m}_{ic}$ is a row vector of $\mathbf{M}_i$ corresponding to class c. It is assumed that the p_{ic} are only affected by skills that correspond to nonzero elements in $\mathbf{q}_i$. The G-DINA model contains the DINA and DINO models (von Davier & Lee, Chap. 1, this volume), compensatory RUM (Stout, Henson, DiBello, & Shear, Chap. 3, this volume), and the log-linear CDM (LCDM; Henson & Templin, Chap. 8, this volume) as submodels. These models can be specified by the `rule` argument in the `CDM::gdina()` function. Polytomous item responses can be handled by a sequential item modeling approach (Tutz, 1997; see also Ma & de la Torre, 2016).

The G-DINA implementation in the CDM package estimates the model based on marginal maximum likelihood (MML) using an expectation maximization (EM) algorithm (de la Torre, 2009b; George & Robitzsch, 2014; George et al., 2016). By employing the argument `method` the user can choose different approaches of the M-step estimation: the unweighted and weighted least squares approach, as originally proposed by de la Torre (2011), or likelihood-based. The parabolic acceleration method (P-EM) is implemented to accelerate the EM algorithm (Berlinet & Roland, 2012). The `CDM::gdina()` function allows using sampling weights and accommodates multiple group estimation (George & Robitzsch, 2014). Missing data in item responses is allowed in the MML estimation and consistent parameter estimates are obtained if the data is missing at random (MAR).

The variance matrix of the estimated item and distribution parameters in the CDM package can be obtained with resampling methods that are particularly suitable for stratified clustered samples (George & Robitzsch, 2015, 2018; see also Hsieh, Xu, & von Davier, 2010; Park, Lee, & Johnson, 2017). In future versions of CDM, we plan to include variance estimates based on the observed log-likelihood Philipp, Strobl, de la Torre, & Zeileis, 2018) and robust maximum likelihood estimation based on the sandwich formula (Liu, Xin, Andersson, & Tian, 2018; White, 1982). It has been demonstrated in item response theory (IRT) models that the variance estimation method of Oakes (1999) is very promising from the perspective of computational efficiency and unbiasedness (Chalmers, 2018; Pritikin, 2017).

Normal prior distributions can be specified for item parameters and, in this case, the MML estimation becomes a maximum posterior (MAP) estimation. Some researchers prefer to use monotone item response functions in the G-DINA model (Templin & Hoffman, 2013; cf. von Davier, 2014), which can also be requested in `CDM::gdina()`. Monotonicity constraints are handled in the EM algorithm with a penalty function approach (Fiacco & McCormick, 1968). A nonnegative penalty results in the case of monotonicity violation.

In order to apply the G-DINA model, the Q-matrix must be specified. Substantial efforts have been made in DCM research to estimate the Q-matrix from data (for example, Chiu, 2013; de la Torre, 2008; DeCarlo, 2012; Desmarais & Naceur, 2013; Liu, Xu, & Ying, 2013; see Liu & Kang, Chap. 12, this volume, for an overview). These methods either assume that the Q-matrix is partially known or that it needs a substantial amount of prior information, or the proposed methods are computationally very demanding. Researchers Xu and Shang (2018, see also Xu, Chap. 16, this volume) emphasized that the problem of Q-matrix estimation is essentially a variable selection problem and proposed a computationally feasible regularization method (Chen, Li, Liu, & Ying, 2017; Chen, Liu, Xu, & Ying, 2015). In a nutshell, a regularization method substitutes the log-likelihood function with an optimization function that is defined as the difference between the log-likelihood function and a penalty function (for an overview of regularization methods see Hastie, Tibshirani, & Wainwright, 2015, and for applications in item response and structural equation modeling, see Huang, Chen, & Weng, 2017; Sun, Chen, Liu, Ying, & Xin, 2016). A penalty function penalizes the occurrence of a large number of item parameters of negligible size in a DCM. Penalty functions that set the values of coefficients of main or interaction effects that are near to zero to exactly zero are of particular interest. The `CDM::gdina()` function allows the regularization of item parameters and the choice of the penalty functions lasso, elastic net, SCAD, MCP, ridge and truncated L_1 (Breheny & Huang, 2011; Fan & Lv, 2010; Hastie et al., 2015; Shen, Pan, & Zhu, 2012). These different penalty functions show different behavior with respect to statistical bias of item parameters. In the M-step of the EM algorithm, the coordinate descent method is applied which successively maximizes the expected log-likelihood function for every item parameter (Xu & Shang, 2018; see also Sun et al., 2016). Even if the Q-matrix is known, we argue that regularization is worth considering as the full G-DINA model is often highly parameterized. Selecting only the important effects stabilizes the estimation and may facilitate the interpretability of the results.

The G-DINA model has also been extended to multiple choice items. The resulting MC-DINA model (de la Torre, 2009a) needs a Q-matrix specification in which the necessary skills have to be determined for every response option of every item. The most flexible parametrization for the MC-DINA model (Chen & Zhou, 2017) is implemented in the `CDM::mcdina()` function and MML can be used for its estimation. Future versions of the CDM package will include some constrained versions of the MC-DINA model (Ozaki, 2015) and regularization methods.

26.1.2 General Diagnostic Model (GDM)

The general diagnostic model (GDM; von Davier, 2008; von Davier, Chap. 6, this volume) is a general item response model that allows for discrete and continuous latent variables. Each component A_d of the D-dimensional skill vector

$\mathbf{A} = (A_1, \ldots, A_D)$ possesses a finite number of values (e.g., -1 and 1 for dichotomous skills or -6, $-5.4, \ldots,$ 5.4, 6 for a variable that should be modeled with a continuous distribution). In the GDM, a log-linear smoothing as a parametric reduction is typically assumed for the skill class probabilities $P(\mathbf{a})$ (Xu & von Davier, 2008b). In more detail, the logarithm of the probabilities is modeled as a linear function: $\log[P(\mathbf{a})] = \mathbf{z_a}^{\mathrm{T}}\boldsymbol{\delta}$, where $\mathbf{Z} = \{\mathbf{z}_a\}$ is an appropriate design matrix linking the class probabilities to a distribution parameter $\boldsymbol{\delta}$. Hence, unidimensional and multidimensional normal distributions can be represented by modeling the first-order and second-order moments of the distribution (Xu & von Davier, 2008a). Moreover, a consideration of moments larger than two allows the specification of skewed latent variables (Xu & von Davier, 2008a).

The GDM also allows for polytomous item responses. However, in order to simplify the presentation, in this chapter we only consider the case of dichotomous responses. The probability of a correct item response, conditional on a skill vector $\mathbf{a}$, is given as

$$\text{logit } P\,(X_i = 1|\mathbf{a}) = \mathbf{h}_i(\mathbf{a}, \mathbf{q}_i)^{\mathrm{T}}\boldsymbol{\gamma}_i \tag{26.1}$$

where the multidimensional function $\mathbf{h}_i$ models the relationship between skill vector $\mathbf{a}$ and the required skills. Equation (26.1) is linear in the item parameter vector $\boldsymbol{\gamma}_i$ but allows for nonlinear functions (such as interactions) due to the flexibility of choosing $\mathbf{h}_i$. It is evident that a G-DINA model with a logistic link function can be rephrased as a specific GDM. The GDM includes many DCMs as special cases (von Davier, 2014). Moreover, the GDM allows for the estimation of multidimensional 2PL models (von Davier, 2008; Xu & von Davier, Chap. 23, this volume) and it has been extended to accommodate mixture distributions (von Davier, 2010).

The GDM is implemented as the `CDM::gdm()` function in the CDM package. The user can either supply a Q-matrix, which results in a linear compensatory DCM incorporating only main effects, or can use the functions $\mathbf{h}_i(\mathbf{a},\mathbf{q}_i)$ as input, to allow for nonlinear functions (e.g., interactions). As in the `CDM::gdina()` function, MML estimation with the option for P-EM acceleration is used. The function can handle multiple groups, sampling weights, and MAR data.

The `CDM::gdm()` function also allows the estimation of the unidimensional and multidimensional located latent class model (Bartolucci, 2007; De Leeuw & Verhelst, 1986) in which the values of the D-dimensional skill vector are estimated. This model can be regarded as a semiparametric item response model in which the trait distribution can have an (almost) arbitrary form. The `CDM::gdm()` function in its current form can also be used to estimate mixture distribution models (e.g., the mixed Rasch model or mixed 2PL model) by making appropriate specifications of the design matrix $\mathbf{Z}$, which represents the skill class distribution and the functions $\mathbf{h}_i$, which appear in the item response functions. We believe that there is a growing need to investigate the GDM and to promote the understanding of it because the corresponding *mdltm* software (Khorramdel, Shin, & von Davier, Chap. 30, this volume) is now in operational use in the PIAAC and PISA studies.

26.1.3 Structured Latent Class Analysis (SLCA)

Structured latent class analysis (SLCA; Formann, 1985, 1992; Formann & Kohlmann, 1998) is a general approach used for estimating restricted latent class models. Again, we focus on dichotomous item responses in order to simplify our presentation. Let $c = 1, \ldots, L$ denote the latent classes and let $p_{ic} = P(X_i = 1 \mid c)$ be the probability that persons in class c solve an item correctly. The class sizes are $\pi_c = P(C = c)$. The SLCA model poses functional restrictions on p_{ic} and the π_c. In more detail, Formann (1985) assumes a logistic transformation of the item response probabilities: logit $p_{ic} = \mathbf{w}_{ic}{}^{\mathrm{T}}\boldsymbol{\gamma}$, where $\mathbf{w}_{ic}$ is a vector relating the item response probability to a general item parameter $\boldsymbol{\gamma}$. All vectors $\mathbf{w}_{ic}$ (for items $i = 1, \ldots, I$ and classes $c = 1, \ldots, L$) can be collected in a three-dimensional array $\mathbf{W}$. As in the G-DINA model and the GDM, the skill class probabilities are represented as $\log \boldsymbol{\pi} = \mathbf{Z}\boldsymbol{\delta}$ with a design matrix $\mathbf{Z}$ and a class distribution parameter $\boldsymbol{\delta}$. Because the logistic transformation of item response probabilities is employed, SLCA is also referred to as linear logistic latent class analysis (Formann, 1992).

The SLCA model is implemented as the `CDM::slca()` function. Again, MML is used by applying an EM algorithm with P-EM acceleration. The `CDM::slca()` function allows for positivity constraints on the $\boldsymbol{\gamma}$ vector. As in the G-DINA model, regularization methods can be applied by defining penalty functions for item parameters. Multiple group estimation, sampling weights, and MAR missing data are accommodated in the `CDM::slca()` function.

It has been emphasized that DCMs are a special case of restricted latent class models (von Davier, 2009). Many DCMs can therefore be reformulated as a SLCA model. All possible values of the skill vector $\mathbf{A} = (A_1, \ldots, A_D)$ can be equivalently represented as latent classes $c = 1, \ldots, L$. In the case of a D-dimensional skill vector containing dichotomous skills, there are $L = 2^D$ latent classes. The application of the G-DINA model to a given Q-matrix partitions the set of L latent classes for each item into classes with equal item response probabilities. These equality constraints can be formulated by specifying an appropriate design matrix $\mathbf{W}$ in the SLCA model (see Templin & Hoffman, 2013, and Sullivan, Pace, & Templin, Chap. 28, this volume, for a similar approach to estimating DCMs as restricted latent class models in the M*plus* software).

The SLCA model also covers unrestricted latent class models, the linear logistic test model, unidimensional and multidimensional IRT models with continuous traits, located latent class analysis, and mixture distribution models such as the mixed Rasch or the mixed 2PL model (Formann, 2007; Formann & Kohlmann, 1998). An extension of the SLCA (Formann & Kohlmann, 2002) contains even more complex IRT models, such as ability-based guessing models (San Martin, del Pino, & De Boeck, 2006) or models that parametrize person misfit (Raiche, Magis, Blais, & Brochu, 2012). Therefore, we consider the SLCA model to be a unifying framework that contains many important latent variable models.

It has been argued for many newly proposed DCMs that maximum likelihood estimation would be difficult to implement and only a Bayesian MCMC approach

would be computationally feasible. Interestingly, there are several useful examples of DCMs that can be reformulated as a SLCA. The multiple strategy DINA model, for example, is a SLCA model for which MML estimation is efficient (de la Torre & Douglas, 2008; Huo & de la Torre, 2014). Further examples include a restricted version of the MC-DINA model (Ozaki, 2015), the continuous conjunctive model of Hong, Wang, Lim, and Douglas (2015), the random-effect DINA model (Huang & Wang, 2014), and DCMs that involve a simultaneous treatment of skills and misconceptions (Kuo, Chen, & de la Torre, 2018; Templin & Bradshaw, 2014). The SLCA model can also be used for analyzing explanatory diagnostic models (Park & Lee, Chap. 10, this volume).

Beyond the core family of DCMs, the HYBRID model (Yamamoto, 1995), the saltus model (Mislevy & Wilson, 1996), ordered latent class models (Croon, 1990; but see also nonparametric approaches, van der Ark, Rossi, & Sijtsma, Chap. 2, this volume), and confirmatory latent class models (Hojtink & Molenaar, 1997; Nussbeck & Eid, 2015; Vermunt, 2001) are also important cases that can be implemented with the `CDM::slca()` function. A useful concept for integrating continuous traits and ordinal skill levels has been proposed in the structured constructs model for analyzing learning progressions (Wilson, 2009). However, some models have item response functions that are nonlinear in item parameters and can therefore not be handled within the SLCA framework (Shin, Wilson, & Choi, 2017).

Although the SLCA model and the GDM may look different in their formulation of the item response probabilities, we believe that the two frameworks are practically equivalent if appropriate conversions of design matrices and design functions are made in the two models.

26.1.4 Regularized Latent Class Analysis (RLCA)

Regularized latent class analysis (RLCA; Chen et al., 2017) models are special cases of penalized latent class models in which the variability in class-specific item probabilities is regularized. DCMs are restricted latent class models in which equality constraints are specified a priori if the Q-matrix is known and a particular DCM specification (e.g., a DINA or the LCDM model) is selected. While these DCMs are similar to the method of confirmatory factor analysis, the RLCA method resembles the method of exploratory factor analysis, which aims to identify a loading structure. Neither the Q-matrix nor the true DCM specification is known, but the unknown restricted latent class model (i.e., the DCM) needs to be estimated. Chen et al. (2017) proposed an estimation method based on regularization. The RLCA model is a latent class model that estimates class-specific item probabilities p_{ic} and class sizes π_c for classes $c = 1, \ldots, L$. The regularization aims to minimize the number of different probabilities p_{ic} for each item i in order to recover the structure of the DCM. The idea is to order the probabilities such that $p_{i(1)} \leq \ldots \leq p_{i(L)}$

and to apply a penalty on the differences of ascending values. The regularization method statistically identifies ascending probabilities that are set equal to each other.

The CDM package implements the RLCA model in the `CDM::reglca()` function. An EM algorithm for MML estimation can be employed in which the coordinate descent method is used in the M-step. The `CDM::reglca()` function can be applied for multiple group models, sampling weights and MAR missing data. The SCAD penalty (see Chen et al., 2017) or the MCP penalty can be chosen by the user to regularize the variability in the class-specific item probabilities. The use of different starting values is recommended as local maxima often occur in latent class models.

In practice, the estimation of the RLCA model necessitates the specification of the number of latent classes L and a regularization parameter λ that controls the amount of regularization of the penalty (as in all regularization methods). The RLCA model is estimated over a grid of values for L and λ and the best fitting model is chosen based on an information criterion (e.g., BIC). Note that the number of parameters can be easily determined by counting the number of different item parameters (and class sizes) in the RLCA model.

Although the proposed RLCA approach is quite recent, we feel that it has great potential for research questions in which the latent structure is unknown or is only partially known. The RLCA method could not only be interesting in research involving DCMs, but could also be of relevance to latent class-based research areas.

26.2 Package Structure and Features of the CDM Package

The R package CDM provides the main functions `CDM::gdina()`, `CDM::mcdina()`, `CDM::gdm()`, `CDM::slca()`, `CDM::reglca()`, and `CDM::din()` for estimating G-DINA, MC-DINA, GDM, SLCA, RLCA, and DINA (DINO) models, respectively. The generic R S3 methods `summary()`, `print()`, `coef()` and `loglik()` can be applied (see George et al., 2016). Furthermore, additional S3 methods have been implemented which can be used for different model classes. The methods `CDM::IRT.likelihood()` and `CDM::IRT.posterior()` extract individual likelihood and individual posterior vectors (i.e., person-wise values evaluated for all skill classes). The methods `CDM::IRT.irfprob()` and `CDM::IRT.expectedCounts()` return the fitted item response functions and expected counts that are evaluated for all skill classes. These extractor functions are helpful in the calculation of fit statistics. The simulation functions `CDM::sim.gdina()` and `CDM::sim.dina()` are provided for simulating data. In Table 26.1, core estimation functions and S3 methods in the CDM package are presented.

The CDM package provides information criteria (AIC, BIC, CAIC; see also the `logLik()` S3 method) as measures of relative model fit (Sen & Bradshaw, 2017). The fit of several DCMs can be conveniently compared using the S3 methods `anova()` or `CDM::IRT.compareModels()`. The CDM package contains a

Table 26.1 Core estimation functions and S3 methods of the CDM package

Function	Description
Estimation functions	
`din`	DINA/DINO model
`gdina`	G-DINA model
`gdm`	General diagnostic model
`slca`	Structured latent class model
`mcdina`	Multiple choice DINA model
`reglca`	Regularized latent class model
S3 methods	
`IRT.compareModels`	Model comparison
`IRT.likelihood`	Individual likelihood
`IRT.posterior`	Individual posterior
`IRT.irfprob`	Item response functions
`IRT.expectedCounts`	Expected counts
`IRT.modelfit`	Model fit
`IRT.RMSD`	RMSD and MD item fit statistics

variety of measures of absolute model fit that are based on the fit of the bivariate frequencies of all item pairs in the test. The differences between the observed and expected covariances or correlations can be assessed by the SRMSR (Maydeu-Olivares & Joe, 2014), the MADCOV (McDonald & Mok, 1995) statistics, or the average of the absolute difference between the observed and expected correlations (MADcor; Chen, de la Torre, & Zhang, 2013; DiBello, Roussos, & Stout, 2007). These statistics can be requested using the `CDM::IRT.modelfit()` method (for overviews see Hu, Miller, Huggins-Manley, & Chen, 2016; Li, Hunter, & Lei, 2016; Han & Johnson, Chap. 13, this volume). Furthermore, the CDM package includes the item fit statistics RMSD (previously denoted as RMSEA) and MD (Yamamoto, Khorramdel, & von Davier, 2013), which are in operational use in the large-scale assessment studies PIAAC and PISA. These statistics are available in the S3 method `CDM::IRT.RMSD()` (for applications see Kunina-Habenicht, Rupp, & Wilhelm, 2009; Liu, Huggins-Manley, & Bulut, 2018). The S–X2 item fit statistic (Orlando & Thissen, 2000) can be found in the `CDM::itemfit.sx2()` function. Moreover, Wald tests can be employed for item-specific choices of different submodels of the G-DINA model with the `CDM::gdina.wald()` function (de la Torre & Lee, 2013; Ma, Iaconangelo, & de la Torre, 2016; Sorrel, Abad, Olea, de la Torre, & Barrada, 2017).

Differential item functioning (DIF) in the G-DINA model (Hou, de la Torre, & Nandakumar, 2014; Li & Wang, 2015; Qiu, Li, & Wang, Chap. 18, this volume) can be assessed with the `CDM::gdina.dif()` function by applying a Wald test (see George & Robitzsch, 2014). This function also contains a DIF effect size that is similar to an unsigned area measure (George & Robitzsch, 2014). As an exploratory measure, DIF can also be assessed by fitting a multiple group DCM with invariant items and applying the `CDM::IRT.RMSD()` function, which results in RMSD and MD statistics being sensitive to DIF. DIF can also be investigated by using

the SLCA framework to define pseudoitems created by combining an original item and the group as data input. For example, if the group is gender, two pseudoitems of an original item are created which contain item responses of the original item of female and male students, respectively. Joint item parameters and DIF effects can then be specified in the SLCA model. The use of regularization methods in DCMs in the SLCA framework could be potentially interesting as it automatically implies a selection procedure for DIF items and DIF-free items. In a similar vein, Tutz and Schauberger (2015) applied the regularization approach with the lasso penalty to study DIF in the Rasch model.

The adequacy of the classification of individual skill profiles can be assessed by computing classification accuracy and classification consistency (Cui, Gierl, & Chang, 2012; see also Sinharay & Johnson, Chap. 17, this volume). The function `CDM::cdm.est.class.accuracy()` is either based on a simulation or a computation that is based on analytical considerations (see also Wang, Song, Chen, Meng, & Ding, 2015). The prediction error quantified by entropy (see Asparouhov & Muthen, 2014) is implemented as the `CDM::entropy.lca()` function. Item-specific reliability measures, based on the concept of Kullback-Leibler information (Henson, Roussos, Douglas, & He, 2008), are implemented in the `CDM::cdi.kli()` function. In addition, the user can skill hierarchies (`CDM::skillspace.hierarchy()`; cf. Leighton et al., 2004; Templin & Bradshaw, 2014), use deterministic classification (`CDM::din.deterministic()`; see Chiu & Kohn, Chap. 5, this volume), investigate person fit (`CDM::personfit.appropriateness()`; Liu, Douglas, & Henson, 2009) or can analyze the ambiguity of skill classes due to nonidentifiability (`CDM::din.equivalent.class()`; Groß & George, 2014; see also Liu & Kang, Chap. 12, this volume, and Xu, Chap. 16, this volume), to name a few of the several smaller subfunctions in the CDM package.

26.3 Data Application

To illustrate the functionality of the CDM package, we used a dataset from the Examination for the Certificate of Proficiency in English (ECPE) developed by the English Language Institute of the University of Michigan. We chose the grammar section, containing 28 multiple-choice items in which syntactically correct sentences are presented with one word omitted. The data, comprising 2992 students, has already been analyzed by Templin and Hoffman (2013), Templin and Bradshaw (2014), von Davier (2014), and George and Robitzsch (2015). Educational experts identified three underlying skills: comprehension of "morphosyntactic rules" (Skill 1), comprehension of "cohesive rules" (Skill 2), and comprehension of "lexical rules" (Skill 3). The experts decided which item requires which skill in order to be solved correctly and thus specified the Q-matrix (see Templin & Hoffman, 2013, for the Q-matrix). The three skills were measured with 13, 6, and 18 items, respectively. The dataset and the Q-matrix are contained in the CDM package as `data.ecpe`.

Item percent correct values ranged between .43 and .90 ($M = .71$) and item-total discriminations ranged between .26 and .51 ($M = .38$). To investigate the degree of multidimensionality in the data, we computed a tetrachoric correlation matrix of all dichotomous items. Applying a singular value decomposition of the tetrachoric correlation matrix revealed that 23.8% of the total variance was explained by the first factor and unidimensionality appears to hold given a large ratio of the first and second eigenvalue of 5.04. Furthermore, we applied a parallel analysis in order to determine the number of dimensions represented in the data. Three dimensions were statistically significant. Finally, we applied a factor analysis with three factors and a promax rotation and determined our exploratory Q-matrix by attributing an item to a dimension if a factor loading exceeded .2. The Q-matrix we obtained had moderate agreement with the original ECPE Q-matrix of Templin and Hoffman (2013), which was indicated by a congruence coefficient of .67. This finding provides some empirical evidence for the validity of the Q-matrix.

Next, we applied a series of G-DINA models using the `CDM::gdina()` function. We used the original Q-matrix and the full dataset. We fitted a model with two latent classes (indicating nonmastery and mastery of a global skill), the DINA model, the compensatory RUM model (CRUM; the G-DINA model with only main effects and the logistic link function), and the full G-DINA model (involving item parameters for interaction effects) with and without monotonicity constraints. Note that the estimation of the G-DINA model without monotonicity constraints resulted in large standard errors for some item parameters indicating that model parameters are weakly identified (von Davier, 2014). We also estimated the G-DINA model with regularization on all item slope parameters with the SCAD penalty function and a range of 0, .01, ..., .20 for the regularization parameter λ.

Table 26.2 contains the measures of the relative model fit (AIC and BIC) and the absolute fit (MADcor, see DiBello et al., 2007; SRSMR, Maydeu-Olivares & Joe, 2014). The DINA model with three skills fitted the data better than a two-class model assuming only one skill. By comparing only models without regularization, the CRUM model showed the best fit in terms of the AIC and BIC. Moreover, posing monotonicity constraints on the G-DINA model (see Templin & Hoffman, 2013) led only to a slight deterioration of model fit. For the fitted G-DINA models with the differing regularization parameter λ, the model with the value of $\lambda = .14$ had the

Table 26.2 Model comparison for the ECPE dataset for different G-DINA models

Model	Deviance	#Npar	AIC	BIC	MADcor	SRMSR
2 classes	85,945.49	57	86,059	86,400	.028	.035
DINA	85,683.27	63	85,809	86,186	.027	.033
CRUM	85,489.64	72	**85,634**	86,064	.025	.032
G-DINA	85,477.13	81	85,639	86,124	.025	.032
G-DINA, monotonicity constraints	85,479.45	81	85,641	86,126	.025	.032
G-DINA, regularized $\lambda = .14$	85,500.73	69	85,639	**86,051**	.026	.032

Note: #Npar number of estimated parameters. The entries with the lowest AIC and BIC are in bold print

lowest BIC value and was chosen for model comparisons. This model was superior to all other models in terms of the BIC.

Table 26.3 shows the item parameters for the G-DINA model with monotonicity constraints and the regularized solution with the optimal $\lambda = .14$. In the latter model, twelve item parameters (nine main effects, three interaction effects) were regularized. Due to the dependence of the parameters of an item, regularizing an effect implies a change in the other item parameters. For example, for Item 1, the main effects were regularized and set to zero (implying that Item 1 follows the DINA rule), which reduced the number the item parameters from four to two and which led to an increased parameter of the interaction effect. Conversely, for Item 7, the interaction effect was set to zero, resulting in changes in the parameters of both main effects. In the regularized model, the skills have marginal class proportions of 39.9% (Skill 1), 54.9% (Skill 2), and 66.4% (Skill 3). The tetrachoric correlations between the three skills were high (Skills 1 and 2: .88; Skills 1 and 3: .80; Skills 2 and 3: .92), although the skills can be statistically separated from each other. Most of the students either possess all skills (skill pattern "111" with a relative frequency of 36.2%) or no skills (pattern "000" with frequency of 30.7%). From the remaining six latent classes, only two skill patterns ("001" and "011") have probabilities substantially different from zero ("100": 0.8%; "010": 1.0%; "001": 11.9%; "110": 1.1%; "101": 1.7%; "011": 16.6%). This implies that the skills were (almost perfectly) ordered in the ECPE dataset and therefore we may deduce a linear hierarchy (Templin & Bradshaw, 2014).[1] However, the computationally more demanding bootstrap method must be applied to obtain valid standard errors from regularized models.

The G-DINA models are restricted latent class models in which the loading structure of the item was known (specified in the Q-matrix). In a next step, we compared this confirmatory approach with an exploratory approach in which we fitted unrestricted latent class models and regularized latent class models. For the regularized latent class models, we again used the SCAD penalty and varied the regularization parameter $\lambda = 0, .01, \ldots, .05$.[2] The models were fitted using the `CDM::reglca()` function.

Table 26.4 provides information about the information criteria of the fitted models. Among the unregularized latent class models, the three-class solution had the best fit in terms of the BIC. Notably, the fit improved compared to the confirmatory G-DINA models. Among the regularized latent class models, for every fixed number of latent classes, the regularization parameter of $\lambda = .02$ had the

[1] Researchers von Davier and Haberman (2014) showed that a linear hierarchy among skills implies a reduced number of identifiable item parameters.

[2] Note that in the G-DINA model, larger regularization parameters were chosen because item parameters were estimated in the logit metric. In the regularized latent class model, item parameters are estimated in the metric of probabilities and, hence, smaller values have to be chosen. For smaller sample sizes, a wider range of λ values should be chosen.

Table 26.3 Item parameters for the ECPE dataset for fitted G-DINA models (based on three skills) with monotonicity constraints and with regularization (for the optimal parameter$\lambda = .14$)

	G-DINA model (monotonicity constraints)								G-DINA model (regularized solution)						
Item	γ_0	γ_1	γ_2	γ_3	γ_{12}	γ_{13}	γ_{23}	#Nreg	γ_0	γ_1	γ_2	γ_3	γ_{12}	γ_{13}	γ_{23}
1	0.83	*0.00*	0.60	–	1.21	–	–	2	0.96	**0.00**	**0.00**	–	1.70	–	–
2	1.03	–	1.25	–	–	–	–	0	1.03	–	1.24	–	–	–	–
3	−0.34	0.76	–	0.35	–	0.53	–	2	−0.19	**0.00**	–	**0.00**	–	1.48	–
4	−0.14	–	–	1.69	–	–	–	0	−0.13	–	–	1.69	–	–	–
5	1.08	–	–	2.02	–	–	–	0	1.09	–	–	2.02	–	–	–
6	0.86	–	–	1.69	–	–	–	0	0.87	–	–	1.69	–	–	–
7	−0.11	2.85	–	0.95	–	*−0.95*	–	1	−0.07	1.91	–	0.87	–	**0.00**	–
8	1.47	–	1.92	–	–	–	–	0	1.49	–	1.85	–	–	–	–
9	0.12	–	–	1.20	–	–	–	0	0.12	–	–	1.21	–	–	–
10	0.05	2.05	–	–	–	–	–	0	0.03	2.00	–	–	–	–	–
11	−0.04	0.82	–	0.96	–	0.77	–	1	−0.05	1.54	–	0.95	–	**0.00**	–
12	−1.77	*0.00*	–	1.29	–	1.52	–	1	−1.75	**0.00**	–	1.25	–	1.50	–
13	0.66	1.63	–	–	–	–	–	0	0.64	1.60	–	–	–	–	–
14	0.17	1.37	–	–	–	–	–	0	0.16	1.36	–	–	–	–	–
15	1.00	–	–	2.11	–	–	–	0	1.01	–	–	2.11	–	–	–

(continued)

Table 26.3 (continued)

Item	G-DINA model (monotonicity constraints)							#Nreg	G-DINA model (regularized solution)						
	γ_0	γ_1	γ_2	γ_3	γ_{12}	γ_{13}	γ_{23}		γ_0	γ_1	γ_2	γ_3	γ_{12}	γ_{13}	γ_{23}
16	−0.10	2.22	–	0.89	–	−0.75	–	1	−0.08	1.48	–	0.83	–	**0.00**	–
17	1.35	–	0.75	0.59	–	–	0.09	2	1.48	–	**0.00**	**0.00**	–	–	1.44
18	0.93	–	–	1.39	–	–	–	0	0.93	–	–	1.39	–	–	–
19	−0.19	–	–	1.85	–	–	–	0	−0.19	–	–	1.85	–	–	–
20	−1.39	0.26	–	0.91	–	1.40	–	1	−1.38	**0.00**	–	0.88	–	1.61	–
21	0.16	1.04	–	1.13	–	0.05	–	1	0.17	1.08	–	1.11	–	**0.00**	–
22	−0.87	–	–	2.24	–	–	–	0	−0.87	–	–	2.25	–	–	–
23	0.66	–	2.07	–	–	–	–	0	0.67	–	2.01	–	–	–	–
24	−0.68	–	1.52	–	–	–	–	0	−0.66	–	1.49	–	–	–	–
25	0.09	1.13	–	–	–	–	–	0	0.09	1.10	–	–	–	–	–
26	0.16	–	–	1.12	–	–	–	0	0.17	–	–	1.11	–	–	–
27	−0.89	1.71	–	–	–	–	–	0	−0.90	1.68	–	–	–	–	–
28	0.57	–	–	1.74	–	–	–	0	0.58	–	–	1.74	–	–	–

Note: *#Nreg* number of regularized parameters per item, γ_0 Item intercept, γ_k Item parameter for main effects ($k = 1,2,3$), γ_{kh} Item parameter for interaction effects ($k, h = 1,2,3; k \neq h$). Item parameters for which a monotonicity constraint had to be applied are in italics. Item parameters that were regularized are in bold print

Table 26.4 Model comparison for the ECPE dataset for different unrestricted and regularized latent class models (for the optimal regularization parameter $\lambda = .02$)

Model	Deviance	#Npar	#Nreg	AIC	BIC
2 Classes	85,945.48	57	–	86,059	86,400
3 Classes	85,095.66	86	–	85,268	**85,782**
4 Classes	84,906.48	115	–	85,136	85,824
5 Classes	84,806.88	144	–	85,095	85,956
6 Classes	84,721.14	173	–	85,067	86,102
7 Classes	84,654.97	202	–	85,059	86,267
8 Classes	84,572.01	231	–	**85,034**	86,415
5 Classes, regularized	84,969.03	99	45	85,167	85,759
6 Classes, regularized	84,820.13	114	59	85,048	85,730
7 Classes, regularized	84,809.65	108	94	85,026	85,671
8 Classes, regularized	84,813.27	106	125	**85,025**	85,659
9 Classes, regularized	84,860.74	99	161	85,059	**85,651**

Note: *#Npar* number of estimated parameters, *#Nreg* number of regularized parameters. The entries with the lowest AIC and BIC are in bold print

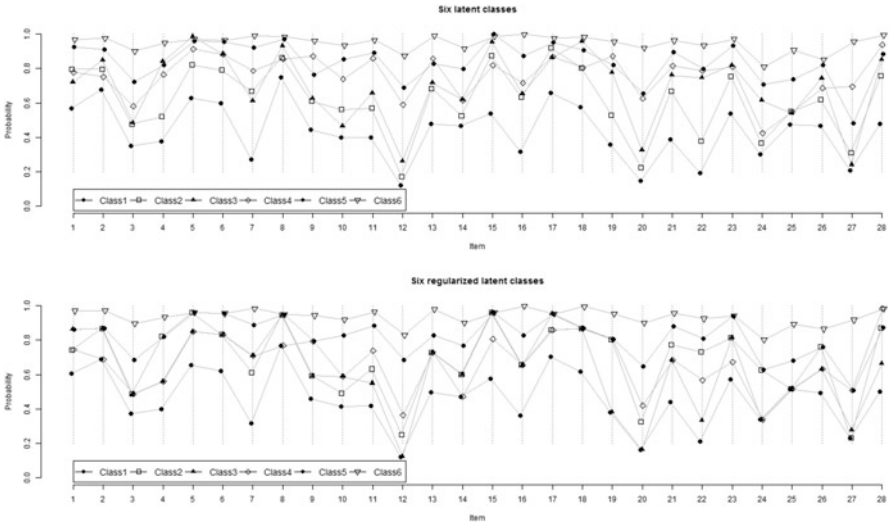


Fig. 26.1 Item parameters for the unregularized and the regularized solution for six classes. Upper panel: unregularized solution. Lower panel: regularized solution

lowest value in terms of the BIC. The number of latent classes would have to be substantially increased if the model choice was based on the minimal BIC.

Notably, the regularized models were superior to the unregularized models with respect to model fit.

The effect of regularization is displayed in Fig. 26.1. In the upper panel of Fig. 26.1, the class-specific item probabilities for the unregularized solution are

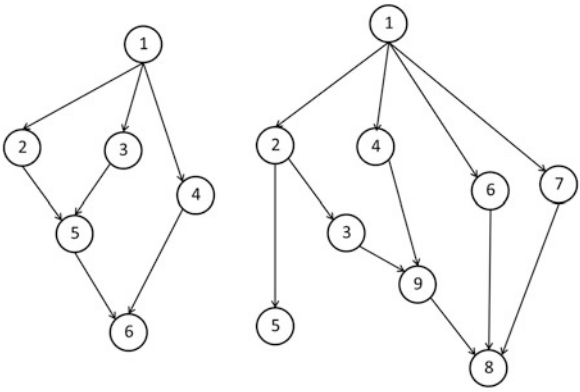


Fig. 26.2 Partial order of the latent classes based on regularized latent class analysis. Left: model with six latent classes. Right: model with nine latent classes

displayed. It is evident that, for every item, there are some latent classes with similar probabilities. As is shown in the lower panel of Fig. 26.1, several probabilities within an item were set equal to each other, which resulted in a more parsimonious model. In total, 59 (out of 168) item parameters were regularized in the six-class model. As a result, the latent classes in the regularized solution imply a partial order (Chen et al., 2017). These partial orders can help to infer the unknown DCM.

In the left-hand panel of Fig. 26.2, the derived partial order for six latent classes is displayed. Skill 1 (relative frequency of 15.0%) is a precursor of all other five skills and Skill 6 (12.1%) is a successor of all other five skills. Skills 2, 3, 4, and 5 are intermediate skills of which Skill 2 (17.6%) and Skill 3 (11.8%) are precursors of Skill 5 (30.4%). The right-hand panel of Fig. 26.2 shows the partial order for the regularized nine-class solution. Although this solution is superior to the six-class solution in terms of model fit, the interpretability of the partial order is somewhat more intricate.

Finally, we fitted the unidimensional latent trait models using the GDM. For the item response functions, we assumed 1PL and 2PL functions. We employed different distributional assumptions for the unidimensional trait. First, a normal distribution was assumed, which can be specified by including the first and second moments in the skill space representation of the GDM. Second, a skewed trait distribution was assumed by fitting the first three moments of the skill space. Third, the trait distribution was fitted by assuming a located latent class model (Bartolucci, 2007). For a model with C latent classes, C locations (values at the θ scale) and $C - 1$ class probabilities are freely estimated. The models were fitted using the `CDM::gdm()` function. Item 24 was used as a reference item and its item difficulty was fixed to zero in the 1PL and the 2PL model and its item slope was fixed to one in the 2PL model.

Table 26.5 shows the measures of relative and absolute model fit for the 1PL and the 2PL model and the different distributional assumptions for the trait. For the 1PL model, the model with a skewed trait ($M = 0.21$, $SD = 1.03$, skewness $= 1.43$, EAP reliability $= .75$) fitted the data significantly better than those with a normally distributed trait ($M = 0.17$, $SD = 0.90$, skewness $= 0$, EAP reliability $= .77$). The

Table 26.5 Model comparison for the ECPE dataset for different unidimensional 1PL and 2PL models estimated with the GDM

Model	Deviance	#Npar	AIC	BIC	MADcor	SRMSR
1PL, normal distribution	85,459.68	28	85,516	85,683	.031	.039
1PL, skewed distribution	85,403.11	30	**85,463**	**85,643**	.031	.039
1PL, 2 classes	86,235.04	30	86,295	86,474	.036	.046
1PL, 3 classes	85,535.73	32	85,600	85,791	.031	.039
1PL, 4 classes	85,403.00	34	85,471	85,674	.031	.039
1PL, 5 classes	85,395.83	36	85,468	85,683	.031	.039
1PL, 6 classes	85,395.19	38	85,471	85,698	.031	.039
2PL, normal distribution	85,094.40	56	85,206	85,541	.018	.022
2PL, skewed distribution	85,041.51	57	**85,156**	**85,496**	.018	.022
2PL, 2 classes	85,945.52	57	86,060	86,400	.028	.035
2PL, 3 classes	85,192.85	59	85,311	85,664	.018	.023
2PL, 4 classes	85,046.84	61	85,169	85,534	.018	.023
2PL, 5 classes	85,034.95	63	85,161	85,538	.018	.022
2PL, 6 classes	85,031.40	65	85,161	85,550	.018	.022

Note: #Npar = number of estimated parameters. The entries with the lowest AIC and BIC are in bold print

located latent class models showed slightly worse performance than the models that assumed a skewed distribution. However, a located latent class model with four classes ($M = 0.19$, $SD = 0.93$, skewness = 0.49, EAP reliability = .78) fitted the data better than a model based on the normal distribution. Estimated class locations were -1.36 (with a frequency of 10.5%), -0.36 (40.7%), 0.64 (37.7%), and 2.14 (11.1%), respectively. The model comparisons for the 2PL models are similar to the 1PL models. The model with a skewed trait distribution ($M = 0.20$, $SD = 0.93$, skewness = 2.60, EAP reliability = .78)[3] fitted the data better than the model with the normal trait distribution ($M = 0.16$, $SD = 0.71$, skewness = 0, EAP reliability = .78) and all located latent class models. If the research question is focused on (diagnostic) classification, then latent class models seem to have some merits. If we compare the information criteria from Table 26.5 with the criteria resulting from the models with discrete latent variables (Tables 26.2, 26.3, and 26.4), we can conclude that simple unidimensional models are competitive with restricted and unrestricted latent class models (see also von Davier, 2014).

[3]The standard deviation of the 2PL model cannot be directly compared with the 1PL model as the value depends on the choice of the reference item.

26.4 Discussion

This chapter introduced the CDM package. We argue that the SLCA model provides a comprehensive framework that includes many important DCMs as special cases. Moreover, some recent developments in exploratory DCMs have been included in recent version of the package. Regularization methods for Q-matrix estimation as well as regularized latent class analysis can be useful statistical tools in diagnostic modeling if the linking of items to skills is (partially) unknown.

In addition to the CDM package, the GDINA package is a comprehensive R package that enables the estimation of the G-DINA model (Ma, Chap. 29, this volume). Overviews of alternative software packages are provided by Li et al. (2016) or Rupp and Templin (2008) and in other chapters of the section "Software, Data, and Tools" in this handbook.

Until now, MML estimation has been the estimation method employed in the CDM package. For a large number of skills, the computation of posterior probabilities and expected counts for all 2^D skill classes is computationally challenging. Bayesian Markov chain Monte Carlo (MCMC) estimation (Culpepper & Hudson, 2018; Zhan, 2017; Liu & Johnson, Chap. 31, this volume) can circumvent this problem when the D dichotomous skills are computationally represented by D underlying continuous multivariate normally distributed skills (Stout et al., Chap. 3, this volume). Statistical inference for model parameters is obtained as a by-product of MCMC estimation and posterior distributions for derived parameters can be easily computed.

Model misfit can occur if some required skills for an item are omitted and, in that case, the loading structure is not correctly specified. Regularization methods can be used to infer unknown entries in the Q-matrix or in the full Q-matrix. However, model misfit can also be due to nonmodeled residual correlations between items that indicate local stochastic dependence. In the same manner as for the loading structure, local dependence can be modeled by applying regularization methods to residual correlations. In this case, MML estimation becomes computationally infeasible and pseudo-likelihood estimation methods have been proposed (Kang, Liu, & Ying, 2017, for DCMs; Chen, Li, Liu, & Ying, 2016, for multidimensional IRT models; Hastie et al., 2015, for general approaches in graphical modeling). Regularization methods can be interpreted in a similar way to prior distributions in Bayesian modeling and can ensure that the regularized set of model parameters remain identified. From a substantive point of view, it could be more useful to represent sources of model misfit in the structure of residual correlations because the interpretation of skills should not be changed by altering Q-matrix entries solely for reasons of statistical criteria.

DCMs assume that latent variables are multidimensional and dichotomous (or polytomous). In principle, polytomous ordered variables with many levels and continuous variables can hardly be distinguished from each other (e.g., von Davier, Naemi, & Roberts, 2012). In empirical applications, multidimensional IRT models with continuous latent traits can often describe data better than models with discrete

latent variables, although this will not always be the case as it depends on the particular application. Multidimensional noncompensatory IRT models can also be applied to model the noncompensatory interplay of skills (Embretson, Chap. 9, this volume), which weakens the borders between DCMs and IRT models.

References

Asparouhov, T., & Muthen, B. (2014). *Variable-specific entropy contribution* (Technical appendix). http://www.statmodel.com/7_3_papers.shtml

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*(2), 141–157.

Berlinet, A. F., & Roland, C. (2012). Acceleration of the EM algorithm: P-EM versus epsilon algorithm. *Computational Statistics & Data Analysis, 56*(12), 4122–4137.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics, 5*(1), 232–253.

Chalmers, R. P. (2018). Numerical approximation of the observed information matrix with Oakes' identity. *British Journal of Mathematical and Statistical Psychology, 71*(3), 415–436.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140.

Chen, J., & Zhou, H. (2017). Test designs and modeling under the general nominal diagnosis model framework. *PLoS One, 12*(6), e0180016.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2016). *A fused latent and graphical model for multivariate binary data*. arXiv:1606.08925.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika, 82*(3), 660–692.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850–866.

Chiu, C. Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Chiu, C.-Y., & Köhn, H.-F. (this volume). Nonparametric methods in cognitively diagnostic assessment. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*(2), 171–192.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 198–138.

Culpepper, S. A., & Hudson, A. (2018). An improved strategy for Bayesian estimation of the reduced reparametrized unified model. *Applied Psychological Measurement, 42*(2), 99–115.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*(3), 163–183.

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353.

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595–624.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*(4), 355–373.

de la Torre, J., & Minchen, N. D. (this volume). The G-DINA model framework. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

De Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11*(3), 183–196.

Decarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468.

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based Q-matrices. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 441–450). Berlin, Germany: Springer.

Dibello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, volume 26, psychometrics* (pp. 979–1030). Amsterdam, The Netherlands: Elsevier.

Embretson, S. E. (this volume). Diagnostic modeling of skill hierarchies and cognitive processes with MLTM-D. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Fan, J., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica, 20*(1), 101–148.

Fiacco, A. V., & McCormick, G. P. (1968). *Nonlinear programming: Sequential unconstrained minimization techniques*. New York, NY: Wiley.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38*(1), 87–111.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87*(418), 476–486.

Formann, A. K. (2007). (Almost) equivalence between conditional and mixture maximum likelihood estimates for some models of the Rasch type. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 177–189). New York, NY: Springer.

Formann, A. K., & Kohlmann, T. (1998). Structural latent class models. *Sociological Methods & Research, 26*(4), 530–565.

Formann, A. K., & Kohlmann, T. (2002). Three-parameter linear logistic latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 183–210). Cambridge, UK: Cambridge University Press.

George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling, 56*(4), 405–432.

George, A. C., & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology, 11*(3), 189–205.

George, A. C., & Robitzsch, A. (2018). Focusing on interactions between content and cognition: A new perspective on gender differences in mathematical sub-competencies. *Applied Measurement in Education, 31*(1), 79–97.

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2), 1–24. https://doi.org/10.18637/jss.v074.i02

Groß, J., & George, A. C. (2014). On prerequisite relations between attributes in noncompensatory diagnostic classification. *Methodology, 10*(3), 100–107.

Han, Z., & Johnson, M. S. (this volume). Global- and item-level model fit indices. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press.

Henson, R., & Templin, J. L. (this volume). Loglinear cognitive diagnostic model (LCDM). In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.

Hojtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*(2), 171–189.

Hong, H., Wang, C., Lim, Y. C., & Douglas, J. (2015). Efficient models for cognitive diagnosis with continuous and mixed-type latent variables. *Applied Psychological Measurement, 39*(1), 31–43.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98–125.

Hsieh, C. A., Xu, X., & von Davier, M. (2010). *Variance estimation for NAEP data using a resampling-based approach: An application of cognitive diagnostic models* (RR-10-26). Princeton, NJ: Educational Testing Service.

Hu, J., Miller, M. D., Huggins-Manley, A. C., & Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnosis models. *International Journal of Testing, 16*(2), 119–141.

Huang, H. Y., & Wang, W. C. (2014). The random-effect DINA model. *Journal of Educational Measurement, 51*(1), 75–97.

Huang, P. H., Chen, H., & Weng, L. J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika, 82*(2), 329–354.

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement, 38*(6), 464–485.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing, 26*(1), 31–73.

Kang, H.-A., Liu, J., & Ying, Z. (2017). *A general diagnostic classification model*. arXiv:1707.06318.

Khorramdel, L., Shin, H. J., and von Davier, M. (this volume). GDM software *mdltm* including parallel EM algorithm. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2009). A practical illustration of multidimensional diagnostic skills profiling: Comparing results from confirmatory factor analysis and diagnostic classification models. *Studies in Educational Evaluation, 35*(2–3), 64–70.

Kuo, B. C., Chen, C. H., & de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Applied Psychological Measurement, 42*(3), 179–191.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule space approach. *Journal of Educational Measurement, 41*(3), 205–237.

Li, H., Hunter, C. V., & Lei, P. W. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing, 33*(3), 391–409.

Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement, 52*(1), 28–54.

Liu, X., & Johnson, M. S. (this volume). Estimating CDMs using MCMC. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Liu, J., & Kang, H.-A. (this volume). Q-matrix learning via latent variable selection and identifiability. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Liu, J., Xu, G., & Ying, Z. (2013). Theory of the self-learning Q-matrix. *Bernoulli, 19*(5A), 1790–1817.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357–383.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*(8), 579–598.

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2018). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology* (in press). https://doi.org/10.1111/bmsp.12134.

Ma, W. (this volume). Cognitive diagnosis modeling using the GDINA R package. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*(3), 253–275.

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*(3), 200–217.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*(1), 23–40.

Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika, 61*(1), 41–71.

Nussbeck, F. W., & Eid, M. (2015). Multimethod latent class analysis. *Frontiers in Psychology | Quantitative Psychology and Measurement, 6*, 1332.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 61*(2), 479–482.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64.

Ozaki, K. (2015). DINA models for multiple-choice items with few parameters: Considering incorrect answers. *Applied Psychological Measurement, 39*(6), 431–447.

Park, Y. S., & Lee, Y.-S. (this volume). Explanatory cognitive diagnostic models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Park, J. Y., Lee, Y.-S., & Johnson, M. S. (2017). An efficient standard error estimator of the DINA model parameters when analysing clustered data. *International Journal of Quantitative Research in Education, 4*(1–2), 159–190.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics, 43*(1), 88–115.

Pritikin, J. N. (2017). A comparison of parameter covariance estimation methods for item response models in an expectation-maximization framework. *Cogent Psychology, 4*, 1279435.

Qiu, X.-L., Li, X., & Wang, W.-C. (this volume). Differential item functioning in diagnostic classification models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raiche, G., Magis, D., Blais, J. G., & Brochu, P. (2012). Taking atypical response patterns into account. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scale assessment in education: Theory, issues and practice* (pp. 238–259). New York, NY: Routledge.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation, 20*(11), 1–12.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2017). *CDM: Cognitive diagnosis modeling*. R package version 6.0-101. https://CRAN.R-project.org/package=CDM

Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262.

San Martin, E. S., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement, 30*(3), 183–203.

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement, 41*(6), 422–438.

Shen, X., Pan, W., & Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association, 107*(497), 223–232.

Shin, H. J., Wilson, M., & Choi, I. H. (2017). Structured constructs models based on change-point analysis. *Journal of Educational Measurement, 54*(3), 306–332.

Sinharay, S., & Johnson, M. S. (this volume). Measures of agreement: Reliability, classification accuracy, and classification consistency. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*(8), 614–631.

Stout, W., Henson, R., DiBello, L., & Shear, B. (this volume). The reparameterized unified model system: a diagnostic assessment modeling approach. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). Latent variable selection for multidimensional item response theory models via L_1 regularization. *Psychometrika, 81*(4), 921–939.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317–339.

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

Tutz, G. (1997). Sequential models for ordered responses. In W. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 139–152). New York, NY: Springer.

Tutz, G., & Schauberger, G. (2015). A penalty approach to differential item functioning in Rasch models. *Psychometrika, 80*(1), 21–43.

van der Ark, L. A., Rossi, G., & Sijtsma, K. (this volume). Nonparametric item response theory and mokken scale analysis, with relations to latent class models and cognitive diagnostic models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Vermunt, J. K. (2001). The use of restricted latent class models for defining and testing non-parametric and parametric item response theory models. *Applied Psychological Measurement, 25*(3), 283–294.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287–307.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research & Perspectives, 7*(1), 67–74.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*(1), 8–28.

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (RR-14-40). Educational Testing Service. Princeton, NJ.

von Davier, M. (this volume). The general diagnostic model. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

von Davier, M., & Haberman, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional 'diagnostic' classification models − A commentary. *Psychometrika, 79*(2), 340–346.

von Davier, M., & Lee, Y.-S. (this volume). Introduction: From latent class analysis to DINA and beyond. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

von Davier, M., Naemi, B., & Roberts, R. D. (2012). Factorial versus typological models: A comparison of methods for personality data. *Measurement: Interdisciplinary Research and Perspectives, 10*(4), 185–208.

Wang, W., Song, L., Chen, P., Meng, Y., & Ding, S. (2015). Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *Journal of Educational Measurement, 52*(4), 457–476.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730.

Xu, G. (this volume). Identifiability and cognitive diagnosis models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Xu, G., & Shang, Z. (2018). Identifying latent structures in restricted latent class models. *Journal of the American Statistical Association, 113*(523), 1284–1295.

Xu, X., & von Davier, M. (2008a). *Comparing multiple-group multinomial log-linear models for multidimensional skill distributions in the general diagnostic model* (RR-08-35). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008b). *Fitting the structured general diagnostic model to NAEP data* (RR-08-27). Educational Testing Service. Princeton, NJ.

Xu, X., & von Davier, M. (this volume). Applying the general diagnostic model to proficiency data from a national skills survey. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models*. Cham, Switzerland: Springer.

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the HYBRID model* (TOEFL TR-10). Educational Testing Service. Princeton, NJ.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013). Chapter 17: Scaling PIAAC cognitive data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)*. Paris, France: OECD.

Zhan, P. (2017). *Using JAGS for Bayesian cognitive diagnosis models: A tutorial*. arXiv:1708.02632.

Chapter 27
Diagnostic Classification Modeling with flexMIRT

Li Cai and Carrie R. Houts

Abstract In this chapter, we will focus on the use of flexMIRT® (Cai L, flexMIRT® version 3.5: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Vector Psychometric Group, LLC, Chapel Hill, 2017) for estimating certain core diagnostic models that have seen practical application, as well as to illustrate the specialized capabilities the software offers. flexMIRT is a commercially available, stand-alone, general purpose item response theory (IRT) software program that is compatible with machines running Windows 7.0 or later. The basic DCM model in flexMIRT is described in Cai, Choi, Hansen, and Harrell (Annu Rev Stat Appl 3:297–321, 2016) as well as in Hansen, Cai, Monroe, and Li (Br J Math Stat Psychol 69:225–252, 2016) in slightly more restricted form. It is an extension of the log-linear cognitive diagnostic model (LCDM) described by Henson, Templin, and Willse (Psychometrika 74:191–210, 2009) with extra random effects to handle cases of possible local dependence.

27.1 Introduction

There are many software options available for estimating diagnostic classification models (DCMs), each with unique features and capabilities (e.g., Chaps. 26, 27, 28, 29, 30 and 31 organized as Part IV in this volume). In this chapter, we will focus on the use of flexMIRT® (Cai, 2017) for estimating certain core diagnostic models that have seen practical application, as well as to illustrate the specialized capabilities the software offers.

L. Cai (✉)
University of California, Los Angeles, CA, USA

Vector Psychometric Group, LLC, Chapel Hill, NC, USA
e-mail: lcai@ucla.edu

C. R. Houts
Vector Psychometric Group, LLC, Chapel Hill, NC, USA
e-mail: crhouts@vpgcentral.com

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_27

27.2 General Software Description

flexMIRT is a commercially available, stand-alone, general purpose item response theory (IRT) software program that is compatible with machines running Windows 7.0 or later. The C++ engine is portable and is compatible with most platforms for which modern C++ compilers exist. It is almost exclusively a syntax-driven program, although some features are available to be specified via a Windows graphical user interface. flexMIRT was first released in 2012, with the module adapting flexMIRT to allow for the estimation of DCMs using model-specific syntax (such as attributes, main effects, and interactions) included as part of the version 2.0 update, released in 2013.

In submitting data to flexMIRT, no specialized data format is used and requirements for the structure of the data file are minimal. Variables need only be tab, comma, or space delimited and missing values must be represented with a numeric value (default is −9). flexMIRT uses full-information maximum marginal likelihood (MML) estimation with the EM algorithm (see e.g., Bock & Aitkin, 1981) to estimate DCMs (or modal Bayes if priors are supplied).

The basic DCM model in flexMIRT is described in Cai, Choi, Hansen, and Harrell (2016) as well as in Hansen, Cai, Monroe, and Li (2016) in slightly more restricted form. It is an extension of the log-linear cognitive diagnostic model (LCDM) described by Henson, Templin, and Willse (2009) with extra random effects to handle cases of possible local dependence. With parameter restrictions implemented in flexMIRT, the LCDM framework can generate a number of "named" DCMs, e.g., the deterministic-input noisy "and" gate (DINA; Junker & Sijtsma, 2001) model.

In developing DCMs and estimation algorithms, conditional independence of item responses given the attributes (e.g., Templin & Henson, 2006) are frequently assumed. In other words, the conditional probability of item response patterns is assumed to factor into a product over items:

$$\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_n|\boldsymbol{a}\right) = P_{\boldsymbol{\theta}}\left(\bigcap_{k=1}^{K} X_k = x_{nk}|\boldsymbol{a}\right) = \prod_{k=1}^{K} P_{\boldsymbol{\theta}}\left(X_k = x_{nk}|\boldsymbol{a}\right), \tag{27.1}$$

where $\boldsymbol{x}_n = (x_{n1}, \ \ldots \ , x_{nK})'$ is a $K \times 1$ vector that contains the observed item responses for case n, and $\boldsymbol{a} = (a_1, \ \ldots \ , a_D)'$ is a $D \times 1$ vector containing the latent attribute values. The subscript $\boldsymbol{\theta}$ is used to emphasize the dependence of the various probability distributions on a vector of freely estimated structural parameters $\boldsymbol{\theta}$, e.g., item intercepts or slopes.

When there is potential residual dependence among subsets of items, a useful strategy is to include additional random effects or latent variables (Cai, Yang, & Hansen, 2011; Gibbons & Hedeker, 1992). Let there be S mutually exclusive groups of items, indexed $s = 1, \ \ldots \ , S$, each dependent on at most one group/specific latent dimension η_s. With these additional random effects, conditional independence implies

$$\begin{aligned}\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_n|\boldsymbol{a},\eta_1,\ldots,\eta_S\right) &= P_{\boldsymbol{\theta}}\left(\bigcap_{k=1}^{K} X_k = x_{nk}|\boldsymbol{a},\eta_1,\ldots,\eta_S\right)\\ &= \prod_{s=1}^{S}\prod_{k\in\mathcal{I}_s} P_{\boldsymbol{\theta}}\left(X_k = x_{nk}|\boldsymbol{a},\eta_s\right),\end{aligned} \tag{27.2}$$

where $\mathcal{I}_s$ is a notational shorthand to stand in for the index to the group of items that load on group/specific dimension s.

Furthermore, assume that the group/specific dimensions are conditionally independent given the attributes, i.e., $g_{\boldsymbol{\theta}}(\eta_1, \ldots, \eta_S|\boldsymbol{a}) = g_{\boldsymbol{\theta}}(\eta_1|\boldsymbol{a})g_{\boldsymbol{\theta}}(\eta_2|\boldsymbol{a}) \ldots g_{\boldsymbol{\theta}}(\eta_S|\boldsymbol{a})$. This allows us to employ a dimension reduction technique to integrate the η's out without a full S-dimensional integral (see Cai et al., 2011; Rijmen, 2009):

$$\begin{aligned}\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_n|\boldsymbol{a}\right) &= \int \prod_{s=1}^{S}\prod_{k\in\mathcal{I}_s} P_{\boldsymbol{\theta}}\left(X_k = x_{nk}|\boldsymbol{a},\eta_s\right) g_{\boldsymbol{\theta}}\left(\eta_1,\ldots,\eta_S|\boldsymbol{a}\right) d\eta_1\cdots d\eta_S\\ &= \prod_{s=1}^{S}\int \prod_{k\in\mathcal{I}_s} P_{\boldsymbol{\theta}}\left(X_k = x_{nk}|\boldsymbol{a},\eta_s\right) g_{\boldsymbol{\theta}}\left(\eta_s|\boldsymbol{a}\right) d\eta_s.\end{aligned} \tag{27.3}$$

This leads to vastly reduced computing time for parameter estimation.

Following the LCDM framework, the specification of the item response probabilities is straightforward. If item k is scored dichotomously, a possible specification is

$$P_{\boldsymbol{\theta}}\left(X_k = 1|\boldsymbol{a},\eta_s\right) = \frac{1}{1+\exp\left[-\left(\alpha_k + \boldsymbol{\beta}_k' h_k\left(\mathbf{q}_k,\boldsymbol{a}\right) + \gamma_k\eta_s\right)\right]}, \tag{27.4}$$

where $h_k(\mathbf{q}_k,\boldsymbol{a})$ is a vector-valued function that can generate the necessary main effects and interactions for this item, using the Q-matrix specification in $\mathbf{q}_k$, and α, $\boldsymbol{\beta}$, and γ are item parameters. For example, an item that assumes a DINA-like conjunctive response process but also depends on an additional random effect may be written as

$$P_{\boldsymbol{\theta}}\left(X_k = 1|\boldsymbol{a},\eta_s\right) = \frac{1}{1+\exp\left[-\left(\alpha_k + 0a_1 + 0a_2 + \beta_k a_1 a_2 + \gamma_k\eta_s\right)\right]}, \tag{27.5}$$

where the two main effects are fixed to 0 and the second-order interaction term is freely estimated. flexMIRT provides specific commands that implement DCM-specific restrictions.

As in de la Torre and Douglas's (2004) analysis, we can model the association among the latent attributes with higher order latent variables. For example,

for dichotomous attributes, we may use a multidimensional extension of the 2-parameter logistic model (Reckase, 2009):

$$P_{\boldsymbol{\theta}}\left(A_d = 1|\boldsymbol{\xi}\right) = \frac{1}{1 + \exp\left[-\left(\lambda_{d0} + \lambda_{d1}\xi_1 + \cdots + \lambda_{dm}\xi_m\right)\right]}. \tag{27.6}$$

Again, under the conditional independence assumption, the distribution of the attributes is

$$\pi_{\boldsymbol{\theta}}\left(\boldsymbol{a}|\boldsymbol{\xi}\right) = \prod_{d=1}^{D}\left[P_{\boldsymbol{\theta}}\left(A_d = 1|\boldsymbol{\xi}\right)\right]^{a_d}\left[1 - P_{\boldsymbol{\theta}}\left(A_d = 1|\boldsymbol{\xi}\right)\right]^{1-a_d}. \tag{27.7}$$

Combining $\pi_{\boldsymbol{\theta}}(\boldsymbol{x}_n|\,\boldsymbol{a})$ from Eq. (27.3) with $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\,\boldsymbol{\xi})$ from Eq. (27.7), the contribution to marginal likelihood from observed response pattern $\boldsymbol{x}_n$ is:

$$\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_n\right) = \int\left[\int \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_n|\boldsymbol{a}\right)\pi_{\boldsymbol{\theta}}\left(\boldsymbol{a}|\boldsymbol{\xi}\right)d\boldsymbol{a}\right] g_{\boldsymbol{\theta}}\left(\boldsymbol{\xi}\right)d\boldsymbol{\xi}, \tag{27.8}$$

where the inner integral in the brackets is a 2^D-term summation over the attribute profile probabilities. The marginal log-likelihood based on all observed response patterns is.

$$l\left(\boldsymbol{\theta}\right) = \sum_{n=1}^{N}\log\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_n\right). \tag{27.9}$$

In addition to fitting a wide variety of DCMs, flexMIRT also has rich statistical features, implemented as part of more general purpose IRT routines that can enhance DCM modeling. For instance, flexMIRT is natively multiple-group ready and has features that permit likelihood-based hypothesis testing, e.g., with likelihood ratio tests or Wald tests. Arbitrary user-supplied constraints may be applied to item parameters. Prior distributions may be provided as well. Further, there are numerous standard error estimation methods for users to choose from with DCM-specific recommendations provided in the flexMIRT documentation.

Due to its origins as a more general purpose IRT/item factor analysis program, flexMIRT also includes many item fit and overall model fit indices. Overall model fit values (−2 log likelihood, AIC, BIC, and when appropriate, the likelihood ratio (G^2) and Pearson X^2 full-information fit statistics) are reported by default and users may optionally request additional model fit values, such as the M_2 family of limited information measures (e.g., Maydeu-Olivares & Joe, 2005; Cai & Hansen, 2013) and the associated RMSEA (e.g., Steiger & Lind, 1980), and Tucker-Lewis Index (Tucker & Lewis, 1973) values. Item fit statistics are available as well, including sum-score based chi-square values (e.g., Orlando & Thissen, 2000; Cai, 2015) and marginal and bivariate chi-square values (Chen & Thissen, 1997).

Finally, flexMIRT can be used to simulate data from any model that it supports. While flexMIRT is a stand-alone program, it can be incorporated into other statistical environments via shell or system calls; analyses can be initiated from outside programs (e.g., R, SAS) making flexMIRT particularly convenient for investigating statistical properties of DCMs via simulation studies where many repeated analyses need to be completed.

As noted previously, flexMIRT is general purpose item response modeling software, but DCM-specific syntax has been incorporated to allow for a more intuitive translation from model to software syntax. We will highlight this aspect of the program with an example.

27.3 Example

Our demonstration of flexMIRT replicates the de la Torre and Douglas (2004) presentation of analyses using the well-known Tatsuoka (2002) fraction subtraction data set. In this example, a DINA model with a higher-order latent variable will be estimated using responses from 536 students (a publicly available subset of the full N = 2144 dataset) to 20 items. For the interested reader, the data set used, and complete flexMIRT syntax and all output files are available on the flexMIRT support page (https://www.vpgcentral.com/software/irt-software/support/).

The Q-matrix (available in de la Torre and Douglas [2004, p. 347]) was such that eight attributes were specified, with items assigned to as many as five attributes. Below we present an excerpt of the flexMIRT syntax for defining this analysis, focusing on aspects specific to the DCM:

```
 <Groups>
      %G%
        Attributes = 8;
        Generate = (4,6,7),(4,7),(2,3,5,7),(2,4,7,8),(1,2,7),
       (2,5,7,8),(2,5,7),(7,8),(2,4,5,7),
         (2,7),(1,7),(2,5,6,7),(1,2,3,5,7),(2,3,5,7);
       %D%
        DM = G;
        Varnames = a1-a8;
<Constraints>
      Fix G,(v1-v20),Slope;
      Free G,(v1),Interaction(4,6,7); // 3rd-order int of attr
     4,6,7
      Free G,(v2),Interaction(4,7); // 2nd-order int of attr 4,7
      Free G,(v3),Interaction(4,7); // 2nd-order int of attr 4,7
      Free G,(v4),Interaction(2,3,5,7); // 4th-order int of attr
     2,3,5,7
      Free G,(v5),Interaction(2,4,7,8); // 4th-order int of attr
     2,4,7,8
      Free G,(v6),MainEffect(7); // main effect of attr 7
     Equal D,(a1-a8),Slope; // constraint for "restricted" higher
     order model
```

In the <Groups> section of the syntax, group %G% will hold the observed data/item responses and group %D% will be used to model the higher-order space, using the attributes constructed within group %G% as the data for analysis. As noted, the number of attributes/main effects for the model was eight, justifying the Attributes = 8; statement of the presented flexMIRT syntax. The Generate statement is then used to construct only those higher-order interactions that will be used in the analysis, as determined by the Q-matrix specifications. The DM = G; statement is used to indicate that the DCM attribute profile probabilities defined in group %G% are modeled in the higher-order portion of the model. In other words, in specifying the higher-order portion of the model, the attributes will be treated as "items" to be fit with the specified model, hence the Varname = a1-a8; statement in group %D%.

In the <Constraints> section, the first line of syntax is to "reset" the loading pattern of items onto attributes (so items begin by loading on no attributes) and we then provide statements for the first 6 items to demonstrate how items are assigned to attributes. Given the desired DINA model, the highest-order interaction for each item is the key term, with all lower-order interactions and main effect parameter being fixed at 0 (and therefore set appropriately by the initial Fix statement). Finally, we specify that the slopes/loadings of the attributes onto the higher-order continuous latent variable are constrained to be equal, mirroring the restricted higher-order model of de la Torre and Douglas (2004). There is additional coding in the full syntax file to assign the remaining items to attributes, but we feel the above give readers a general sense of the intuitive manner, using relatively simple syntax, by which DCMs are specified in flexMIRT, even those with complex structured latent variable portions.

Even using only the available subset of data, the flexMIRT estimates of the slipping and guessing parameters for the items were extremely similar to those originally reported by de la Torre and Douglas (2004). The flexMIRT analysis completed in approximately 45 s on a personal computer (dual-core processing at 2.8 GHz with 4GB of RAM); this processing time included the item parameter estimation, estimation of the model-implied proportions of attribute profile memberships, estimates of latent class memberships for each observation, the default "item" and test information function values, as well as an estimate of the marginal reliability for the higher-order portion of the model and, lastly, default and additionally requested overall model fit values and optionally requested individual item fit/local dependence indices. This highlights the efficiency that can be achieved when estimating DCMs through MML estimation in a comprehensive item analysis package.

27.4 Discussion

flexMIRT® is a commercial general purpose multilevel and multidimensional item factor analysis software program that also implements specific features to accommodate DCM analysis. Its major advantage is the integration of DCM analysis with its rich statistical features available for general IRT modeling. For academic

researchers who teach, flexMIRT is currently offered freely for students, with full functionality. For researchers and users at operational assessment programs, flexMIRT's ability to handle operational-grade analysis while offering flexibility as a research tool may be an attractive feature.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Cai, L. (2015). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika, 80*, 535–559.

Cai, L. (2017). *flexMIRT® version 3.5: Flexible multilevel multidimensional item analysis and test scoring [Computer software]*. Chapel Hill, NC: Vector Psychometric Group, LLC.

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application, 3*, 297–321.

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*, 245–276.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full information bifactor analysis. *Psychological Methods, 16*, 221–248.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57*, 423–436.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of fit testing of diagnostic classification item response theory models. *British Journal of Mathematical and Statistical Psychology, 69*, 225–252.

Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009–1020.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50–64.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Rijmen, F. (2009). *Efficient full-information maximum likelihood estimation for multidimensional IRT models* (Technical Report No. RR-09-03). Princeton, NJ: Educational Testing Service.

Steiger, J. H., & Lind, J. M. (1980, June). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics), 51*, 337–350.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

Chapter 28
Using *Mplus* to Estimate the Log-Linear Cognitive Diagnosis Model

Meghan Fager, Jesse Pace, and Jonathan L. Templin

Abstract In this chapter, we present the software package *Mplus* (Muthén LK, Muthén BO, Mplus User's Guide. 8th edn. Los Angeles, Muthén & Muthén. https://www.statmodel.com/, 2017) with the Log-linear Cognitive Diagnosis Model (LCDM), a general model for diagnostic assessment (Henson RA, Templin JL, Willse JT, Psychometrika, 74(2):191–210, 2009; see also Chap. 8 in this volume). We devote most of this chapter to presenting relevant features of the LCDM as implemented in *Mplus* with a conceptual example using fraction subtraction data (Tatsuoka KK, Analysis of errors in fraction addition and subtraction problems, Report NIE-G-81-0002. University of Illinois, Computer-based Education Research Library, Urbana, 1984, Tatsuoka C, J R Stat Soc Series C (Appl Stat), 51:337 350. https://doi.org/10.1111/1467-9876.00272, 2002) to illustrate syntax composition, output evaluation, and assessment refinement.

28.1 Introduction

In this chapter, we present the software package *Mplus* (Muthén & Muthén, 2017) with the Log-linear Cognitive Diagnosis Model (LCDM), a general model for diagnostic assessment (Henson, Templin, & Willse, 2009; see also Chap. 8 in this volume). We devote most of this chapter to presenting relevant features of the LCDM as implemented in *Mplus* with a conceptual example using fraction

M. Fager
National University, Precision Institute, La Jolla, CA, USA
e-mail: mfager@nu.edu

J. Pace
University of Kansas, Lawrence, KS, USA
e-mail: Jesse.Pace@ku.edu

J. L. Templin (✉)
Educational Measurement and Statistics Program, University of Iowa, Iowa City, IA, USA
e-mail: jonathan-templin@uiowa.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_28

subtraction data (Tatsuoka, 1984, 2002) to illustrate syntax composition, output evaluation, and assessment refinement.

As you have learned from earlier chapters in this volume, diagnostic models are confirmatory latent class models of discrete latent traits (e.g., mastery/non-mastery). These models require specific constraints to align with model interpretations: (1) a pre-specified number of attributes or skills measured by the assessment, (2) a Q-matrix (i.e., item-by-skill association mapping; Tatsuoka, 1985), and (3) model constraints, such as equality restrictions and monotonically increasing item response functions (i.e., masters of skills have a higher probability of correct item responses than non-masters). Imposing further constraints on LCDM item parameters allows users to estimate other models subsumed by the LCDM, such as the DINA, DINO, NIDO, and C-RUM. For *Mplus,* we specify syntax and model estimation procedures following latent class model conventions with these implied constraints. The information presented in this chapter is meant to describe the sections of *Mplus* syntax and translate the traditional terminology of latent class models into the necessary components in estimating the LCDM.

28.2 Software Overview

Mplus is software for estimating models with observed and latent variables with multiple types of distributions available. Though presented here in the context of diagnostic models, *Mplus* is used for a variety of research situations that often arise in the social and behavioral sciences. In addition to more basic analyses, *Mplus* offers advanced capabilities such as multiple imputation for missing data, Monte Carlo simulation, and bootstrapped standard errors for estimated model parameters. Moreover, *Mplus* offers several measures of model fit that aid in evaluating models and deciding between competing models. *Mplus* estimation methods also accommodate both Frequentists' and Bayesians' perspectives by offering both maximum likelihood and Bayesian estimators with Markov chain Monte-Carlo methods.

Users can use *Mplus* on Windows, Mac, and Linux operating systems. Single user licenses start at $695 (at the time of writing) for base versions and are discounted to $195 for students. University licenses are also available and start at $595. For added features, *Mplus* offers package add-ons for an additional cost, such as the mixture add on (needed for the LCDM), a multi-level add-on for complex data structures, and a combination add-on. Software can be purchased on DVD or as a digital download. For ordering information and the *Mplus* user guide, we refer readers to the *Mplus* website (Muthén & Muthén, 2017).

28.3 Example Data

The fraction subtraction test has frequently appeared in diagnostic assessment publications (e.g., DeCarlo, 2011; de la Torre, 2008, 2009; de la Torre & Douglas, 2008; Chiu & Douglas, 2013; Mislevy, 1996; Tatsuoka, 1984, 2002; Templin &

Henson, 2006), and is publicly available. To download the data and Q-matrix, users must use R software (R Core Team, 2017) with the *CDM* package (Robitzsch, Kiefer, George, & Ünlü, 2018). This dataset is composed of 536 middle school students' responses to 20 test items. The test items measure skills such as "simplify before subtracting" and "convert a whole number to a fraction" and are Bernoulli distributed (i.e., scored correct or incorrect). Although the fraction subtraction data has eight skills, we restrict our analysis to two skills: borrow from a whole number part (alpha 5) and subtract numerators (alpha 7). This results in $2^2 = 4$ possible skill profiles depending on the mastery status of our skills (e.g., masters and non-masters). We also limit our analysis to seven items instead of the complete set for ease of exposition.

28.4 *Mplus* Syntax

Next, we describe core *Mplus* syntax for estimating the LCDM item, structural (saturated log-linear structural model), and respondent parameter estimates in the context of our example data. We will start with a brief introduction to basic concepts involved in using *Mplus* followed by a brief description of composing syntax files and the different options available for estimating the LCDM.

28.4.1 Basic Concepts

Mplus uses standard ASCII-text data files for input; common formats accepted include '.csv' and '.dat' data types. Input syntax files, like data files, are ASCII-text files. These input files can be created in a text editor. When using *Mplus* as the input file editor, the program interface allows the user to view syntax with a coloring scheme for *Mplus* key terms such as commented lines, key commands, and user input (e.g., variable names). Commented lines can be included anywhere in the input syntax and are designated by an exclamation point at the line start. Though general text files can continue lines indefinitely, *Mplus* requires that no line exceed 90 characters in width, and that variable names be no longer than eight characters. Furthermore, *Mplus* requires that syntax statements end with a semicolon to separate command options, with the exception of the title.

28.5 Command Syntax

The core sections of *Mplus* syntax, termed commands, have a specific purpose and associated set of options. There are a total of ten syntax commands, but *Mplus* choses specific defaults for several of these commands to minimize user input for common types of analyses. In our case, we rely on six preliminary commands to set up our analysis: (1) TITLE, (2) DATA, (3) VARIABLE, (4) ANALYSIS, (5)

OUTPUT, and (6) SAVEDATA, and a primary MODEL command specifically for the LCDM. For a more detailed description of other commands and options not listed here, we refer the reader to the *Mplus* User's guide (Muthén & Muthén, 2017).

28.5.1 TITLE Command

At the beginning of the syntax file, we can name our analysis with the TITLE command. A useful convention to adopt is titling each analysis according to (1) the dataset used, (2) the number of skills and items, and (3) the model applied. If we chose to exclude higher-order interaction terms, we would note this here. This convention helps in keeping an organized record of models run since *Mplus* requires individual input files per model estimated. For our example dataset, we use the title: "LCDM for inputdata.dat with 2 skills and 7 items: Saturated structural model with 2 levels of interactions." After we run our input syntax file, our title will appear at the beginning of our output file.

28.5.2 DATA Command

Mplus model syntax uses the DATA command to specify the data file and its structure. If the input file is contained in the same directory as the data file, *Mplus* defaults to the same folder and the option FILE = only needs the data file name. Otherwise, the file location can be specified with the path to the directory on the hard disk (e.g., FILE = "C:\LCDM\input.dat"). The structure of the data file requires that any column names or headers be removed. The default importing format of data is free-format rows which can be changed with the FORMAT option to describe other types of formatting schemes, such as a fixed format (i.e., equal spaced columns).

28.5.3 VARIABLE Command

The next section of the syntax file is the VARIABLE command. This command is required to provide *Mplus* with observed variable names, observed variable types (e.g., categorical, nominal), and observed variables involved in the analysis. In our example dataset, our complete data has twenty items. Under the VARIABLE command, the syntax NAMES = ID X1-X20 specifies a unique identifier of respondents and all variables in the data file. However, since we are only using a subset of the items in our example, we limit our USEVARIABLES statement to list only the items we wish to use. Because our data are scored dichotomously (0/1), we have categorical item responses which *Mplus* labels as CATEGORY 1 for

values of 0 and CATEGORY 2 for values of 1. This option would be specified as TYPE = CATEGORICAL in a subsequent line after we define our variables.

The final option we include in our VARIABLE command section is the number of latent classes. Recall that as constrained latent class models, the LCDM specifies the number of classes according to the number of skills and the number of levels they assume. For our data with two skills, this results in $2^2 = 4$ latent classes because we have two mastery statuses. The syntax CLASSES = c(4) specifies the class label as "c" and the number of classes as (4). If we had missing data codes, we would also specify a missing value code (e.g., −999).

28.5.4 ANALYSIS Command

Our next command in our *Mplus* input file is the ANALYSIS command. Here, we can indicate estimation options and the type of model to run. Since constrained latent class models such as the LCDM fall under the framework of finite mixture models (e.g., McLachlan & Peel, 2000), the option noted here is TYPE = MIXTURE. Otherwise, we will use the default options for our analysis: robust (marginal) maximum likelihood (ML) estimation and automatically generated starting values for all parameters.

28.5.5 OUTPUT and SAVEDATA Commands

The final two commands to set up our analysis are the OUTPUT and SAVEDATA commands to obtain additional needed output for our model. For OUTPUT, we request the TECH10 option for added model fit information. We also request that each respondent's estimated class membership probabilities be stored in a separate file by specifying the FILE = option under the SAVEDATA command.

28.5.6 The MODEL Command: Syntax for the LCDM

The next and final command necessary to estimate the LCDM with *Mplus* is the MODEL command. *Mplus* uses this command to define LCDM latent classes, item parameters, and order constraints. Syntax first defines each latent class by the items' class-specific response probabilities, however *Mplus* does not directly specify item response probabilities as defined in the LCDM. Instead, as a method to analyze categorical data, *Mplus* uses thresholds to model each item's different functioning between classes. Thresholds provide an intercept in the model for the probability of each observed item category, starting with the lowest and ending with one less than the total number of categories.

Thresholds are modeled on the logit metric. Since our items are scored with two categories, there is only one threshold per item to represent the difference between an incorrect and correct response. The item parameter thresholds are constrained to be equal between different latent classes to coincide with the LCDM model interpretation. That is, the LCDM keeps item response probabilities equal for respondents in different latent classes.

The *Mplus* syntax for the MODEL command, %c#1%, indicates that all following commands correspond to latent class 1 (i.e., $c_1 = [0, 0]$). Each item has a line of syntax per each class that specifies the threshold and its corresponding label: [x1$1] (T1_1). The first term contained in the bracket refers to the threshold, $1, for an item, x1. The second term within parentheses (T1_1), is arbitrarily assigned to label the threshold. This label is then used to allow the thresholds to be predicted by the corresponding item parameters involved.

To map thresholds to the LCDM, we begin by predicting the threshold from the LCDM item parameters present, as a function of each latent class. To translate thresholds to LCDM item response probabilities, the sum of the item parameters predicting the threshold is multiplied by negative one to model the probability of a one (i.e., a correct response) rather than the probability of a zero, the *Mplus* default. To illustrate the MODEL CONSTRAINT syntax, we will use item 10 as an example. Since item 10, X_{10}, measures both skills contained in skill profile $\boldsymbol{a_c}$, there are four different combinations of possible response probabilities that depend on the attribute profile:

$$P(\alpha_1) = \frac{exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(0) + \lambda_{10,1,(7)}(0) + \lambda_{10,2,(5*7)}(0)(0)\right)}{1 + exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(0) + \lambda_{10,1,(7)}(0) + \lambda_{10,2,(5*7)}(0)(0)\right)} \quad (28.1)$$

$$P(\alpha_2) = \frac{exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(0) + \lambda_{10,1,(7)}(1) + \lambda_{10,2,(5*7)}(0)(1)\right)}{1 + exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(0) + \lambda_{10,1,(7)}(1) + \lambda_{10,2,(5*7)}(0)(1)\right)} \quad (28.2)$$

$$P(\alpha_3) = \frac{exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(1) + \lambda_{10,1,(7)}(0) + \lambda_{10,2,(5*7)}(1)(0)\right)}{1 + exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(1) + \lambda_{10,1,(7)}(0) + \lambda_{10,2,(5*7)}(1)(0)\right)} \quad (28.3)$$

$$P(\alpha_4) = \frac{exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(1) + \lambda_{10,1,(7)}(1) + \lambda_{10,2,(5*7)}(1)(1)\right)}{1 + exp\left(\lambda_{10,0} + \lambda_{10,1,(5)}(1) + \lambda_{10,1,(7)}(1) + \lambda_{10,2,(5*7)}(1)(1)\right)} \quad (28.4)$$

where α_1, α_2, α_3 and α_4 are the latent classes (vectors indicating attribute mastery) [0,0], [0,1], [1,0], and [1,1], respectively.

Here, we have four different thresholds for item X_{10} because we have four different probabilities of correctly responding to the item, $P(X_{10} = 1|\,\boldsymbol{\alpha_c})$. These item response probabilities are functions of the parameters contained within the equation: the intercept, $\lambda_{0,10}$, two main effects for each skill, $\lambda_{10,1,(5)}$ and $\lambda_{10,1,(7)}$, and the interaction term, $\lambda_{10,2,(5_*7)}$, included when both skills are mastered. Thresholds, like item response probabilities, are generated according to the different skill statuses: t10_1 for class 1, t10_2 for class 2, and so on. For a factorially simple item, such as Item 1, only two thresholds would be generated as there are only two possible item

response probabilities involving two possible LCDM parameters: an intercept, $\lambda_{i,0}$ and a main effect, $\lambda_{i,1(\alpha)}$, for the skill measured by the item.

28.6 *Mplus* Output and Results

Upon completing the syntax file, the analysis can be completed in *Mplus* in the following ways: using the RUN button found atop the graphical interface menu bar or in batch mode via a Windows command prompt or a Linux/MacOS terminal. As it runs, messages will appear describing the calculation of the model, either in a secondary window or in the terminal/command prompt, until successful completion of the estimation. The output is returned in a new file with a ".out" file type that contains the model results.

At the top of the output file, the title and full syntax of the input file precedes the results of the analysis. If there were any errors, they will be noted after the syntax and should be diagnosed before proceeding. We next describe how to interpret results and present key findings of our example analysis.

28.6.1 Model Fit Information

After *Mplus* terminates, it gives verbose output about the data characteristics and the estimated model. But before proceeding, it is necessary to ensure that our data adequately fits our proposed model. For our example, we have estimated 27 parameters from 536 respondents. Our loglikelihood, labeled H0 Value, is -1579.995 and scaled by robust maximum likelihood with a correction factor of 1.011 to augment likelihood calculations. Our relative fit statistics for model comparisons are AIC = 3213.99, BIC = 3329.66, and sample-size adjusted BIC = 3242.96. These can be used for non-nested model comparisons, or the loglikelihood can be used to make nested model comparisons with a likelihood ratio test. This is done by performing Chi-square tests of the difference in -2 loglikelihoods, with the difference in the model degrees of freedom between the models under comparison as the test degrees of freedom. When robust ML is used for estimation, a scaled version of the likelihood ratio test can be found via methods described on *Mplus*'s website.

The next piece of model fit information is the Chi-square test of model fit for the binary and ordered categorical (ordinal) outcomes. Here, the Pearson Chi-square and likelihood ratio Chi-square tests are tests of global model fit that compare the current model to the saturated model where each response pattern has an estimated probability. The degrees of freedom are the same for both tests and equal 100 in our example. This number is found by taking the total response patterns possible (i.e., $2^7 = 128$ for 7 items) minus the number of estimated parameters (i.e., 27). The null hypothesis for both tests is that the model fits the data as well as the saturated model. We found that both tests were nonsignificant, therefore we retained the null

and concluded that our model fits our data according to absolute and relative fit statistics.

Another important piece of model fit information comes from the univariate and bivariate tests requested from the TECH10 option of the OUTPUT command. The bivariate section is of primary interest for the LCDM to evaluate local model fit for each pair of items. Compared to global tests of model fit which rely on all possible response options, bivariate fit can help diagnose specifically which items are causing misfit.

Bivariate information is evaluated using two-way contingency tables for the response options of a pair of items. These are used to compare the observed frequencies with those expected from the model. The hypothesis test for each pair of items is a one degree of freedom Chi-square test with the null hypothesis being that the model predicted values from the item parameters are equal to those observed in the data. In our example, the bivariate Pearson Chi-square and Log-Likelihood Chi-square tests for all pairs of items were nonsignifincant, indicating that we have another piece of evidence supporting the fit of our model. The overall test aggregates the bivariate test statistics from each pair of items to a total Pearson chi-square of 3.670 and a log-likelihood chi-square of 3.694. Since these values are far below the number of pairs of items we have in the test, (i.e., 21 pairs), we can conclude that our model fits the data well.

28.6.2 Final Class Counts and Estimated Proportions

The next section of the *Mplus* output offers the number of estimated respondents in each latent class and a converted proportion. Here, we can tell that most of our sample is expected to have either mastered all or none of the skills because most respondents fall in either class 1 ($\alpha_1 = [0, 0]$; 38%) or class 4 ($\alpha_4 = [1, 1]$; 41%). The remaining 21% of our sample are predicted to have mastered one of the skills, but not the other.

28.6.3 New/Additional Parameters

The next section of interest is the new/additional parameters section. This section outputs our estimated item parameters (defined in the MODEL CONSTRAINT section) that follow our LCDM parameterization. These parameters can be used to further refine the design of our assessment and continue model calibration. The output includes five columns, including the parameter labels we defined previously, the estimated parameter values, the standard error of the estimates, a test statistic for each parameter (EST./S.E.), and a two-tailed p-value determining whether the parameter is significantly different than zero (not-needed) or nonzero (needed). If the item parameter is nonsignificant (disregarding the intercept), this may indicate that the item does not measure the skill (i.e., item main effects), or there is not an

Table 28.1 Fraction Subtraction Subsetted Data Q-Matrix and LCDM Item Parameter Estimates

	Skill 5	Skill 7	$\lambda_{i,0}$	$\lambda_{i,1,(5)}$	$\lambda_{i,1,(7)}$	$\lambda_{i,2,(5,7)}$
Item 1	0	1	−2.933*		5.201*	
Item 2	0	1	−2.667*		6.316*	
Item 10	1	1	−5.085*	5.116*	3.212*	−1.760
Item 11	1	1	−2.641*	4.247*	0.522	0.785
Item 17	1	1	−3.411*	4.635*	0.980	−0.024
Item 18	1	1	−3.108*	4.593*	2.620*	−2.338*
Item 20	1	1	−4.936*	6.178*	0.000	0.480

*$p < 0.01$

additional bump in the probability of a correct response for having mastered both skills measured by the item (i.e., item interactions).

Results of our example analysis are presented alongside the Q-matrix in Table 28.1. As indicated by the *p*-values, most parameters are significant, however four interactions are nonsignificant. This indicates that the interaction term does not contribute to our model because the probability of correctly responding to the item is not enhanced when the interacting skills measured are both mastered. We can remove these terms and keep the main effects as they are sufficient for describing our data.

For example, in light of the nonsignificant interaction term for item 10, we may have specified the item 10 thresholds as follows, with a strikethrough indicating what would be deleted to remove the nonsignificant interaction term:

```
! ITEM 10:
  NEW (l10_0 l 10_12 l10_11 ~~l10_212~~);
   t10_1=-(l10_0); !intercept
   t10_3=-(l10_0+l10_11); !intercept + main effect for
  first attribute (skill 5)
   t10_2=-(l10_0+l10_12); !intercept + main effect for
  second attribute (skill 7)
   t10_4=-(l10_0+l10_11+l10_12~~+l10_212~~); !intercept + two
  main effects ~~+ interaction~~
```

Removing the interaction term in the definition of new parameters and the summation that defines the thresholds is all that is necessary. In the case of some items, we could also remove nonsignificant main effects, such as for skill 7 for items 11, 17, and 20, because having mastery of skill 7 does not contribute significantly to mastering the item. If both a main effect and an interaction were to be removed, such as in item 11, then all that would remain is the single main effect. The entire t11_2 and t11_4 would be deleted. Any calls to label t11_2 would be deleted as it is not estimated to measure attribute 7, and any calls to t11_4 (i.e., in profile [1,1]) would instead be labeled with t11_3, as that is all that remains of that effect.

Table 28.2 Estimated Respondent Parameters for Five Respondents

Response Pattern	α_{r1}	α_{r2}	α_{r3}	α_{r4}	$\max(\alpha_r)$
[0,0,1,1,0,1,1]	0.00001	0.00000	0.98217	0.01782	3
[0,1,1,1,1,1,1]	0.00000	0.00000	0.03691	0.96308	4
[0,1,0,0,0,0,0]	0.76235	0.23752	0.00010	0.00003	1
[1,1,1,1,0,1,1]	0.00000	0.00024	0.00055	0.99921	4
[0,0,0,0,0,0,0]	0.99930	0.00056	0.00013	0.00000	1

28.6.4 Saved Estimated Respondent Latent Class Memberships

The previous sections focused on model output relevant to diagnosing the fit of the model and its estimated parameters. Once this is achieved, we can examine the posterior probabilities of skill mastery for respondents. Table 28.2 lists entries from the exported file for five respondents. The seven elements of the vectors contained in the response pattern column correspond to the respondents scores on the seven items. Following this, estimated probabilities (i.e., EAPs) for each respondent are given for each latent class. The final column, $\max(\alpha_r)$, indicates which class is the most likely according to the highest probability of each class (i.e., MAP) for each respondent. These estimates are useful for giving individual, fine-grained feedback for respondents to assess their levels of mastery of multiple skills.

28.7 Discussion

DCMs are a family of latent class models which contain discrete latent variables; these variables are used to classify respondents. The purpose of this chapter has been to explicate the use of *Mplus* software in the estimation of DCMs. Other software exists for this purpose, such as flexMIRT (Houts, & Cai, 2016; Chap. 27 in the volume) and the CDM package in R (George, Robitzsch, Kiefer, Gross, & Ünlü, 2016; Chap. 26 in the volume).

As we have demonstrated in this chapter, *Mplus* can be used to model one of the most general of DCMs, the LCDM, with relatively simple input from the user. We highlighted the core syntax involved in estimating the LCDM as well as the primary output of interest for our model and purpose. We note once more that it is a relatively simple matter to impose constraints in *Mplus* on the LCDM to generate more restrictive DCMs such as DINA or DINO. Due to both the relatively easy input as well as its flexibility in modeling both general and constrained DCMs, the *Mplus* software as presented here is a valuable resource for any researcher interested in using DCMs in their work.

References

Chiu, C.-Y., & Douglas, J. A. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification, 30*, 225–250. https://doi.org/10.1007/s00357-013-9132-9

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*, 343–362. https://doi.org/10.1111/j.17453984.2008.00069.x

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130. https://doi.org/10.3102/1076998607309474

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595–624. https://doi.org/10.1007/s11336-008-9063-2

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement, 35*, 8–26. https://doi.org/10.1177/0146621610377081

George, A. C., Robitzsch, A., Kiefer, T., Gross, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*, 1–24. https://doi.org/10.18637/jss.v074.i02

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Houts, C. R., & Cai, L. (2016). *flexMIRT R user's manual version 3.5: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416. https://doi.org/10.1111/j.17453984.1996.tb00498.x

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén. https://www.statmodel.com/

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2018). CDM: Cognitive diagnosis modeling. *R package version 6*. 2–91. https://CRAN.R-project.org/package=CDM

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics), 51*, 337–350. https://doi.org/10.1111/1467-9876.00272

Tatsuoka, K. K. (1984). *Analysis of errors in fraction addition and subtraction problems* (Report NIE-G-81-0002). Urbana, IL: University of Illinois, Computer-based Education Research Library.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconception by the pattern classification approach. *Journal of Educational Statistics, 10*, 55–73. https://doi.org/10.2307/1164930

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Chapter 29
Cognitive Diagnosis Modeling Using the GDINA R Package

Wenchao Ma

Abstract The GDINA R package (Ma and de la Torre, GDINA: The generalized DINA model framework. R package version 2.3.2. Retrieved from https://CRAN.R-project.org/package=GDINA: 2019) provides psychometric tools for estimating a range of cognitive diagnosis models (CDMs) and conducting various CDM analyses. The package is developed in the R programming environment (R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/: 2018). This chapter describes the main features of the package and presents an exemplary analysis of data to illustrate the use of the package.

29.1 Introduction

The `GDINA` R package, (Ma & de la Torre, 2019) provides psychometric tools for estimating a range of cognitive diagnosis models (CDMs) and conducting various CDM analyses. The package is developed in the `R` programming environment (R Core Team, 2018) building on earlier work of de la Torre, who initially wrote many pieces of source code for CDM estimation and analyses in Ox (Doornik, 2009). These pieces of source code are integrated into the `GDINA` R package and have been further extended in various ways. The `GDINA` package is publicly available under the version 3 of the GNU General Public License and can be used free of charge on multiple platforms including Unix/Linux, Windows and Mac OS. By using a few additional R packages, data and Q-matrix prepared using ASCII, Excel, SPSS, SAS, STATA, or other popular statistical software programs, can be easily imported into `R`, and analyzed using the `GDINA` package. The goal of this chapter is twofold: (1) to provide an overview of the `GDINA` package, and (2) to present an exemplary analysis of data to illustrate the use of the package.

W. Ma (✉)
The University of Alabama, Tuscaloosa, AL, USA
e-mail: wenchao.ma@ua.edu

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_29

29.2 Psychometric Models

The G-DINA model (de la Torre, 2011; Chap. 7 in this volume) and its extensions are the bases of the `GDINA` package, hence its name. The `GDINA` package can calibrate the G-DINA model and many reduced models it subsumes, including the DINA model (Haertel, 1989), DINO model (Templin & Henson, 2006), reduced reparametrized unified model (RRUM; Hartz, 2002), *additive*-CDM (*A*-CDM; de la Torre, 2011), and linear logistic model (LLM; Maris, 1999). By defining design matrices and specifying link functions, new models within the G-DINA model framework can be estimated as well.

The `GDINA` package can also accommodate polytomous attributes and responses. Chen and de la Torre's (2013) polytomous G-DINA model can be used to handle expert-defined polytomous attributes, and Ma and de la Torre's (2016) sequential G-DINA model is available for calibration of ordinal and nominal responses. In either case, the G-DINA model can be further constrained to the reduced CDMs mentioned above. It is also straightforward to specify different CDMs for different items using the package.

Additionally, the `GDINA` package can accommodate independent, saturated, higher-order, loglinear smoothed and hierarchical attribute structures. For the saturated structure, the population proportions of latent classes are treated as parameters and estimated directly. For the higher-order structure (de la Torre & Douglas, 2004), the Rasch model, one-parameter logistic model with a common slope parameter, or two-parameter logistic model (see de Ayala, 2013) can be employed. For the loglinear approach, a loglinear model (Xu & von Davier, 2008) can be specified. For hierarchically structured attributes, linear, convergent, divergent, or unstructured attributes (Leighton, Gierl, & Hunka, 2004) can be specified.

Last, the `GDINA` package can fit multiple-group models (Ma, Terzi, Lee, & de la Torre, 2017), where different groups can have distinct attribute structures. The Bugs-DINA and DINO models (Kuo, Chen, Yang, & Mok, 2016) for diagnosing the presence of misconceptions, the multiple-strategy DINA model (de la Torre & Douglas, 2008) and the diagnostic tree model (Ma, 2019) for accommodating multiple strategies for dichotomous and polytomous responses, respectively, can also be calibrated. Note that the loglinear CDM (LCDM; Henson, Templin, & Willse, 2009; Chap. 8 in this volume) is equivalent to the G-DINA model in the logit link, and thus its item parameters can be obtained using the `GDINA` package as well. Von Davier (2014) shows that LCDM is a special case of the general diagnostic model (GDM; von Davier, 2008; Chap. 6 in this volume) and can also be estimated using `mdltm` (Chap. 30 in this volume).

29.3 Parameter Estimation

Item parameters of the CDMs discussed above are estimated via an expectation-maximization (EM) implementation of the marginal maximum likelihood estimation algorithm (Bock & Aitkin, 1981) in the `GDINA` package, which, for reduced CDMs, differs from the two-step procedure via the least square method introduced in de la Torre (2011). Users can specify monotonic constraints for item parameter estimation (i.e., mastering an additional attribute will not result in lower success probability). The E step of the EM algorithm, as well as other computationally intensive functions, was written in `C++` through the `Rcpp` (Eddelbuettel & Francois, 2011) and `RcppArmadillo` (Eddelbuettel & Sanderson, 2014) packages to speed up program execution. In the M step, various optimization techniques can be used depending on the fitted models and parameter constraints. Multiple starting values or user specified starting values can be used to minimize the likelihood of obtaining solutions based on local maxima. The `GDINA` package also allows users to estimate the parameters for some items while fixing the parameters for others. This is a common practice when some items need to be added to a calibrated item bank. Standard errors of item parameters can be estimated using the outer product of gradient approximations (Philipp, Strobl, de la Torre, & Zeileis, 2018), or bootstrap approaches. Person attribute profiles are estimated using maximum likelihood estimation (MLE), expected a posteriori (EAP), or maximum a posteriori (MAP) estimations (Huebner & Wang, 2011) after the item parameters are estimated or provided by users.

29.4 Other Statistical Procedures

Apart from model calibration, additional procedures are available in the `GDINA` package. First, the Q-matrix can be validated using de la Torre and Chiu's (2016) ς^2 method or Ma and de la Torre's (2019) stepwise method. Both methods can be used along with the G-DINA model and the sequential G-DINA model. The mesa plot (de la Torre & Ma, 2016) based on the proportion of variance accounted for (PVAF) by each candidate q-vector (de la Torre & Chiu, 2016) provides a way to visually pinpoint the best q-vector candidates for each item. Second, the Wald test, likelihood ratio (LR) test, or score test can be used to evaluate whether, for each item, the G-DINA model can be replaced by a reduced model without a significant loss of model-data fit (de la Torre & Lee, 2013; Ma, Iaconangelo & de la Torre, 2016; Sorrel, Abad, Olea, de la Torre, & Barrada, 2017a; Sorrel, de la Torre, Abad, & Olea, 2017b). In addition, absolute model-data fit can be evaluated using the M_2 statistic, RMSEA_2 and SRMSR (Maydeu-Olivares, 2013; Liu, Tian, & Xin, 2016). The log odds ratio and Fisher-transformed correlations (Chen, de la Torre, & Zhang, 2013) provide more detailed absolute fit information for item pairs, which may be used to identify the sources of misfit. The deviance, AIC, and BIC are

available for evaluating relative model-data fit. To compare nested models at the test level, the LR test can be employed. However, the LR test may not be valid if the least restrictive model is misspecified (Maydeu-Olivares & Cai, 2006). Furthermore, the Wald test (Hou, de la Torre, & Nandakumar, 2014) and LR test (Ma, Terzi, Lee, & de la Torre, 2019) are available for detecting differential item functioning. Last, to be more accessible, the package offers a graphical user interface via R package `shiny` (Chang, Cheng, Allaire, Xie, & McPherson, 2017), and a wrapper function for conducting the Q-matrix validation, item-level model selection, and model calibration sequentially in a single run. Table 29.1 summarizes the major features of the package.

Table 29.1 Summary of the features of the `GDINA` R package (version 2.3.2)

Model Structures	Measurement model
	G-DINA model, polytomous and sequential G-DINA models
	DINA, DINO, *A*-CDM, LLM, and RRUM
	Bugs-DINA and DINO models
	Models defined using design matrix and link function
	Multiple-strategy DINA model and generalized-multiple strategy CDMs
	Diagnostic tree model
	Structural model
	Independent model
	Saturated model
	Higher-order model
	Loglinear model
	Hierarchical model
Model Estimation	MMLE/EM (Item parameters)
	MLE, MAP, EAP (Person parameters)
Fit Indices	M_2, $RMSEA_2$, SRMSR
	Log odds ratio and Fisher-transformed correlation for item pairs
	Deviance, AIC, BIC
Model Comparison	LR test, AIC, BIC (Test level)
	Wald test, LR test, score test (Item level)
Q-matrix Validation	ς^2 method, stepwise method
	Mesa plot
Complex Sampling	Missing by design and at random
Designs	Multiple-group estimation
Other Features	DIF detection
	Classification accuracy evaluation
	Data simulation
	Graphical user interface

29.5 Real Data Illustration

Responses of 2922 students to 28 items measuring three attributes in the Examination for the Certificate of Proficiency in English Grammar section from 2003 to 2004 were analyzed using the `GDINA` package. This data set has been analyzed by many researchers (e.g., Liu, Douglas, & Henson, 2009; Templin & Bradshaw, 2014; von Davier, 2014; Chap. 26 in this volume).

For illustration, the Q-matrix was first empirically validated based on de la Torre and Chiu's (2016) ς^2 approach, which requires fitting the G-DINA model to the data using the current Q-matrix. Monotonic constraints were imposed for the model calibration because of the identifiability concerns (von Davier, 2014), and as expected, the results from the G-DINA model estimation using the `GDINA` function were virtually identical to these from the LCDM (Templin & Hoffman, 2013) and the GDM (von Davier, 2014) after parameter transformations because these models are equivalent. Based on the G-DINA estimates, the Q-matrix was validated using the `Qval` function. Results showed that the original q-vectors for Items 9 and 13 had PVAFs less than 0.95, indicating that they may need further examination. Figure 29.1 gives the mesa plots for these two items. The mesa plot is a line chart, where the x-axis is the q-vectors with the highest PVAF for different numbers of required attributes, whereas the y-axis gives the corresponding PVAFs. Note that $\mathbf{0} = (0,0,0)$ is not a valid q-vector, but is still shown on the x-axis for reference. The mesa plot is similar to the scree plot in factor analysis, and the q-vector on the edge of the "mesa" is believed the correct q-vector for the item (de la Torre & Ma, 2016). This makes specifying a cutoff for PVAF as in de la Torre and Chiu (2016) unnecessary. The mesa plots showed that the original q-vectors, which were indicated using solid red dots, were on the mesa edges for both items with PVAFs of about 0.9. Thus, we can reasonably believe that they are appropriate for these two items.

After validating the Q-matrix, several CDMs were refitted to the data. In addition to the DINA model and RRUM employed by Liu, Douglas, and Henson (2009),

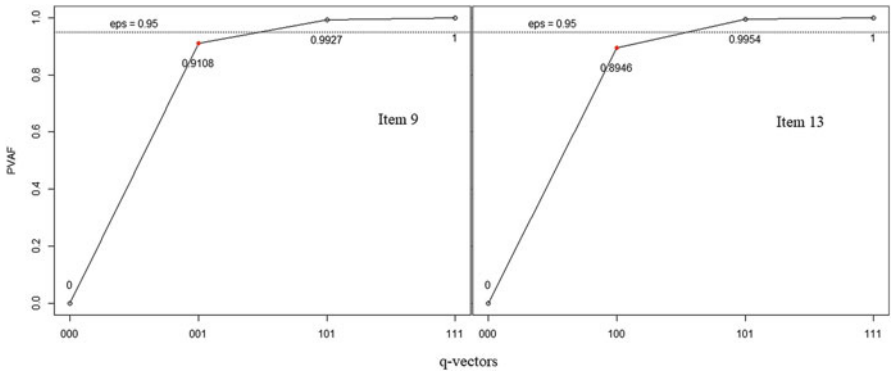


Fig. 29.1 Mesa plots for Items 9 and 13

the G-DINA model, DINO model, *A*-CDM, and LLM were considered as well. Furthermore, the Wald test was also used to examine whether the G-DINA model can be replaced by reduced models, and the corresponding p values for items requiring two or more attributes are given in Table 29.2. When multiple reduced models are retained for an item, the rule in Ma, Iaconangelo and de la Torre (2016) was adopted to determine the most appropriate model. Specifically, if the DINA or DINO model is retained, the one with larger p value is selected; if both DINA and DINO models are rejected, but any of the *A*-CDM, LLM and RRUM is retained, the one with the largest p value is selected; if all reduced models are rejected, the G-DINA model is used. As shown in Table 29.2, the DINO model, *A*-CDM, LLM and RRUM were selected as the most appropriate models.

According to AIC and BIC statistics, as shown in Table 29.3, the CDMs chosen by the Wald test are preferred. However, by assessing the absolute model-data fit using the `modelfit` function, the CDMs chosen by the Wald test produced a M_2 statistic of 528.30 ($df = 335$) with $p < 0.001$, indicating an inadequate model-data fit.

To further explore if any misfit can be identified at the level of item pairs, the `itemfit` function was used to calculate the log odds ratio and Fisher-transformed correlation based on observed and predicted item responses. Figure 29.2 displays

Table 29.2 The p values of the Wald test for item-level model comparison

	p values					
Item	DINA	DINO	*A*-CDM	LLM	RRUM	Selected model
1	.01	<.01	.39	.12	.48	RRUM
3	.02	<.01	.50	.38	.75	RRUM
7	<.01	<.01	.08	.66	<.01	LLM
11	<.01	<.01	.92	.27	.72	*A*-CDM
12	<.01	<.01	.01	.18	.49	RRUM
16	<.01	<.01	.13	.57	.02	LLM
17	.08	.13	.81	.96	.77	DINO
20	<.01	<.01	.02	.07	.39	RRUM
21	<.01	<.01	.43	.95	.23	LLM

Table 29.3 Test level model comparisons

CDMs	No. of Parameters	Likelihood	AIC	BIC
G-DINA	81	−42739.71	85641.42	86125.81
DINA	63	−42841.49	85808.98	86185.72
DINO	63	−42920.37	85966.75	86343.49
A-CDM	72	−42745.49	85634.98	86065.54
LLM	72	−42744.76	85633.51	86064.08
RRUM	72	−42745.64	85635.29	86065.85
CDMs selected By the Wald test	71	−42744.06	**85630.11**	**86054.69**

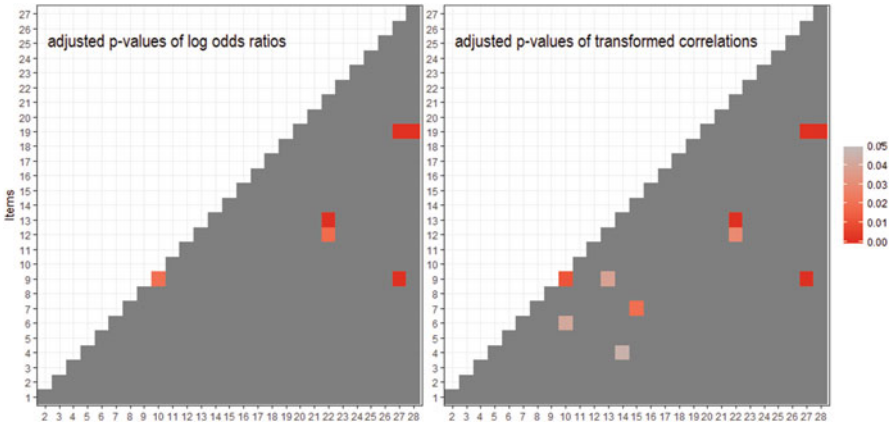


Fig. 29.2 Heatmap plots of adjusted p-values

a heatmap plot showing the *p*-values adjusted using the Bonferroni method. In the heatmap plot, x-axis and y-axis represent items, and the first item is dropped on x-axis and the last is dropped on y-axis. The adjusted *p* values for all item pairs are displayed in the lower right shading area, where gray squares represent *p* values greater than .05 (i.e., statistically adequate fit) and red squares represent *p* values less than .05 (i.e., statistically inadequate fit). Significant misfit can be observed for some pairs of items (e.g., items 9 and 27, and items 19 and 28), and thus may need further examination by domain experts.

29.6 Discussion

Given the growing interest in CDMs in recent years, the `GDINA` package aims to offer many ready-to-use functions to facilitate both research and operational work with CDMs. However, as any software, the package in its current form has some limitations. First, only the G-DINA model and its extensions, as well as CDMs they subsume, are available in this package. Researchers who intend to use other models may consider other software packages discussed in this handbook, such as the `CDM` package (Chap. 26 in this volume) or `mdltm` (Chap. 30 in this volume). Second, the model calibration using the `GDINA` package may fail when the number of attributes is very large. For example, on a workstation with 128 GB RAM, the package seems to be able to handle at most 22 attributes. Last, like most R packages, the `GDINA` package is still under development. Many functions, such as considering sample weights in a complex sample design, will be incorporated in the future.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443–459.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). *Shiny: Web application framework for R*. R package version 1.0.5, Retrieved from https://CRAN.R-project.org/package=shiny.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*, 419–437.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123–140.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

de Ayala, R. J. (2013). *The theory and practice of item response theory*. New York: Guilford Publications.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253–273.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595–624.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355–373.

de la Torre, J., & Ma, W. (2016, August). *Cognitive diagnosis modeling: A general framework approach and its implementation in R*. New York: A Short Course at the Fourth Conference on Statistical Methods in Psychometrics, Columbia University.

Doornik, J. A. (2009). *Object-oriented matrix programming using Ox (Version 6) [Computer software]*. London: Timberlake Consultants Press.

Eddelbuettel, D., & Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software, 40*, 1–18.

Eddelbuettel, D., & Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis, 71*, 1054–1063.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321.

Hartz, S. M. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practicality*. (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191–210.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*, 98–125.

Huebner, A., & Wang, C. (2011). A note on comparing examinee classification methods for cognitive diagnosis models. *Educational and Psychological Measurement, 71*, 407–419.

Kuo, B. C., Chen, C. H., Yang, C. W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology, 36*, 1115–1133.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The Attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205–237.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*, 579–598.

Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*, 3–26.

Ma, W. (2019). A diagnostic tree model for polytomous responses with multiple strategies. *British Journal of Mathematical and Statistical Psychology, 72*, 61–82.

Ma, W., & de la Torre, J. (2019). *An empirical Q-matrix validation method for the sequential G-DINA model*. Advanced online publication. https://doi.org/10.1111/bmsp.12156.

Ma, W., & de la Torre, J. (2019). GDINA: The generalized DINA model framework. *R package version*, 2.3.2. Retrieved from https://CRAN.R-project.org/package=GDINA

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection and attribute classification. *Applied Psychological Measurement, 40*, 200–217.

Ma, W., Terzi, R., Lee, S., & de la Torre, J. (2017, April). *Multiple group cognitive diagnosis models and their applications in detecting differential item functioning*. Paper presented at the annual meeting of the American Educational Research Association, San Antonio, TX.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*, 253–275.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 11*, 71–101.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*, 55–64.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics, 43*, 88–115.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017a). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*, 614–631.

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017b). Two-step likelihood ratio test for item-level model comparison in cognitive diagnosis models. *Methodology, 13*, 39–47.

Templin, J. L., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*, 317–339.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305.

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*, 37–50.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287–307.

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series, 2014*, 1–13.

Xu, X., & von Davier, M. (2008). Fitting the structured general diagnostic model to NAEP data. ETS Research Report Series, 2008, 1–18.

Chapter 30
GDM Software *mdltm* Including Parallel EM Algorithm

Lale Khorramdel, Hyo Jeong Shin, and Matthias von Davier

Abstract This chapter illustrates the use of the software *mdltm* (von Davier, A general diagnostic model applied to language testing data. ETS Research Report No. RR-05-16, Educational Testing Service, Princeton, 2005), for multidimensional discrete latent trait models. The software *mdltm* was designed to handle large data sets as well as complex test and sampling designs, providing high flexibility for operational analyses. It allows the estimation of many different latent variable models, includes different constraints for parameter estimation, and provides different model and item fit statistics as well as multiple methods for proficiency estimation. The software utilizes an computationally efficient parallel EM algorithm (von Davier, New results on an improved parallel EM algorithm for estimating generalized latent variable models. In van der Ark L, Wiberg M, Culpepper S, Douglas J, Wang WC (eds) Quantitative psychology. IMPS 2016. Springer Proceedings in Mathematics & Statistics, vol 196. Springer, New York, 2017) that allows estimation of high-dimensional diagnostic models for very large datasets. The software is illustrated by applying diagnostic models to data from the programme for international student assessment (PISA).

30.1 Introduction

Many diagnostic classification models (DCMs) turn out to be special cases of the General Diagnostic Model (GDM; von Davier, 2005, 2013, 2014; von Davier & Rost, 2016). These models aim to provide additional information beyond overall test scores, as typically obtained by classical test theory (CTT) or unidimensional

L. Khorramdel · M. von Davier (✉)
National Board of Medical Examiners (NBME), Philadelphia, PA, USA
e-mail: LKhorramdel@nbme.org; mvondavier@nbme.org

H. J. Shin
Educational Testing Service, Princeton, NJ, USA
e-mail: hshin@ets.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_30

item response theory (IRT). While a test score describe levels of test-takers overall proficiency with respect to the latent construct, DCMs may help to better understand what specific strengths and weaknesses are at play. For example, the latent construct might consist of different subscales aimed at obtaining competency profiles based on subscale scores. DCMs may be helpful to either confirm a subscale model or to find meaningful subscales in exploratory analyses should they exist. Moreover, they can be used to find latent classes of test-takers who differ systematically with regard to these subscales. Such approaches can be helpful to better understand quantitative and qualitative performance differences among test-takers, not only for individual-level score reporting but also for group-level score reporting.

This chapter illustrates the use of the software *mdltm* (von Davier, 2005), for multidimensional discrete latent trait models, in applying diagnostic models to data from an international large-scale group score assessment (e.g., Mazzeo & von Davier, 2008, 2013). We apply different DCMs using data from the Programme for the International Student Assessment (PISA; OECD, 2017) collected in 2015. In the PISA 2015 cycle, the major domain was "Scientific Literacy" (Science, for short), in which a new framework and new items were administered in addition to trend items from previous cycles (OECD, 2016). Items were deliberately allocated and designed according to a new science framework. The example provided here demonstrates a confirmatory approach that is typical of DCMs. A unidimensional IRT model (as a basis for comparison) is compared to multidimensional and mixture multidimensional IRT models specified in the GDM framework.

The software *mdltm* was designed to handle large data sets as well as complex test and sampling designs, providing high flexibility for operational analyses. It allows the estimation of many different latent variable models, includes different constraints for parameter estimation, and provides different model and item fit statistics as well as methods for proficiency estimation. In addition, it can handle missing data by design and non-response, as well as multiple populations and weights to account for complex sampling (e.g., Rutkowski, von Davier, Joncas, & Gonzales, 2010). Furthermore, IRT linking can be easily accomplished in *mdltm,* which allows for a wide range of customary linking approaches (von Davier & von Davier, 2007; von Davier, González, & von Davier, 2013; Xu & von Davier, 2008a). Moreover, the author of the software developed a parallel EM (expectation-maximization) algorithm (von Davier, 2016) that allows for much faster parameter estimation. This is especially helpful in the analysis of large data sets or high dimensional models.

The next sections provide more information about the software *mdltm* and the parallel EM algorithm. We illustrate how certain DCMs can be estimated with *mdltm* and how they can be interpreted using empirical examples based on PISA data.

30.2 *mdltm*

The software is based on the mixture general diagnostic modeling framework (MGDM; von Davier, 2008, 2010; von Davier & Rost, 2016). The software can

be requested for research purposes by contacting the author of the program. The software can be used for estimating the parameters and examining the goodness of fit for a wide range of latent variable models:

- Unidimensional and multidimensional IRT (MIRT) models based on the Rasch Model and two-parameter logistic model (2PLM) for dichotomous responses
- IRT and MIRT models based on the partial credit model (PCM) and the generalized partial credit model (GPCM) for polytomous responses
- Latent class models and multiple-classification latent class models
- Unidimensional and multidimensional located latent class models
- Diagnostic classification models with dichotomous or ordinal skill variables
- Mixture distribution IRT and mixture diagnostic classification models (DCMs)
- Growth mixture models, hierarchical latent class models
- Hierarchical diagnostic models
- Multiple-group IRT models

The family of models included in the MGDM framework is covered in detail in Chap. 6 of this volume. Readers interested in model equations and statistical details, including parameter estimation and fit assessment, please refer to Chap. 6 and the references therein.

The software provides marginal maximum likelihood (MML) estimates obtained using customary expectation-maximization (EM) methods. Due to the use of optimized code, the software often provides estimates within seconds for small and moderately sized datasets and typically converges within minutes even for large data sets (with samples of 200,000 or more respondents) based on multiple test forms with a combined coverage of hundreds of items. In addition to maximum likelihood estimates of model parameters, the software provides logic and parsing routines for model specification and data processing. Moreover, *mdltm* can be used on all major operating systems, for example on Microsoft Windows, Linux, and Apple OS X platforms.

Recently, parallel processing was enabled in the program, the parallel-E parallel-M (PEPM) algorithm, which further improved the performance of the software (von Davier, 2016). The PEPM algorithm is based on a direct implementation of distributed parallelism that allows the utilization of all processor cores. This allows for more efficient computation, with a reduction in time by a factor of 6 or even 20 for some examples (von Davier, 2016).

In addition to efficient computation, *mdltm* is flexible in handling missing observations as well as multiple populations. In the software, missing responses are handled directly, without needing to collapse categories or recode data. The estimation of skill distributions for multiple populations is conducted simultaneously, thus enabling the comparison of parameters across multiple populations. Various constraints can be imposed on item parameter estimates, such as equality constraints typically needed for linking purposes across items or populations, as well as fixed parameter linking constraints using values from previous calibrations.

Hence, the software is suitable and has been used for operational analysis of data from large-scale assessment programs, such as the Programme for International Student Assessment (PISA; OECD, 2017) and the Programme for the International

Assessment of Adult Competencies (PIAAC; OECD, 2013). It has also been used for research based on data from the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), as well as data from large-scale testing programs including the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE).

The software output provides tabulations of observed quantities (item category frequencies, item-total correlations, etc.), parameter estimates, standard errors, and expected counts. In addition, a variety of goodness-of-fit indices are generated for each estimated model. Information criteria (Akaike, 1974; Schwarz, 1978) and related quantities, such as the log-penalty (Gilula & Haberman, 1994), are available, as well as other fit diagnostics, such as item fit based on pseudo-counts and person fit indices based on observed and expected counts. The following table gives an overview of the features provided in *mdltm* (Table 30.1).

The software allocates memory dynamically, so there is no inherent limitation of the number of items. Diagnostic models with 12–14 attributes ($2^{12} \sim 2^{14}$ attribute patterns) have been successfully estimated with *mdltm*.

30.3 Parallel EM Algorithm in mdltm

To improve the performance of *mdltm* for large data sets and complex model estimations, von Davier (2016) developed the PEPM algorithm which parallelizes both the E step and the M step. The parallel estimation algorithm was further improved and tested for additional gains by von Davier (2017). The PEPM algorithm is based on a direct implementation of parallelism using a paradigm that allows the distribution of work among all available processor cores of a PC. It leads to a substantial reduction in time in the most calculation-intense parts of the program. A server with 32 physical cores executes the parallel-E step algorithm up to 20 times faster than a single-core computer or the equivalent nonparallel algorithm (note, that modern desktop computers as well as many laptops use processors that contain four cores and potentially twice the number of virtual cores).

Parallel computing for psychometric modeling with general latent variable models can provide analyses based on the full data set without shortcuts such as relying on subsamples and/or approximations and simplifications based on the model structure (e.g., Cai, 2010a, 2010b; Rijmen, Jeon, Rabe-Hesketh, & von Davier, 2014) or computational approximations (e.g., Jeon & Rijmen, 2014; Jeon, Rijmen, & Rabe-Hesketh, 2013; Rijmen & Jeon, 2013). Therefore, the advantages of the PEPM algorithm for psychometric modeling and parameter estimation are profound.

A move to special-purpose hardware for further speedup of the algorithm presented by von Davier (2016, 2017) appears to be straightforward. Parallel algorithms can utilize special-purpose graphical processing units (GPUs) that provide a much larger number of specialized cores or, alternatively, can make use of multicore coprocessors (such as the Xeon-Phi series) for further speedup. It should be noted that the so-called hyper-threading technology does not provide further speedup, as it does not double the number of physical cores but rather arranges them into virtual

Table 30.1 Features of the *mdltm* software

Class of features	Feature	Implementation
Basic characteristics	Type of Software	Stand-alone software for Windows, Apple OSX, Unix, Linux, Solaris, BSDs
	Required primary software	None
	Fee for software	None
	Programming language used	C
	GUI for input	Yes
	If not GUI, specify input format	ASCII script
	GUI for output	No
	If not GUI, specify output format	ASCII files
	Estimation approach	MML/EM
	Ability to modify estimation parameters	Various convergence criteria, acceleration, starting values, model constraints, linking constraints, parameter fixation, ability distribution constraints, etc.
Input Characteristics	Number of Response Variables	1000+ (models with 1000 or more items or response variables were estimated successfully)
	Scale Types	Nominal, dichotomous, polytomous
	File formats	Raw data
Model structures	Compensatory DCMs	Yes
	Non-compensatory DCMs	Yes
	Different DCMs for different items	Yes
	Number of Latent Variables	16
	Scale Types	Nominal, dichotomous, polytomous, continuous
	Maximum Number of Scale Points	Unlimited
	Structural modeling of attribute space	Yes
	Models for attribute space	Saturated, log-linear models, independence models, multiple group log-linear models
Estimates	Item parameters	Yes
	SEs for item parameters	Yes
	Person parameters	Yes
	SEs for person parameters	Yes
	Classification reliability	Yes
	Information indices	Yes
	Traditional CTT statistics for subscales	Yes
Fit indices	Item fit	Yes
	Person fit	Yes
	Absolute model fit	Yes
	Relative model fit	Yes
	Q-matrix misspecification	Yes

cores only. However, even on a customary four-core laptop computer, a significant increase in estimation speed can be gained by applying the PEPM algorithm.

The PEPM algorithm was used for operational analyses for the PISA 2015 and PISA 2018, and involved unidimensional and multidimensional models. The gains achieved with this algorithm allowed the full scaling analysis to be run within 1 day whenever new data files became available. The data used for the PISA operational scaling consisted of up to two-million students combined from four PISA cycles (i.e., 2006, 2009, 2012, 2015) and three core scales (Mathematics, Reading, and Science), with additional scales for some countries (e.g., Financial Literacy, Collaborative Problem Solving).

30.4 Input File Preparation and Output

30.4.1 Data and Instrument

To illustrate the estimation of DCMs with *mdltm*, we use the data from the United States collected in the PISA[1] 2015 cycle (N = 5677), applying different psychometric models to the Science domain. In the PISA 2015 cycle, the major domain was Science, in which a new assessment framework was developed for extension of the construct through new interactive items (OECD, 2016). Trend and new items were deliberately allocated and designed according to the new Science framework based on the following subscales:

- Science Knowledge: content, procedural, and epistemic
- Science Competency: explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically
- Science System: physical, living, and earth and space

These subscales allow for investigations of different aspects within the Science domain and, thus, for exploring further the variability of skills within and across countries participating in PISA. Table 30.2 gives an overview of the distributions of 85 trend and 99 new items (184 in total) to the three main subscales, Knowledge, Competency, and System, as well as the eight underlying subscales. It should be noted that the three Science subscale types are based on a three-way classification of the same 184 items (distributed into the $2 + 3 + 3 = 8$ subscales).

[1] PISA is a major international academic student survey that assesses the proficiencies of 15-year-old school populations (students in grade 7 or higher) in the domains of mathematics, reading, and science (sometimes accompanied by additional cognitive domains of interest such as collaborative problem solving and financial literacy). PISA is administered every 3 years since 2000 with the aim of monitoring students' ability to use their knowledge and skills for meeting real-life challenges and to provide trend measures over time. In each cycle, one of the three domains is featured as major domain and consists of trend and new items, while the others serve as minor domains and consist of trend items only.

Table 30.2 Distribution of 85 trend and 99 new items to the Science subscales

Knowledge			Competency			System		
Subscales	Trend	New	Subscales	Trend	New	Subscales	Trend	New
Content	51	47	Explain Phenomena Scientifically	42	47	Physical	28	33
Procedural and Epistemic	34 (24 + 10)	52 (36 + 16)	Evaluate and Design Scientific Enquiry	16	23	Living	39	35
–	–	–	Interpret Data and Evidence Scientifically	27	29	Earth and Space	18	31
Total no. of trend/new items	85	99		85	99		85	99
Total no. of items	184			184			184	

While PISA 2015 test scores (plausible values[2]) were generated only for the three main subscales (Knowledge, Competency, System), we make use of different item classifications to demonstrate the utilization of DCMs.

30.4.2 IRT and Diagnostic Classification Models

In this chapter, we fit different IRT and MIRT models, implemented as discrete latent trait models, which are equivalent to DCMs with polytomous ordered skill variables. These models are fit to the PISA 2015 Science data from the United States. All models are based on the two-parameter logistic model (2PLM; Birnbaum, 1968) for dichotomous data and the generalized partial credit model (GPCM; Muraki, 1992) for polytomous data, and are compared to a unidimensional (1D) 2PLM/GPCM as the baseline. The models account for different Science subscales and hypotheses. More precisely, we estimate the following models:

[2]Plausible values are multiple imputations drawn from a posterior distribution obtained from a latent regression model (also referred to as population modeling or conditioning model) using IRT item parameters from the cognitive PISA assessment and principal components from the PISA Background Questionnaire. In PISA, each respondent receives 10 plausible values for each cognitive domain that can be used as test scores to produce group level statistics (never as individual test scores). For more information on plausible values and population modeling in large-scale assessments, see Mislevy and Sheehan (1987), von Davier, Gonzalez and Mislevy (2009), von Davier, Sinharay, Oranje, and Beaton (2006) or Yamamoto, Khorramdel, and von Davier (2013, updated 2016).

IRT Models:

(a) 1D model (baseline model): All Science items (y) are assigned to one overall dimension (θ).
(b) 3D model: The Science items are assigned to the three Competency subscales (explain θ_1, evaluate and design θ_2, interpret θ_3), which aim to assess different cognitive processes. Because each item measures only one dimension, this model accounts for between-item multidimensionality (Adams, Wilson, & Wang, 1997).
(c) 3D mixture distribution models: As in the 3D model, the Science items are assigned to the three Competency subscales. But instead of assuming homogeneity among respondents (one class), we test for multiple latent classes (heterogeneity among respondents who show different response patterns) and compare these additional models to the model with one class. For more information on mixture distribution models, see von Davier and Carstensen (2006), for example.

Diagnostic Classification Model (DCM):

(d) 3D/Bifactor model: As in the 3D model, the Science items are assigned to the three Competency subscales. In addition to these specific dimensions, all items are also assigned to a general dimension or skill (θ_g). This means that each item is assigned to two dimensions, one of the specific Competency subscales and a general dimension. Thus, this model accounts for within-item multidimensionality. For more information about the bifactor model, see Gibbons and Hedeker, (1992), for example.

Each multidimensional model can be represented as a DCM with binary latent variables, when only assuming mastery/non-mastery of the skills is deemed sufficient. Alternatively, the GDM can be specified with multiple levels of proficiency per latent variable, thus generalizing the DCM approach to latent variables with polytomous, ordinal skills. The 1D, 3D and 3D/Bifactor models are illustrated in Figs. 30.1 and 30.2; the 3D mixture distribution models are an extension of 3D

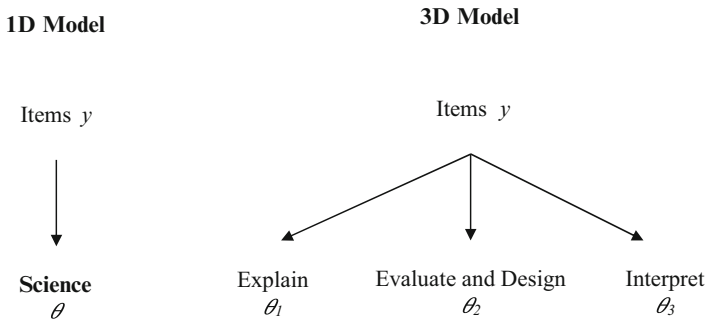


Fig. 30.1 1-dimensional (1D) and 3-dimensional (3D) models for the domain of Science and its Competency subscales

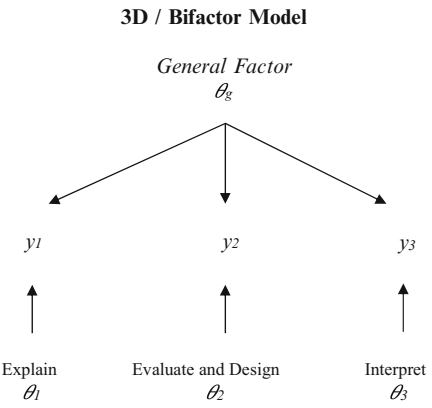


Fig. 30.2 3D/Bifactor model accounting for the Science Competency subscales as specific factors and a general factor

model to multiple latent classes. Given past experiences with multidimensional IRT (MIRT) models and DCMs used in PISA, TIMSS and PIAAC, we expect that the Science subscales are highly correlated skills, and that the models are not significantly distinguishable from the 1D baseline model in terms of model fit. Hence, the MIRT models and DCMs are expected to provide at best an alternative, more complex description of the data.

30.5 Estimating DCMs and Other Psychometric Models in *mdltm*

There are different types of *mdltm* files: the data file (*.dat), the syntax or input file (*.inp), the main output file (*.lst), and specific output files (*.items, *.pparm, *.status). Data and input files must be prepared in advance for analyses while output files will be generated during and after the analyses by the software, if requested in the input file. In the following sections, we illustrate what a simple input file looks like for IRT models and DCMs, how simple and complex Q-matrixes can be defined, and we give an example for the output generated by the *mdltm* software.

30.5.1 Preparing the Input (Syntax) File

This section illustrates the basic commands used in *mdltm* input or syntax files for estimating IRT models and DCMs. A more comprehensive list of commands and details can be found in the *mdltm* user manual, available upon request from the authors of this chapter. Most commands are not case-sensitive, but some are (for example group IDs; the input file has to use the same IDs as they appear in the data file to avoid mis-assignment). Some commands are sensitive to the order in which they are listed; for example, the number of items in the model has to

be specified before individual item IDs and locations are defined, and multiple groups or populations (if used) as well as items must be defined before constraints or releases of constraints on items or groups can be applied, in order to avoid unidentified assigments.

First, we must specificy the data file using the command:

```
data = PISA2015_data_Science.DFILE
```

From this data file, we select the variables we want to use for our analysis and specify some information about the variables we are selecting. We start by defining the maximum number of response categories the items can have and the maximum number items we want to use:

```
maxnrcat=4
nitems=184
```

If the items we select have fewer categories or if we use fewer items than we defined, we do not have to change the information above since we defined a maximum allowance for these numbers. However, we cannot use more items or items with a higher number of categories unless we adjust these maximum values.

In the next step, the software allows users to select and define known groups through group IDs (e.g., different countries or variables defined in the data to identify reporting groups such as gender, or ethnicity). In this chapter, we use a data file that only contains the cases for one single population, or group (the United States), and we do not have to select or define any group IDs. In addition to specify known groups, the software also allows for the estimation of parameters for latent classes. The following command can be used to specify the number of latent classes: ncl. If we want to allow for multiple latent classes, we can set the number to 2 or more (e.g., ncl=2). If we do not want to allow for latent classes, we tell the software to assign all cases to the same class:

```
ncl=1
```

This command is also used to specify the number of known groups in addition to the use of group IDs (the procedure for this can be found in the software manual). Next, we define the number of dimensions or skills we want to estimate in our model. If we want to estimate a unidimensional model, we specify:

```
nskills=1
```

Larger numbers would be used for DCMs or MIRT models. We also must define the maximum number of skill levels for any skill we are estimating using the following command:

```
maxlevel=11
```

The number of skill levels reflects the number of intervals used in the numerical integration over the latent distribution. For scales that contain many items, in particular if respondents answer all items without any omissions or missing by design, more skill levels can be chosen. The skill levels used here can be viewed analogous to the quadrature points used in major MML based IRT estimation

software packages. However, in *mdltm*, skill levels can be fixed and user defined, as done in MML numerical integration over fixed quadrature points, or estimated and adjusted while iterations take place. Fewer quadrature points lead to shorter run times, and often do not harm model-data fit (Haberman, von Davier, & Lee, 2008). For diagnostic models, we can use as few as two levels, which is the typical number for DINA, LCDM, and G-DINA, while the GDM also allows models with more than 2 skill levels. In case of a unidimensional model and a large number of items, a larger number, such as 41 or more skill levels/quadrature points can be used.

Next, we define each skill, assign a name, indicate the actual number of levels (again, we can use fewer than the maximum number of levels defined above – in our case 9 – but never a larger number), and set the range of the ability distribution on the logit scale – in our case from −3 to 3. For a unidimensional model, this can look like:

```
skill=1,"Science",9,-3:3
```

For a multidimensional model with three dimensions or skills, we may define:

```
skill=1,"Explain-Phenomena",5,-3:3
skill=2,"Evaluate-Design",5,-3:3
skill=3,"Interpret-Data-Evidence ",5,-3:3
```

If the skills have a different item proportions, we can also use a different number of levels or quadrature points for each skill. For example, 5 for the first skill with most items, 3 for the second skill with fewer items, and 2 for the third skill with the smallest number of items:

```
skill=1,"Explain-Phenomena",5,-3:3
skill=2,"Evaluate-Design",3,-3:3
skill=3,"Interpret-Data-Evidence ",2,-3:3
```

The number of maximally supported skill levels is described in research on semi-non-parametric IRT models, references are provided for example by Haberman et al. (2008).

Further, we define whether we want to estimate a model based on the Rasch Model (Rasch, 1960) or PCM (Masters, 1982), or based on the 2PLM (Birnbaum, 1968) or GPCM (Muraki, 1992). The following commands are used to specify the 2PLM:

```
doslopes=yes
centerslope=items
centerscale=items
```

To remove the indeterminancy of IRT scales, we could also use centerslope = cases and centerscale = cases (to set the latent ability distribution to have a mean of 0 and a standard deviation of 1) instead of items (where the mean of item difficulties is set to 0 and the mean of item slopes is set to 1). For multidimensional models or for the models with multiple groups or latent classes, a reference dimension or a reference group can be specified (the procedure and a detailed description of the commands used for this can be found in the software manual).

For loglinear constrained (i.e., smoothed) latent distributions (Xu & von Davier, 2008b, 2008c), which is recommended to be used in most cases, we set:

```
traitdistribution=loglin
```

We could use traitdistribution = saturated for unconstrained latent distribution or traitdistribution = independence for uncorrelated factors, and traitdistribution = bifactor for estimating a Bifactor model. To approximate normal distributions, we specify the maximum of fitted marginal moments of the loglinear distribution by using:

```
maxmoment=2
```

To allow more flexible shapes in the distribution, we can use higher order moments >2.

We can also define in which columns the respondent ID can be found in the data file (e.g., from column 15 to 33), and we can specify the use of sampling weights (e.g., found in columns 51 to 60 in the data file):

```
subjectid=15.33
weight=51.60
```

A very important part of the input file is the definition of items along with a Q-matrix. The following item definitions tell the program the columns in the data file in which the items are located, which items we want to select for the analysis, and to which dimension or skill each item should be assigned (Q-matrix). An example of item definition that includes the Q-matrix is:

```
item=301.301,2PL,"DS269Q01C",1
item=302.302,2PL,"DS269Q03C",1
item=303.303,2PL,"CS269Q04S",1
item=304.304,2PL,"CS408Q01S",1
item=305.305,2PL,"DS408Q03C",1
item=306.306,2PL,"CS408Q04S",1
item=307.307,2PL,"CS408Q05S",1
item=308.308,2PL,"CS521Q02S",1
item=309.309,2PL,"CS521Q06S",1
item=310.310,2PL,"DS519Q01C",1
item=311.311,2PL,"CS519Q02S",1
item=312.312,2PL,"DS519Q03C",1
item=313.313,2PL,"CS527Q01S",1
item=314.314,2PL,"CS527Q03S",1
item=315.315,2PL,"CS527Q04S",1
.....
```

First, we define in which column an item can be found; in this example, the first item can be found in column 301 in the data file (items that span multiple columns are possible, but are not encountered frequently, so that the format "starting_column.ending_column" for each item will typically contain the same entry twice). We then define the IRT model as "2PL" (or "Rasch" for the Rasch Model), assign an ID to the item as it should be printed in the output files, and determine to which dimension/skill each item should be assigned (0 for not assigned and 1 for assigned). The Q-matrix above is defined for a unidimensional model

where all items are assigned to the same dimension/skill. Below, we give examples of Q-matrices for multidimensional models with multiple dimensions/skills.

We can also define missing and omitted responses in *mdltm*. Missing values can be a result of planned missingness in complex rotated booklet designs where not all items are administered to all examinees by design, or due to respondent behavior when examinees do not reach the items at the end of a test when running out of time (i.e. the item is not presented to the examinees). Omitted responses occur when an item is presented but the respondent chooses not to provide an answer. Every missing response followed by a valid response to the next item is typically defined as an omitted response. Missing values by design and not reached items may be coded with 9 in the data file and omitted responses could be coded with 8, but other values are possible, as long as these are distinct from observed response codes. In the *mdltm* input file we include the following commands to indicate which responses are to be considered missing and omitted, respectively:

```
missing=9
omit=8
```

Omitted responses can be further be recoded to incorrect responses and assigned a value of 0 if we have the hypothesis that omitted responses may be the result of low proficiency (like it is done in PISA):

```
recode = ALL:8>0
```

This way, omitted responses are included in the likelihood function. However, if omitted responses are missing at random, are not related to proficiency given other observed variables, and, hence, should not be included in the likelihood function, we can recode them to missing values:

```
recode = ALL:8>9
```

The recode command can also be used to recode other values in the data file for all items (in this case, use ALL) or for selected items (in this case, use the item ID).

In a last step, we can set the number of iterations and the convergence criteria, for example:

```
iterations=999, 0.01
```

In this example, the iterations will stop either when the number of iterations reaches 999 or when the change in likelihoods between two consecutive iterations is smaller than 0.01.

With the following command, we define a maximum stepwidth that governs the adjustment of parameters in the maximization step. Any number between 0 and 1 can be used, for example:

```
maxstepwidth=0.9
```

With the following command, we specify that a *.pparm file will be generated that includes person parameter estimates, most likely class memberships, response residuals (optional) and person fit statistics:

```
personparameter=yes
```

We can also define which type of proficiency estimates (WLE for weighted likelihood estimates, MAP, MLE for maximum a posteriori or maximum likelihood estimates, and EAP for expected a posteriori estimates) we want to generate and print:

```
printskillprob=WLE
```

The software *mdltm* allows for user specified changes to defaults by way of many more possible commands and constraints. For example, item parameters can be fixed to values obtained from a prior estimation (fixed item parameter linking) or set to be equal across different groups in multiple group models for concurrent calibrations. Starting values can be set for the item parameter estimation, and single items can be excluded for either all groups in multiple group models, by deleting the item from the input file, or for selected groups, by setting the slope and difficulty parameter to zero for these groups. Please see the user manual for more information.

30.5.2 Q-matrix for Multidimensional Models

In the following, we illustrate specification of a Q-matrix for the different multidimensional models that were estimated. Table 30.3 shows the Q-matrix for a simple

Table 30.3 Q-matrix for the simple structure 3D model and the Bifactor model

	3D Model			*3D/Bifactor Model*			
	Skills/Dimensions (Science Competency Subscales)						
Items	**1**	**2**	**3**	**G**	**1**	**2**	**3**
DS269Q01C	1	0	0	1	1	0	0
DS269Q03C	1	0	0	1	1	0	0
CS269Q04S	1	0	0	1	1	0	0
CS408Q01S	1	0	0	1	1	0	0
DS408Q03C	1	0	0	1	1	0	0
CS408Q04S	1	0	0	1	1	0	0
CS408Q05S	0	1	0	1	0	1	0
CS521Q02S	1	0	0	1	1	0	0
CS521Q06S	1	0	0	1	1	0	0
DS519Q01C	0	0	1	1	0	0	1
CS519Q02S	1	0	0	1	1	0	0
DS519Q03C	0	1	0	1	0	1	0
CS527Q01S	0	0	1	1	0	0	1
CS527Q03S	1	0	0	1	1	0	0
CS527Q04S	0	0	1	1	0	0	1
…..							

Note: The first three skills are subscales of the Science Competency scale and the fourth skill is the general factor in the Bifactor model; skill 1 = Explain Phenomena Scientifically, skill 2 = Evaluate and Design Scientific Enquiry, skill 3 = Interpret Data and Evidence Scientifically, skill 4 = General Dimension (Science)

structure 3D model where each item is assigned to one of three dimensions, and the 3D/Bifactor model where each item is assigned to two dimensions out of four (a specific dimension and the general dimension). The matrices are illustrated for the first 15 Science items (out of 184 items).

In the *mdltm* syntax file, the matrix for the 3D model would look like this:

```
item=301.301,2PL,"DS269Q01C",1,0,0
item=302.302,2PL,"DS269Q03C",1,0,0
item=303.303,2PL,"CS269Q04S",1,0,0
item=304.304,2PL,"CS408Q01S",1,0,0
item=305.305,2PL,"DS408Q03C",1,0,0
item=306.306,2PL,"CS408Q04S",1,0,0
item=307.307,2PL,"CS408Q05S",0,1,0
item=308.308,2PL,"CS521Q02S",1,0,0
item=309.309,2PL,"CS521Q06S",1,0,0
item=310.310,2PL,"DS519Q01C",0,0,1
item=311.311,2PL,"CS519Q02S",1,0,0
item=312.312,2PL,"DS519Q03C",0,1,0
item=313.313,2PL,"CS527Q01S",0,0,1
item=314.314,2PL,"CS527Q03S",1,0,0
item=315.315,2PL,"CS527Q04S",0,0,1
.....
```

The Q-matrix for the Bifactor model would look like this within the item specification:

```
item=301.301,2PL,"DS269Q01C",1,1,0,0
item=302.302,2PL,"DS269Q03C",1,1,0,0
item=303.303,2PL,"CS269Q04S",1,1,0,0
item=304.304,2PL,"CS408Q01S",1,1,0,0
item=305.305,2PL,"DS408Q03C",1,1,0,0
item=306.306,2PL,"CS408Q04S",1,1,0,0
item=307.307,2PL,"CS408Q05S",1,0,1,0
item=308.308,2PL,"CS521Q02S",1,1,0,0
item=309.309,2PL,"CS521Q06S",1,1,0,0
item=310.310,2PL,"DS519Q01C",1,0,0,1
item=311.311,2PL,"CS519Q02S",1,1,0,0
item=312.312,2PL,"DS519Q03C",1,0,1,0
item=313.313,2PL,"CS527Q01S",1,0,0,1
item=314.314,2PL,"CS527Q03S",1,1,0,0
item=315.315,2PL,"CS527Q04S",1,0,0,1
.....
```

30.5.3 Output Files

The software *mdltm* provides a main output file (*.lst), and different specific output files (*.items, *.pparm, *.status).

– The *.items file provides an overview of all item parameter estimates (slope and intercept or location parameter estimates) as well as the Fisher information for each parameter.

– The *.pparm file provides the group or class membership for each examinee, the probability of belonging to a certain class (in latent class models), the percent of correct responses, the person parameter or latent ability estimate for each examinee and for each dimension or skill, and the standard error (SE) associated with the person ability estimate. Response residuals can be printed upon request in this file as well with the command printresponses = yes.
– The *.status file provides information about the convergence status of the model and shows the log for the iterations. This file is helpful to determine whether the model has fully converged or is close to converging.
– The *.lst file contains an overview of the main results, such as information about the data, descriptive statistics, and almost all results for the model of interest. It is not possible to list the whole *.lst file in this chapter (see the manual for more examples); in the following, we illustrate some of the most important results of the 1D 2PLM.

The main output, the *.lst file, starts with an overview of descriptive statistics, such as the number of responses[3] per response category for each item, the correlation between the item score and the test score r(itm,skill), and the logit transformed probability of receiving a correct response (logits; log(p(1-p)), and adjacent category logit for the polytomous items[4]):

```
     item                tried   0       1      2      r(itm,skill)|  logits
 1   DS269Q01C  ncat: 2  535.21  331.04  204.17        | 0.50890 |    -0.48139
 2   DS269Q03C  ncat: 2  535.21  332.28  202.93        | 0.54530 |    -0.49121
 3   CS269Q04S  ncat: 2  535.21  381.81  153.40        | 0.32015 |    -0.90797
 4   CS408Q01S  ncat: 2  535.21  246.16  289.05        | 0.46850 |     0.16001
 5   DS408Q03C  ncat: 2  532.72  323.16  209.56        | 0.36984 |    -0.43149
 6   CS408Q04S  ncat: 2  533.50  228.36  305.14        | 0.24222 |     0.28873
 7   CS408Q05S  ncat: 2  532.72  386.48  146.24        | 0.34811 |    -0.96757
 8   CS521Q02S  ncat: 2  532.72  262.02  270.70        | 0.22903 |     0.03244
 9   CS521Q06S  ncat: 2  531.56   72.17  459.39        | 0.42833 |     1.83936
10   DS519Q01C  ncat: 3  525.11  272.98   96.72 155.40 | 0.41146 |    -1.03092  0.47030
11   CS519Q02S  ncat: 2  527.80  243.79  284.00        | 0.27116 |     0.15209
12   DS519Q03C  ncat: 2  524.27  422.35  101.92        | 0.32622 |    -1.41428
13   CS527Q01S  ncat: 2  520.06  433.11   86.95        | 0.35832 |    -1.59654
14   CS527Q03S  ncat: 2  518.80  193.96  324.85        | 0.35298 |     0.51366
15   CS527Q04S  ncat: 2  515.87  230.07  285.80        | 0.45539 |     0.21608
.....
```

After the summary of iterations, the output file provides the number of estimated model parameters, the log-likelihood, deviance, different global model fit indices (e.g., AIC, BIC, CAIC), and information about the estimated skill distribution (mean, SD) per group/class:

```
Number of estimated item thresholds                197
Number of estimated item slopes                    184
Minus determinacy constraints on items (negative) -   2
```

[3]Note that decimals in the category frequency counts are due to the use of sample weights in the analyses.

[4]For the details about adjacent category logit, including various types of parameterization for the polytomous responses, please refer to Agresti (2002).

```
Number of estimated skill distribution parameters     2 (  2)
Minus skills determinacy constraints (negative)       0
Number of class-size and cluster parameters           0 (percent classified: 0)
Total number of parameters:                         381


Trait Distribution:  log-linear model min. moment 1 max. moment 2

Likelihood:     -85018.9138  Deviance:          170037.82755

AIC  penalty term:    762.00     AIC    Index:     170799.82755
AICc penalty term:    825.49     AICc   Index:     170863.32150
BIC  penalty term:   3242.48     BIC    Index:     173280.31200

BIC_sp*   penalty:   2555.78     BIC_sp Index:     172593.60759
BIC_nP    penalty:   1300.26     BIC_nP Index:     171338.08950
BIC_NP    penalty:   3856.04     BIC_NP Index:     173893.86954
CAIC penalty term:   3623.48       CAIC Index:     173661.31200

Penalty Factor BIC:     8.51, BIC_sp*:     6.71, BIC_nP:     3.41

Model based log penalty per item:      0.5640627 ..  Akaike:  0.5665905
Independence log penalty per item:     0.6421668
Average Model based odds P(X|M)/P(X|I)  1.0812352

Model based  log-lik per respondent:        -17.1186952
Independence log-lik per respondent:        19.4890706

Unweighted N: 5677 Average # of items per respondent:   30.35

Iterations        230

Estimated skill distribution(s):
Class: POP001 size:       1.0000000
Scale: 1
Estimated Skill Mean:      -0.0464442  .. Stdev:       0.7086600

-3.0000 :       0.0000714
-2.2500 :       0.0033576
-1.5000 :       0.0515205 ****
-0.7500 :       0.2579311 ***********************
 0.0000 :       0.4213047 ***************************************
 0.7500 :       0.2245212 ********************
 1.5000 :       0.0390379 ***
 2.2500 :       0.0022145
 3.0000 :       0.0000410
```

The AIC (Akaike, 1974) and the BIC (Schwarz, 1978) use the maximum likelihood value (L) of a model, the number of estimated model parameters (k), and the sample size. The AIC is computed as follows:

$$\text{AIC} = -2\log L + 2k \tag{30.1}$$

While the number of model parameters in the AIC is weighted with 2, the BIC uses the logarithm of the sample size (N) as weight:

$$\mathrm{BIC} = -2 \log L + (\log N)\, k \tag{30.2}$$

Thus, the BIC penalizes overparameterization more than the AIC as soon as $\log(N) > 2$.

This information is followed by the final item parameter estimates (a is the slope, b is the difficulty, d is the intercept related to the difficulty), standard errors, and the expected category frequencies[5] per item:

```
Final Item Parameter Estimates
item  itemlabel   a-param | intercept|      | b-param | d-step | d-step |
1     DS269Q01C   1.30806 | -0.65396 |              | 0.29409 |
2     DS269Q03C   1.44684 | -0.70355 |         |  0.28604 |
3     CS269Q04S   0.72904 | -1.02784 |         |  0.82932 |
4     CS408Q01S   1.01714 |  0.25332 |         | -0.14650 |
5     DS408Q03C   0.80686 | -0.48730 |         |  0.35527 |
6     CS408Q04S   0.49957 |  0.33635 |         | -0.39605 |
7     CS408Q05S   0.85017 | -1.14310 |         |  0.79091 |
8     CS521Q02S   0.42838 |  0.05369 |         | -0.07373 |
9     CS521Q06S   1.69735 |  3.05137 |         | -1.05748 |
10    DS519Q01C   0.55789 | -0.91470 | 0.23076 |  0.36057 | -0.60388 | 0.60388 |
11    CS519Q02S   0.53866 |  0.18829 |         | -0.20561 |
12    DS519Q03C   0.87731 | -1.68657 |         |  1.13084 |
13    CS527Q01S   1.06038 | -2.02111 |         |  1.12119 |
14    CS527Q03S   0.78029 |  0.65389 |         | -0.49294 |
15    CS527Q04S   1.04855 |  0.34135 |         | -0.19149 |
.....
Final Item Parameter Standard Errors
item       itemlabel       a-param  |  intercept |  |
1            DS269Q01C       0.11696  |    0.10727 |
2            DS269Q03C       0.12469  |    0.11037 |
3            CS269Q04S       0.09205  |    0.10227 |
4            CS408Q01S       0.10134  |    0.09920 |
5            DS408Q03C       0.09259  |    0.09710 |
6            CS408Q04S       0.08009  |    0.09110 |
7            CS408Q05S       0.09627  |    0.10583 |
8            CS521Q02S       0.07783  |    0.08938 |
9            CS521Q06S       0.11300  |    0.15189 |
10           DS519Q01C       0.05501  |    0.09704 |   0.10592 |
11           CS519Q02S       0.08139  |    0.09151 |
12           DS519Q03C       0.10158  |    0.11908 |
13           CS527Q01S       0.10760  |    0.12941 |
14           CS527Q03S       0.09092  |    0.09885 |
15           CS527Q04S       0.10427  |    0.10198 |
.....
```

[5]The expected category frequencies (for multiple groups) and conditional proportions correct P(+|group) are statistics given separately for each group (e.g. a state, country or language). For latent class models, mixture IRT models and diagnostic models, the expected category frequencies are expected proportions correct per latent class, which are estimates of these proportions, given the classifications of respondents (proportionally assigned using posterior distribution of class membership given observed responses) into these classes.

```
Expected category frequencies by class and scale:
seq  itemlabel ncat nresponses     0    1       2      expected
1     DS269Q01C     2      535.21    0.6185   0.3815          0.3815
2     DS269Q03C     2      535.21    0.6208   0.3792          0.3792
3     CS269Q04S     2      535.21    0.7134   0.2866          0.2866
4     CS408Q01S     2      535.21    0.4599   0.5401          0.5401
5     DS408Q03C     2      532.72    0.6066   0.3934          0.3934
6     CS408Q04S     2      533.50    0.4280   0.5720          0.5720
7     CS408Q05S     2      532.72    0.7255   0.2745          0.2745
8     CS521Q02S     2      532.72    0.4919   0.5081          0.5081
9     CS521Q06S     2      531.56    0.1358   0.8642          0.8642
10    DS519Q01C     3      525.11    0.5199   0.1842 0.2959   0.7761
11    CS519Q02S     2      527.80    0.4619   0.5381          0.5381
12    DS519Q03C     2      524.27    0.8056   0.1944          0.1944
13    CS527Q01S     2      520.06    0.8328   0.1672          0.1672
14    CS527Q03S     2      518.80    0.3739   0.6261          0.6261
15    CS527Q04S     2      515.87    0.4460   0.5540          0.5540
.....
```

Towards the end of the main output file, we find the test characteristic curve (TCC), and the item fit statistics for each item by group based on the chi-square item fit value per item, root mean square deviation (RMSD), and mean deviation (MD):

```
Model based expected total (TCC) and skill score (STCCs) given skill pattern
in class: POP001

Skill 1| E(Tot) |--------|---------|----- ----|---------|---------|
-3.000 |   4.75 |*       |         |          |         |         |
-2.250 |   9.81 |  *     |         |          |         |         |
-1.500 |  22.47 |      * |         |          |         |         |
-0.750 |  51.71 |        | *       |          |         |         |
0.000  |  97.83 |        |         |       *  |         |         |
0.750  | 143.48 |        |         |          |       * |         |
1.500  | 171.76 |        |         |          |         | *       |
2.250  | 185.01 |        |         |          |         |     *   |
3.000  | 190.87 |        |         |          |         |       *|
               0--------|---------|--- 98---|---------|--------197
item chi-square:
item     label           POP001
1        DS269Q01C        0.70
2        DS269Q03C        0.07
3        CS269Q04S        1.61
4        CS408Q01S        0.70
5        DS408Q03C        1.00
6        CS408Q04S        0.69
7        CS408Q05S        0.88
8        CS521Q02S        2.64
9        CS521Q06S        0.37
10       DS519Q01C       14.25
11       CS519Q02S        4.32
12       DS519Q03C        2.51
13       CS527Q01S        0.58
14       CS527Q03S        0.52
15       CS527Q04S        0.31
.....
```

```
item rmsea:
 item       label          POP001
 1          DS269Q01C      0.0113
 2          DS269Q03C      0.0093
 3          CS269Q04S      0.0244
 4          CS408Q01S      0.0233
 5          DS408Q03C      0.0268
 6          CS408Q04S      0.0244
 7          CS408Q05S      0.0191
 8          CS521Q02S      0.0453
 9          CS521Q06S      0.0134
 10         DS519Q01C      0.0550
 11         CS519Q02S      0.0630
 12         DS519Q03C      0.0468
 13         CS527Q01S      0.0117
 14         CS527Q03S      0.0147
 15         CS527Q04S      0.0182
 .....

item mean deviation:
 item     label            POP001
 1        DS269Q01C -      0.0001
 2        DS269Q03C -      0.0001
 3        CS269Q04S -      0.0001
 4        CS408Q01S -      0.0001
 5        DS408Q03C -      0.0001
 6        CS408Q04S -      0.0001
 7        CS408Q05S -      0.0001
 8        CS521Q02S -      0.0002
 9        CS521Q06S        0.0011
 10       DS519Q01C -      0.0002
 11       CS519Q02S -      0.0001
 12       DS519Q03C -      0.0000
 13       CS527Q01S -      0.0001
 14       CS527Q03S -      0.0001
 15       CS527Q04S -      0.0001
 .....
```

The basic idea of item fit statistics is to compare the observed item characteristic curve (ICC) using pseudo counts from the E-step with the model-based ICC, and monitor whether there is any item that exhibits a considerably large gap between those two curves. Both fit statistics, the MD and RMSD, quantify the magnitude and direction of deviations in the observed data from the estimated ICC for each item. While the MD is most sensitive to the deviations of observed item difficulty parameters from the estimated ICC, the RMSD is sensitive to the deviations of both the item difficulty parameters and item slope parameters.

The MD is the weighted sum of differences between the observed item response curve $P_o(\theta)$ and the expected item response function $P_e(\theta)$ over the range of the latent distribution $\int P_e(\theta)$. It linearly relates to the proportion of correct responses and is calculated as:

$$MD = \int \left(P_o\left(\theta\right) - P_e\left(\theta\right)\right) f\left(\theta\right) d\theta \tag{30.3}$$

The RMSD indicates the absolute difference between two ICCs by squaring the differences, multiplying the proficiency distribution as weights, and taking the square root of the total sum. It is calculated as:

$$RMSD = \sqrt{\int \left(P_o\left(\theta\right) - P_e\left(\theta\right)\right)^2 f\left(\theta\right) d\theta} \tag{30.4}$$

The *.lst file provides more information about the estimated model than the examples above capture, for example, the estimated skill distribution correlations in multidimensional models with multiple skills, and the class membership probabilities in mixture distribution models; please refer to the *mdltm* user manual for more details and output results.

30.6 Model Comparison Results

30.6.1 Overall Model Fit and Subscale Profiles

This section summarizes results and model fit statistics for the models described above estimated for the PISA USA Science data. To compare the overall model fit of the different models, we use the AIC and the BIC as well as the model-based log penalty. We also calculate the relative model fit improvements based on the log penalty measure between the different models. The log penalty (Gilula & Haberman, 1994) provides the negative expected log likelihood per observation. The percent of improvement compares the log-penalties of the models relative to the difference between the most restrictive and the most general model. Table 30.4 gives an overview of the results.

Results in Table 30.4 indicate that, based on the AIC and log penalty, the relatively best fitting model among the first group is the Bifactor model, which accounts for the three Science Competency subscales measuring different cognitive processes as specific factors and one general factor. According to the BIC, the 3D simple structure model shows the best fit. The AIC is known to choose more complex models, particularly with large samples. The BIC tends to reduce this effect by integrating the log of the sample size into the penalty term.

The differences in model fit improvements, based on the Gilula and Haberman (1994) log penalty measure, appear to be small. The 1D model reaches 96.94% of the model fit improvement obtained by the 3D/Bifactor model, and the 3D model without an additional general factor reaches 97.67%. Hence, the 1D model describes the data sufficiently well, and it is defendable to provide one overall Science test score for examinees and countries, as done for reporting in PISA 2015.

Table 30.4 Overall model fit of the uni- and multidimensional IRT models with between and within item multidimensionality

	Likelihood	AIC-penalty	AIC	BIC-penalty	BIC	Log penalty	% Improve-ment
unidimensional and multidimensional models							
Independence						0.64217	0.00%
1D Model	−85018.91	762	170,800	3242.48	173,280	0.56406	96.94%
3D Model	−84930.51	776	170,637	3302.06	**173,163**	0.56348	97.67%
3D/Bifactor Model	−84647.11	1142	**170,436**	4859.47	174,154	0.56160	100.00%
3D mixture distribution models							
Independence						0.64217	0.00%
1 class	−84930.51	776	170,637	3302.06	**173,163**	0.56348	88.51%
2 classes	−84303.64	1558	170,165	6629.65	175,237	0.55932	93.19%
3 classes	−83839.59	2340	170,019	9957.24	177,636	0.55624	96.66%
4 classes	−83391.56	3122	**169,905**	13284.82	180,068	0.55327	100.00%
5 classes	−83030.59	3904	169,965	16612.41	182,674	0.55087	

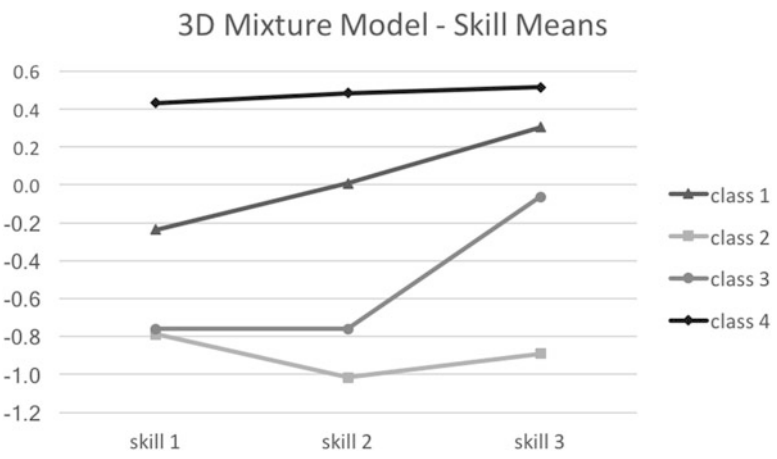


Fig. 30.3 Skill means for the three Science Competency subscales (S1, S2, S3) obtained from the mixture 3D model with four classes

Results for the 3D mixture distribution IRT models based on the AIC and log penalty indicate that a model assuming four latent classes fits the data relatively best, while the BIC shows the best fit for the 3D Model with one class only. Figure 30.3 displays the means for the three Science Competency subscales for each of the four classes obtained from the 3D mixture distribution models. These classes can be compared based on their Science subscale profiles. Depending on researcher's interest, further investigations across the classes using external variables (e.g., school type or gender) can be pursued.

30.7 Summary

The software *mdltm* (von Davier, 2005) was designed to enable estimation of a wide range of latent variable models using small and large data sets, such as data from international large-scale assessments, and can be used on all major operating systems, in particular, on Microsoft Windows, Linux, and Apple OS X platforms. The software allows estimation of unidimensional and multidimensional IRT (MIRT) models based on the Rasch Model and two-parameter logistic model (2PLM) or based on the generalized partial credit model (GPCM). Moreover, it can be used to estimate latent class models and multiple-classification latent class models, unidimensional and multidimensional located latent class models, diagnostic classification models with dichotomous or ordinal skill variables, mixture distribution IRT and diagnostic classification models (DCMs), growth mixture models, hierarchical latent class models, hierarchical diagnostic models, and multiple-group IRT models.

Furthermore, it offers a wide variety of different model constraints that allow for the application of different statistical and linking methods, such as fixed item parameter linking, concurrent calibration in IRT modeling, and the use of already-established item parameters as starting values for the estimation of item parameters in a new model. It also allows users to select a subset of samples/cases or items from the data set and to recode responses without needing to change the data file. The software provides useful information in different output files, including descriptive statistics based on the classical test theory, different proficiency estimates (EAP, MAP, WLE), model fit statistics (e.g., AIC, BIC, CAIC, log penalty) and item fit statistics (e.g., RMSD/RMSEA and MD), test characteristic curves (TCC), IRT based marginal reliabilities for each estimated scale/skill, latent correlations between different scales/skills in multidimensional models, and more.

The software provides marginal maximum likelihood (MML) estimates obtained using customary expectation-maximization (EM) methods, with optional acceleration, and it operates within seconds for small and moderately sized datasets and within minutes for large data sets. With the use of a recently developed PEPM algorithm (von Davier, 2016), parallel processing was enabled in the software, which further improved the performance of the software by achieving much more efficient computation with a reduction in time by a factor of 6 or even 20 for some examples.

In this chapter, we illustrated the use of *mdltm* for comparing simple structure with multidimensional and complex latent variable models. We estimated a unidimensional 2PLM, a 3-dimensional model, a Bifactor model and 3-dimensional mixture 2PLMs to describe the PISA Science scale using data from PISA 2015. It was illustrated how to prepare the *mdltm* syntax or input file to estimate the different models, and we showed examples of the main output file. We also showed how DCMs can be used to provide information that supplements the overall test score obtained from a unidimensional model.

Despite the advantages and flexibilities offered by *mdltm* in estimating DCMs, the usefulness of such models depends on the data set being analyzed. Therefore, users should carefully examine whether a DCM provides additional value over a simpler model. In a number of cases, such models might provide at best an alternative, more complex description of the data.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*, 33–57.

Cai, L. (2010b). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*, 307–335.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*, 645–656.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). Comparision of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions. *ETS Research Report Series* (pp. 1–25). https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Jeon, M., & Rijmen, F. (2014). Recent developments in maximum likelihood estimation of MTMM models for categorical data. *Frontiers in Psychology, 5*, 269. https://doi.org/10.3389/fpsyg.2014.00269

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics, 38*, 32–60.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results* (OECD Working Paper EDU/PISA/GB (2008) 28). Paris, France: OECD. Retrieved from https://edsurveys.rti.org/pisa/documents/mazzeopisa_test_designreview_6_1_09.pdf

Mazzeo, J., & von Davier, M. (2013). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Boca Raton, FL: CRC Press.

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177.

OECD. (2016). *PISA 2015 Assessment and analytical framework: Science, reading, mathematic and financial Literacy*. Paris, France: PISA, OECD Publishing.

Organisation for Economic Co-Operation and Development. (2013). *Chapter 17: Technical report of the Survey of Adult Skills (PIAAC)* (pp. 406–438). Retrieved from the OECD website: http://www.oecd.org/site/piaac/Technical%20Report_17OCT13.pdf

Organisation for Economic Co-Operation and Development. (2017). *PISA 2015 technical report*. Paris, France: OECD Publishing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche (Expanded Edition, Chicago, University of Chicago Press, 1980).

Rijmen, F., & Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Annals of Operations Research, 206*, 647–662.

Rijmen, F., Jeon, M., Rabe-Hesketh, S., & von Davier, M. (2014). A third order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics, 38*, 32–60.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142–151.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). The mixture general diagnostic model. In G. R. Hancock & K. M. Samuelson (Eds.), *Advances in latent variable mixture models*. Information Age Publishing.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

von Davier, M. (2013). The DINA model as a constrained general diagnostic model – Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology, 67*(1), 49–71. https://doi.org/10.1111/bmsp.12003

von Davier, M. (2014). *The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM)* (Research Report No. ETS RR-14-40). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12043

von Davier, M. (2016). High-performance psychometrics: The Parallel-E Parallel-M algorithm for generalized latent variable models. *ETS Research Report Series ISSN, 2016*, 2330–8516.

von Davier, M. (2017). New results on an improved parallel EM algorithm for estimating generalized latent variable models. In L. van der Ark, M. Wiberg, S. Culpepper, J. Douglas, & W. C. Wang (Eds.), *Quantitative Psychology. IMPS 2016. Springer Proceedings in Mathematics & Statistics* (Vol. 196). New York, NY: Springer.

von Davier, M., & Carstensen, C. H. (2006). *Multivariate and mixture distribution rasch models: Extensions and applications*. New York, NY: Springer.

von Davier, M., Gonzalez, E. & Mislevy, R. (2009) What are plausible values and why are they useful? In *IERI Monograph series: Issues and methodologies in large scale Assessments,* vol. 2. Retrieved from: http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M., González, J. B., & von Davier, A. A. (2013). Local equating using the Rasch Model, the OPLM, and the 2PL IRT Model—or—What is it anyway if the model captures everything there is to know about the test takers? *Journal of Educational Measurement, 50*(3), 295–303. https://doi.org/10.1111/jedm.12016

von Davier, M., & Rost, J. (2016). Logistic mixture-distribution response models. In W. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, 2nd ed., pp. 393–406). Boca Raton, FL: CRC Press.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics*. Amsterdam, The Netherlands: Elsevier.

von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology, 3*(3), 115–124.

Xu, X. & von Davier, M. (2008a). Linking with the General Diagnostic Model. *ETS Research Report No. RR-08-08*, Princeton, NJ: Educational Testing Service. http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2008.tb02094.x/full

Xu, X. & von Davier, M. (2008b). Fitting the structured general diagnostic model to NAEP data. *ETS Research Report No. RR-08-27*, Princeton, NJ: Educational Testing Service.

Xu, X. & von Davier, M. (2008c). Comparing multiple-group multinomial loglinear models for multidimensional skill distributions in the general diagnostic model. *ETS Research Report No. RR-08-35*, Princeton, NJ: Educational Testing Service.

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013, updated 2016). Chapter 17: Scaling PIAAC cognitive data. In OECD (2013), *Technical Report of the Survey of Adult Skills (PIAAC)* (pp. 406–438), PIAAC, OECD Publishing. Retrieved from http://www.oecd.org/site/piaac/All%20PIACC%20Technical%20Report%20final.pdf

Chapter 31
Estimating CDMs Using MCMC

Xiang Liu and Matthew S. Johnson

Abstract In this chapter, we provide a brief survey of Markov chain Monte Carlo (MCMC) methods used in estimating Cognitive Diagnostic Models (CDMs). MCMC techniques have been widely used for the Bayesian estimation of psychometric models. MCMC algorithms and general purpose MCMC software has been facilitating the development of modern psychometric models that are otherwise difficult to fit (Levy R, J Probab Stat 1–18, 2009. Retrieved from http://www.hindawi.com/journals/jps/2009/537139/, https://doi.org/10.1155/2009/537139). We introduce a Gibbs sampler for fitting the saturated Log-linear CDM model (LCDM, Henson RA, Templin JL, Willse JT, Psychometrika, 74(2):191–210, 2009. Retrieved from https://doi.org/10.1007/s11336-008-9089-5). The utility of Bayesian inference is demonstrated by analyzing the Examination for the Certificate of Proficiency in English (ECPE) dataset.

31.1 Introduction

In the past two decades, Markov chain Monte Carlo (MCMC) techniques have been widely used for the Bayesian estimation of psychometric models. Not only an alternative to other estimation methods, MCMC algorithms and general purpose MCMC software has been facilitating the development of modern psychometric models that are otherwise difficult to fit (Levy, 2009). In this chapter, we provide a brief survey of MCMC methods used in estimating Cognitive Diagnostic Models

X. Liu (✉)
Department of Human Development, Teachers College, Columbia University, New York, NY, USA
e-mail: xl2438@tc.columbia.edu

M. S. Johnson (✉)
Educational Testing Service, Princeton, NJ, USA
e-mail: msjohnson@ets.org

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*, Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4_31

(CDM). In addition, a Gibbs sampler for fitting the saturated Log-linear CDM model (LCDM; Henson et al., 2009) is introduced. The utility of Bayesian inference is demonstrated by analyzing the Examination for the Certificate of Proficiency in English (ECPE) dataset.

To help understand the motivation of developing MCMC methods, consider the following general statistical inference problem. Given a set of observed data $\boldsymbol{X} = \boldsymbol{x}$, we would like to model the data with a probabilistic model $p(\boldsymbol{x}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the model parameter vector. Under the Bayesian framework, a prior is assigned to the parameters, i.e., $p(\boldsymbol{\theta})$. Then we are interested in the posterior distribution of the model parameters given the observed data, i.e.

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{31.1}$$

In some cases the closed form of the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ can be analytically derived. However, under other circumstances, the posterior distribution must be approximated numerically. Difficulty arises from the numerical evaluation of the integral in the denominator of (31.1). If $\boldsymbol{\theta}$ is unidimensional, the integral can be approximated by using k quadrature points fairly efficiently. But in general, evaluating the multiple integral $\int\int\cdots\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\theta_1 d\theta_2 \cdots d\theta_d$ requires a high-dimensional grid of k^d points in $\mathbb{R}^d$. As the number of dimensions d grows, integration by quadrature quickly becomes infeasible. This problem is also referred to as the "curse of dimensionality". Instead of deterministically evaluating the high-dimensional integral, MCMC algorithms stochastically sample from the posterior distribution by constructing a Markov chain whose stationary distribution is the target posterior distribution. For a detailed review of MCMC, refer to Gelman et al. (2013), Neal (1998), and Brooks, Gelman, Jones, and Meng (2011).

Despite its importance to Bayesian inference, it should be noted that MCMC methods are not limited to Bayesian applications. High-dimensional integrals also arise from computing marginal maximum likelihood estimates in some models. As a result, MCMC as a class of efficient stochastic numerical integration algorithms is also used in frequentist applications. In fact, such applications have been developed in psychometrics. For example, Cai (2010) adapted the Metropolis-Hastings Robbins-Monro algorithm to estimate the high-dimensional item factor analysis model by marginal maximum likelihood. Given much improved computing power and the availability of general purpose Bayesian inference software, the CDM literature, flourishing in recent years, also saw a wide range of applications of MCMC methods.

31.2 MCMC Background

In this section, we provide a brief background and intuition of MCMC for readers who might not be familiar with the concept. A Markov chain is a *series* of random variables, $\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(t)}, \boldsymbol{\Theta}^{(t+1)}, \ldots$, where the state at time $t+1$ depends

only on the immediate previous state at t. In other words, the distribution of $\boldsymbol{\theta}^{(t+1)}$ is independent of everything else given $\boldsymbol{\Theta}^{(t)} = \boldsymbol{\theta}^{(t)}$, i.e.,

$$P(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t)}) = P(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}). \tag{31.2}$$

This is often referred to as the Markov property. Additionally, the state space, that is the range of $\boldsymbol{\theta}$, is common across all time points. In practice, it implies the parameter space of the model cannot be changed. However, there exist MCMC methods that can handle models with variable parameter space – for example, the reversible jump MCMC. This topic is significantly more advanced and out of the scope of this chapter. Interested readers may refer to Green (1995). Observing the aforementioned Markov property, it is clear that, in order to define a Markov chain, we need to specify the probability of an initial state $\boldsymbol{\theta}$ – $p_0(\boldsymbol{\theta}) = P(\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta})$ and the transition probabilities between consecutive states – $T_t(\boldsymbol{\theta}, \boldsymbol{\theta}') = P(\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}'|\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta})$ for $t = 0, 1, 2, \ldots$. Then the distribution of $\boldsymbol{\theta}$ at time $t + 1$ can be determined by

$$p_{t+1}(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{\theta}}} p_t(\tilde{\boldsymbol{\theta}}) T_t(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}). \tag{31.3}$$

For *homogeneous* Markov chains, the transition probabilities stay the same across all time points, i.e., $T_t(\boldsymbol{\theta}, \boldsymbol{\theta}') = T(\boldsymbol{\theta}, \boldsymbol{\theta}'), \forall t$. A Markov chain is said to have reached its *stationary* or *invariant* distribution – $\pi(\boldsymbol{\theta})$ if the distribution of $\boldsymbol{\theta}$ does not change according to time points t any more. Specifically, there exists some $\tilde{t}$ such that $p_{\tilde{t}}(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})$ and

$$\pi(\boldsymbol{\theta}) = \sum_{\tilde{\boldsymbol{\theta}}} \pi(\boldsymbol{\theta}) T_t(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}), \forall t \geq \tilde{t}. \tag{31.4}$$

The purpose of using MCMC in Bayesian inference is to help us sample from an otherwise difficult to evaluate posterior distribution. To achieve this goal, we are interested in constructing a Markov chain where the target posterior distribution is invariant. Often, we choose reversible homogeneous Markov chains in which the probability of a transition from the state $\boldsymbol{\theta}$ to the state $\boldsymbol{\theta}'$ is the same as the probability of a transition from $\boldsymbol{\theta}'$ to $\boldsymbol{\theta}$ under the distribution of states π. Equivalently,

$$\pi(\boldsymbol{\theta}) T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}') T(\boldsymbol{\theta}', \boldsymbol{\theta}). \tag{31.5}$$

The above condition is usually called *detailed balance*. It is straightforward to show that detailed balance implies invariance, i.e.,

$$\sum_{\boldsymbol{\theta}'} \pi(\boldsymbol{\theta}') T(\boldsymbol{\theta}', \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}) \sum_{\boldsymbol{\theta}'} T(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}). \tag{31.6}$$

It should be noted that detailed balance is a sufficient but not necessary condition for a distribution to be invariant (Neal, 1998).

Detailed balance ensures that once a Markov chain reaches its invariant distribution, subsequent states are samples from this invariant distribution. However, we generally do not know this invariant distribution which is the target posterior distribution. Instead, we hope the distribution of states at time t converges in distribution to its invariant distribution π as $t \to \infty$ regardless of its initial probability distribution of states $p_0(\boldsymbol{\theta})$. The Markov chain is *ergodic* if it holds this property. For a homogeneous Markov chain with an invariant distribution π, it is ergodic if the chain can traverse the entire support of π, i.e.,

$$v = \min_{\boldsymbol{\theta}} \min_{\boldsymbol{\theta}':\pi(\boldsymbol{\theta}')>0} T(\boldsymbol{\theta}, \boldsymbol{\theta}')/\pi(\boldsymbol{\theta}') > 0. \tag{31.7}$$

For a proof of this theorem, readers can refer to Neal (1998).

The simplest MCMC algorithm is perhaps the Gibbs sampler (Geman & Geman, 1984; Gelfand & Smith, 1990). Suppose we are interested in sampling from a joint distribution given by $p(\theta_1, \theta_2, \ldots, \theta_K)$ which is our target posterior distribution. Gibbs sampler works by repeatedly sampling each θ_k from their full conditional distributions. At the tth iteration, we

- sample $\theta_1^{(t)}$ according to the distribution given by $p(\theta_1^{(t)}|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)})$;
- sample $\theta_2^{(t)}$ according to the distribution given by $p(\theta_2^{(t)}|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)})$;

$$\vdots$$

- sample $\theta_K^{(t)}$ according to the distribution given by $p(\theta_K^{(t)}|\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{K-1}^{(t)})$.

The above steps together form a transition of state from $\boldsymbol{\theta}^{t-1}$ to $\boldsymbol{\theta}^t$ with probabilities $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ that leaves the target distribution invariant. Starting from an initial state $\boldsymbol{\theta}^{(0)}$, after simulating the Markov chain long enough, subsequent draws of $\boldsymbol{\theta}^{(t)}$ are treated as samples from the target posterior distribution.

31.3 Applications of MCMC in CDM

Similar to other types of psychometric modeling, instances of applications of MCMC in CDM are numerous. By no means the brief survey in this section is exhaustive, but rather to give readers flavors of the existing literature. The applications of MCMC in CDM can be traced back to earlier papers on the topic. In Junker and Sijtsma (2001), one of the earlier papers on CDM, the authors fit the deterministic inputs, noisy "and" gate (DINA) model and the noisy inputs, deterministic "and" gate (NIDA) model using the BUGS (Bayesian inference Using Gibbs Sampling) software (Thomas, Spiegelhalter, & Gilks, 1992).

While de la Torre (2008) provides a description of for estimating the DINA model by marginal maximum likelihood using the expectation-maximization (EM; Dempster, Laird, & Rubin, 1977) algorithm; the development of the EM algorithm for the higher-order DINA (HO-DINA) model is not trivial. As a result, in de la Torre and Douglas (2004), the HO-DINA model is estimated by a blocked Gibbs sampler (Geman & Geman, 1984; Gelfand & Smith, 1990; Gelman et al., 2013). The full-conditional distributions for HO-DINA do not have closed forms and are not easy to sample from directly. Therefore, the authors adopted the Metropolis algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). Instead of directly sampling from the full-conditional distributions, at each iteration, the Metropolis algorithm draws a proposed sample value or vector from a proposal distribution (usually a Gaussian distribution), and accepts or rejects it with an appropriately defined acceptance probability. To calculate the acceptance probability, only the unnormalized full conditional density function is required. This circumvents the difficulty of obtaining the normalizing constant when it cannot be derived analytically. The combination of Gibbs sampler and Metropolis algorithm is usually referred to as the Metropolis-within-Gibbs which is implemented in many general purpose Bayesian inference softwares (e.g., OpenBUGS; Lunn, Spiegelhalter, Thomas, & Best, 2009, JAGS; Plummer, 2005).

One difficulty of using the Metropolis algorithm is the tuning of the sampler. If the variance of the proposal distribution is large, the proposed sample is more likely to be further away from the current sample, which leads to low acceptance probabilities. Consequently, a large number of proposed samples are rejected before an acceptance, and the sampler rarely moves. On the other hand, small variance of the proposal distribution leads to high acceptance probabilities. But the proposed samples tend to be close to the current ones. As a result, the sampler moves slowly and does not explore the posterior distribution very efficiently. Therefore tuning is required so that the Markov chain is mixing at an optimal rate (Roberts, Gelman, & Gilks, 1997). Tuning a sampler could be a tedious task. Culpepper (2015) derived the closed forms of full-conditional distributions for DINA model so that the parameters can be directly sampled without using the Metropolis algorithm. In the same paper, the author also shows that the monotonicity assumption of the DINA model can be enforced by sampling the item parameters from a truncated bi-variate Beta distribution.

In the applications discussed so far, the Q-matrix (Tatsuoka, 1983) needs to be specified before the model can be estimated. In reality, the specification of the Q-matrix is not always straightforward and elements of the Q-matrix can be uncertain. Recognizing this limitation, DeCarlo (2012) proposes a Bayesian model to handle the uncertainty. Instead of treating all elements of the Q-matrix as fixed, the author specifies some of them as Bernoulli distributed random parameters, and assigns a Beta prior to the Bernoulli probabilities. The uncertain elements of the Q-matrix are recovered from examining the posterior distributions. OpenBugs software (Spiegelhalter, Thomas, Best, & Lunn, 2014) is used to estimate the model under the reparameterized DINA (RDINA; DeCarlo, 2012) model. DeCarlo and Kinghorn (2016) extend the approach to the case where none of the Q-matrix elements is fixed.

Furthermore, there has also been some other effort developing exploratory Bayesian methods for estimating CDM models without any prior knowledge of the Q-matrix except for the dimensions. Chung (2014) derives a Gibbs sampler for the DINA model and a Metropolis-within-Gibbs algorithm for the rRUM (reduced reparameterized unified model; Hartz, 2002) that include all elements of the Q-matrix as model parameters. The distribution of attribute patterns for examinees is modeled by a saturated categorical distribution, and the probabilities of the categories are given a Dirichlet prior. Thanks to the categorical-Dirichlet conjugacy, the probabilities of attribute patterns can be directly sampled from Dirichlet posterior distributions. By using a saturated categorical distribution, the author did not assume a particular factorization of the joint distribution of the attributes. Correlated attributes with different structures can be modeled in addition to independent attributes. However, the trade-off is the large number of parameters needs to be estimated. For a Q-matrix with K attributes, there are $2^K - 1$ probabilities for the attribute patterns. The Q-matrix is estimated similarly by using a categorical distribution. Item parameters for the DINA model can be sampled from truncated Beta distributions respecting the monotonicity assumption. Unfortunately, the full-conditional distributions of the item parameters for the rRUM model do not have closed forms. Thus, the Metropolis algorithm is used. Another example of the exploratory Bayesian approach can be found in Chen, Culpepper, Chen, & Douglas (2018). The paper deals with the same problem of estimating the DINA model without knowing the elements of the Q-matrix. Building on the development in understanding the identifiablity of the DINA model (Chen, Liu, Xu, & Ying, 2015; Liu, Xu, & Ying, 2012, 2013; Xu & Zhang, 2016), Chen et al. (2018) constrain the Q-matrix to be identified in their estimation procedure.

MCMC also aids the development and applications of more complex CDM models. For example, Li, Cohen, Bottge, and Templin (2016) introduce a longitudinal model that incorporates learning into CDM models. The attribute patterns for each student can change over time. It is modeled by a latent transition model. The transition matrix indicates the probability of transition from one attribute pattern to another. In this paper, several models with different transition matrices are fitted and compared using deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002).

Not only useful in estimating CDM models, MCMC also provides some of the most intuitive ways in checking model fit. Using the posterior samples, the posterior predictive model check (PPMC) method (Rubin, 1984; Gelman et al., 2013) calculates posterior distributions of various fit measures. It has been used in assessing the fit of IRT models (Sinharay, 2005). In CDM, Park, Johnson, and Lee (2015) examines the performance of PPMC using observed total-scores distribution, association of item pairs, and correlation of attribute pairs in assessing model fit.

As mentioned earlier, the review in this section is far from exhaustive. As more and more elaborate CDM models are developed in literature, we will certainly see more applications of MCMC.

31.4 A Gibbs Sampler for the Saturated Log-Linear CDM Model

In this section, we propose a new Gibbs sampler for the LCDM model. We analyze the ECPE data set as an illustration.

31.4.1 The Log-Linear CDM Model

The LCDM is similar to the Generalized DINA (GDINA; de la Torre, 2011) model in the sense that they all provide a flexible and general framework that encompasses many specific CDM models and can be viewed as a special case of the general diagnostic model (GDM; von Davier, 2008, 2014).

Under the LCDM, the probability of nth person answering the kth item correctly is

$$\text{logit}\, P_k(\boldsymbol{a}_n) = \lambda_{k0} + \sum_{d=1}^{K} \lambda_{kd} a_{nd} q_{kd} + \sum_{d=1}^{K} \sum_{d'>d} \lambda_{kdd'} a_{nd} a_{nd'} q_{kd} q_{kd'} + \ldots. \quad (31.8)$$

$a_{nd} \in \{0, 1\}$ with $a_{nd} = 1$ being the nth person has the dth attribute, and $a_{nd} = 0$ otherwise. Similarly, q_{kd} is 1 if the kth item measures the dth attribute, and 0 otherwise. λ_{k0} is the intercept, so a person who does not possess any of the skills measured in the test would have the probability of $\text{logit}^{-1}(\lambda_{k0})$ getting the kth item correct. λ_{kd} is the main effect for the dth attribute. And $\lambda_{kdd'}$ is the interaction effect for the dth and d'th attributes. Depending on the Q-matrix, some of the terms in (31.8) may be dropped. If an item only measures one attribute, there is only the intercept and one main effect. It should be noticed that some specific CDM models are nested within (31.8). For example, if only the highest order interaction and the intercept are retained, the LCDM reduces to the DINA. A saturated model includes the intercept, all main effects of the measured attributes, and all interaction terms associated with those attributes.

31.4.2 A Bayesian Formulation of the Reparameterized Saturated LCDM

For a general CDM with three attributes, there can be $2^3 = 8$ latent classes defined by the attribute patterns $\boldsymbol{a}$; therefore, under the unrestricted latent class model, there would be 8 item response probabilities that would need to be estimated for each item. The Q-matrix restricts the probabilities by enforcing certain equality constraints on the item response probabilities. For example, under the saturated LCDM, the probability of giving a correct response to an item by different people

who possess different subsets of the required attributes may be different. Suppose an item requires the first two attributes but not the third, so the kth row of the Q-matrix is $\boldsymbol{q}_k = (1, 1, 0)$. Then three people with attribute patterns $\boldsymbol{a} = (1, 1, 0)$, $(1, 0, 0)$, and $(0, 1, 0)$ may potentially receive different probabilities of giving a correct answer to this item. However, a person with the attribute pattern $(1, 1, 1)$ would have the same probability of giving a correct response as someone whose attribute pattern is $(1, 1, 0)$ due to the fact that the third attribute is not required by the item. As a result, there are $2^2 = 4$ probabilities associated with this item. Except for this restriction, the saturated model LCDM does not make any further constraints.

In the following Bayesian specification of the item-saturated LCDM, we use the natural probabilities as the model parameters rather than using the linear coefficients. To aid in the description of the model, we define the condensed attribute pattern $\boldsymbol{\omega}_{nk}$ for each individual n and item k, as the subvector of $\boldsymbol{a}_n$ corresponding to only the dimensions or attributes required by item k, i.e., $\boldsymbol{\omega}_{nk} = (\boldsymbol{e}_{d_1}, \ldots, \boldsymbol{e}_{d_m})^\top \boldsymbol{a}_n$, where $\boldsymbol{e}_d$ is the standard unit vector for dimension d with a 1 for element d and a zero everywhere else, and the $\boldsymbol{d}$ is an ordered index set $\boldsymbol{d} = \{m : q_{km} = 1\}$. In our three attribute example, with only the first two attributes required for an item, we have

$$\boldsymbol{\omega}_{nk} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \boldsymbol{a}_n.$$

Then the item response probability can be denoted by $p_{k(\boldsymbol{\omega}_{nk})} = P(X_{nk} = 1 | \boldsymbol{A} = \boldsymbol{a})$.

Formally, suppose we observe an N by K response matrix from N subjects answering K items and a K by D Q-matrix, then our Bayesian hierarchical formulation of the LCDM assumes

$$x_{nk} | \boldsymbol{\omega}_{nk}, \boldsymbol{p}_k \sim \text{Bernoulli}(p_{k(\boldsymbol{\omega}_{nk})}), \tag{31.9}$$

$$p_{k(\boldsymbol{\omega}_{nk})} \sim \text{Beta}(\alpha_k, \beta_k), \tag{31.10}$$

$$\boldsymbol{\alpha}_n | \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}), \tag{31.11}$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\boldsymbol{v}). \tag{31.12}$$

Conditional on the latent attributes required by a particular item $\boldsymbol{\omega}_{nk}$, a person gives a correct response with the probability $p_{k(\boldsymbol{\omega}_{nk})}$. We assume a Beta prior distribution for the vector of item response probabilities $\boldsymbol{p}_k$. A non-informative prior can be specified by giving the uniform Beta(1, 1); while, a Beta(0.5, 0.5) may be used if a researcher believes the item might have a higher discrimination among those with and without the required skills.

We do not assume a particular factorization of the joint distribution of the attributes. Instead, each of the possible 2^D attribute patterns is treated as a category. Then each person's attribute pattern follows a categorical distribution with probabilities of each possible attribute pattern governed by parameters $\boldsymbol{\pi} =$

$(\pi_1, \pi_2, \ldots, \pi_{2^D})^\top$. A Dirichlet hyper-prior with concentration parameters $\boldsymbol{v}$ is given to the categorical distribution parameters.

31.4.3 Monotonicity Constraint

The monotonicity assumption specifies a set of constraints that ensures the interpretability of CDM models in addition to the specification of the Q-matrix. Under the monotonicity assumption, mastering additional attributes would not lower the probability of giving a correct response, i.e.

$$P(X_{nk} = 1|\boldsymbol{a}_{n_1}) \geq P(X_{nk} = 1|\boldsymbol{a}_{n_2}), \tag{31.13}$$

whenever $\boldsymbol{\omega}_{n_1kd} \geq \boldsymbol{\omega}_{n_2kd}$ for all $d = 1, \ldots, D_k$, where D_k is the number of skills required by item k. Thus the item parameters in our Bayesian hierarchical formulation must satisfy

$$p_{k(\boldsymbol{\omega}_{n_1k})} \geq p_{k(\boldsymbol{\omega}_{n_2k})}, \text{ if } \omega_{n_1kd} = 1 \forall d \text{ s.t.} \omega_{n_2kd} = 1. \tag{31.14}$$

For the log-linear model, it is equivalent to constraining all main effects to be nonnegative and the coefficient of any interaction term to be no less than -1 times the largest main effect involved in the interaction (LCDM; Henson et al., 2009; Templin & Bradshaw, 2014).

31.4.4 A Gibbs Sampler

Conditional on the observed data for the kth item and class assignment for all people on this item, the item parameter is independent of everything else. So its full conditional distribution is

$$P(p_{k(\boldsymbol{w})}|\boldsymbol{x}_{(k)}, \boldsymbol{\omega}_{(k)}, \alpha_k, \beta_k) \propto \prod_{S_{k\boldsymbol{w}}=\{n:\, \boldsymbol{\omega}_{nk}=\boldsymbol{w}\}} p_{k(\boldsymbol{w})}^{x_{nk}} (1 - p_{k(\boldsymbol{w})})^{1-x_{nk}} P(p_{k(\boldsymbol{w})}|\alpha_k, \beta_k), \tag{31.15}$$

where $\boldsymbol{x}_{(k)}$ denotes the vector of all item responses to item k and $\boldsymbol{\omega}_{(k)}$ denotes the set of item-specific attribute patterns for item k.

Due to the standard Bernoulli-Beta conjugacy, (31.15) has a closed form, i.e.

$$p_{k(\boldsymbol{w})}|\boldsymbol{x}_{(k)}, \boldsymbol{\omega}_{(k)}, \alpha_k, \beta_k \sim \text{Beta}\left(\alpha_k + \sum_{n \in S_{k\boldsymbol{w}}} x_n, \beta_k + |S_{k\boldsymbol{w}}| - \sum_{n \in S_{k\boldsymbol{w}}} x_n\right). \tag{31.16}$$

The monotonicity constraint in (31.14) implies that $p_{j(\boldsymbol{\omega}_{ij})}$ is bounded above by

$$U_{p_{k(\boldsymbol{w})}} = \inf_{\boldsymbol{w}'}\{p_{k(\boldsymbol{w}')} : \ w'_d \geq w_d \ \forall \ d \in \{1, 2, \ldots, D_k\}\}, \tag{31.17}$$

and bounded below by

$$L_{p_{k(\boldsymbol{w})}} = \sup_{\boldsymbol{w}'}\{p_{k(\boldsymbol{w}')} : \ w'_d \leq w_d \ \forall \ d \in \{1, 2, \ldots, D_k\}\}. \tag{31.18}$$

It follows that the full conditional distribution in (31.16) should be truncated, i.e.,

$$p_{k(\boldsymbol{w})}|\boldsymbol{x}_{(k)}, \boldsymbol{\omega}_{(k)}, \alpha_k, \beta_k \sim \text{Beta}\left(\alpha_k + \sum_{n \in S_{k\boldsymbol{w}}} x_{nk}, \ \beta_k + |S_{k\boldsymbol{w}}| - \sum_{n \in S_{k\boldsymbol{w}}} x_n\right)$$
$$I_{(L_{p_{k(\boldsymbol{w})}}, U_{p_{k(\boldsymbol{w})}})}(p_{k\boldsymbol{w}}), \tag{31.19}$$

where $I_{(u,\ell)}(p)$ indicates the distribution is truncated to the interval (u, ℓ).

The full conditional distribution for $\boldsymbol{a}_n$ is

$$P(\boldsymbol{a}_n|\boldsymbol{x}_n, \boldsymbol{p}, \boldsymbol{\pi}) \propto \prod_{k=1}^{K} P(x_{nk}|\boldsymbol{p}_{k(\boldsymbol{\omega}_{nk})}) P(\boldsymbol{a}_n|\boldsymbol{\pi}). \tag{31.20}$$

Since the distribution is discrete, (31.20) can be easily normalized:

$$P(\boldsymbol{a}_n|\boldsymbol{x}_n, \boldsymbol{p}, \boldsymbol{\pi}) = \frac{\prod_{k=1}^{K} P(x_{nk}|\boldsymbol{p}_{k(\boldsymbol{\omega}_{nk})}) P(\boldsymbol{a}_n|\boldsymbol{\pi})}{\sum_{\boldsymbol{a}_n} \prod_{k=1}^{K} P(x_{nk}|\boldsymbol{p}_{k(\boldsymbol{\omega}_{nk})}) P(\boldsymbol{a}_n|\boldsymbol{\pi})}. \tag{31.21}$$

And the closed form full-conditional distribution is

$$\boldsymbol{a}_n|\boldsymbol{x}_n, \boldsymbol{p}, \boldsymbol{\pi} \sim \text{Categorical}(u_1, u_2, \ldots, u_{2^D}), \tag{31.22}$$

where the probabilities $u_1, u_2, \ldots, u_{2^D}$ are given in (31.21).

Finally, the full conditional distribution for hyper-parameters $\boldsymbol{\pi}$ is

$$P(\boldsymbol{\pi}|\boldsymbol{a}, \boldsymbol{v}) \propto \prod_{n=1}^{N} P(\boldsymbol{a}_n|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\boldsymbol{v}). \tag{31.23}$$

The standard categorical-Dirichlet conjugacy leads to the closed form:

$$\boldsymbol{\pi}|\boldsymbol{a}, \boldsymbol{v} \sim \text{Dirichlet}(\boldsymbol{v} + (c_1, c_2, \ldots, c_{2^d})), \tag{31.24}$$

where the elements of the vector $(c_1, c_2, \ldots, c_{2^D})$ are the counts of observations in each class.

Update steps for each iteration of the Gibbs sampler are:

1. Draw the item parameters $p_{k\boldsymbol{w}}$ for each item and item-specific attribute pattern $\boldsymbol{w}$ from the full conditional distributions in (31.19);
2. Draw the latent class assignment $\boldsymbol{a}_n$ for each person from the full conditional distributions in (31.22);
3. Draw the hyper-parameter $\boldsymbol{\pi}$ from the full conditional distribution in (31.24).

31.4.5 Linear Transformation of Model Parameters

The model parameters from the reparameterized saturated model can be easily transformed back to the log-linear model parameters by solving a linear system of equations. For simplicity, consider the case where there are $d = 2$ attributes. Under the saturated log-linear model, $2^2 = 4$ linear coefficients are needed. The logit link links the probabilities to the linear combinations of the attributes, i.e.

$$\boldsymbol{T}\boldsymbol{\lambda}_k = \text{logit}\, \boldsymbol{p}_k, \tag{31.25}$$

where

$$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{k0} \\ \lambda_{k1} \\ \lambda_{k2} \\ \lambda_{k12} \end{bmatrix}, \text{logit}\, \boldsymbol{p}_k = \begin{bmatrix} \text{logit}\, p_{k(00)} \\ \text{logit}\, p_{k(10)} \\ \text{logit}\, p_{k(01)} \\ \text{logit}\, p_{k(11)} \end{bmatrix}.$$

In the above notations, $\boldsymbol{t}_{m\cdot}$ denotes the mth row of the T matrix. Multiplying the inverse of the attribute pattern matrix to both sides of (31.25) gives the log-linear model parameters, i.e.,

$$\boldsymbol{\lambda}_k = \boldsymbol{T}^{-1}\, \text{logit}\, \boldsymbol{p}_k. \tag{31.26}$$

To get the posterior distribution of the log-linear model parameters, simply apply the linear transformation in (31.26) to the posterior samples of the reparameterized saturated model parameters.

31.5 A Bayesian Analysis of the ECPE Dataset

In this section, we analyze the ECPE dataset as a demonstration. The ECPE dataset is available in the *R* CDM package (George, Robitzsch, Kiefer, Groß, & Ünlü, 2016). It has been analyzed in previous research (e.g., Templin & Bradshaw, 2014; Templin & Hoffman, 2013; von Davier, 2014). The dataset consists of the binary

responses from 2922 examinees to 28 items. Three attributes are specified in the Q-matrix: morphosyntactic rules, cohesive rules, and lexical rules. However, none of the items measures all three attributes. Among the 28 items, 9 measure two attributes, and the rest measure one. We fit the reparameterized saturated model and finally transformed parameters back to the log-linear model parameterization. Non-informative priors are used: uniform Beta(1, 1) for item parameters, and Dirichlet(1, 1, . . . , 1) for the hyper-prior of class allocations. Furthermore, the monotonicity is enforced by imposing constraints to item parameters as in (31.17) and (31.18).

Diagnosing the convergence of the Markov chains is important in applications of MCMC. The MCMC theory guarantees that the Gibbs sampler will eventually converge to the target posterior distribution as the number of draws goes to infinity. But, in reality, the number of draws we can afford is always finite and often limited. Therefore, we need to assess whether we can treat MCMC draws approximately as samples from the posterior distribution after a certain number of initial draws. Over the years, many MCMC convergence diagnostics have been proposed. Some of the popular examples include the potential scale reduction factor (PSRF; Gelman & Rubin, 1992), the multivariate PSRF (MPSRF; Brooks & Gelman, 1998), and the Geweke convergence diagnostic (Geweke, 1992). Here we use two common graphical methods to assess the convergence of our Gibbs sampler. Four parallel chains with different starting values are simulated. We run each chain for 5000 iterations. To demonstrate the evidence of convergence, Fig. 31.1 shows the trace of the first 500 iterations of each chain for two parameters. The plots suggest that the chains quickly converged to their target stationary distributions regardless of different starting values. We can also monitor the convergence by examining the k-lag autocorrelation functions. The k-lag autocorrelation is the correlation between every draw and its kth lag. Intuitively, a Markov chain that generates highly correlated samples would take a long time to explore the entire target distribution.

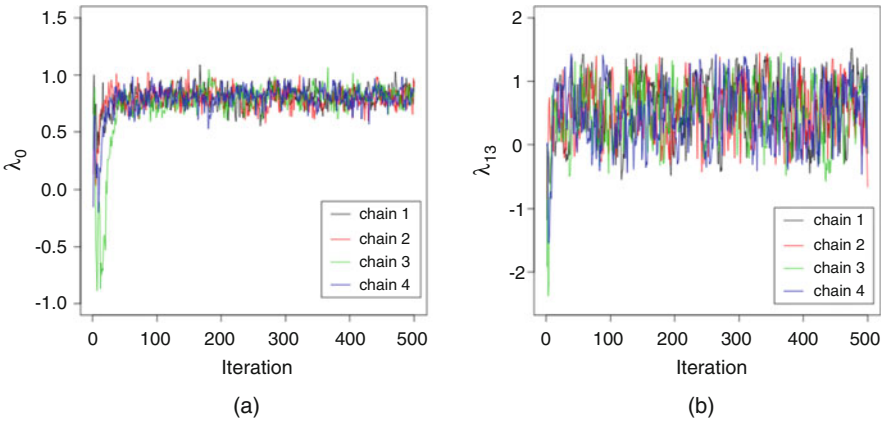


Fig. 31.1 k-lag autocorrelation of two parameters. (**a**) λ_0 – Item 1. (**b**) λ_{13} – Item 11

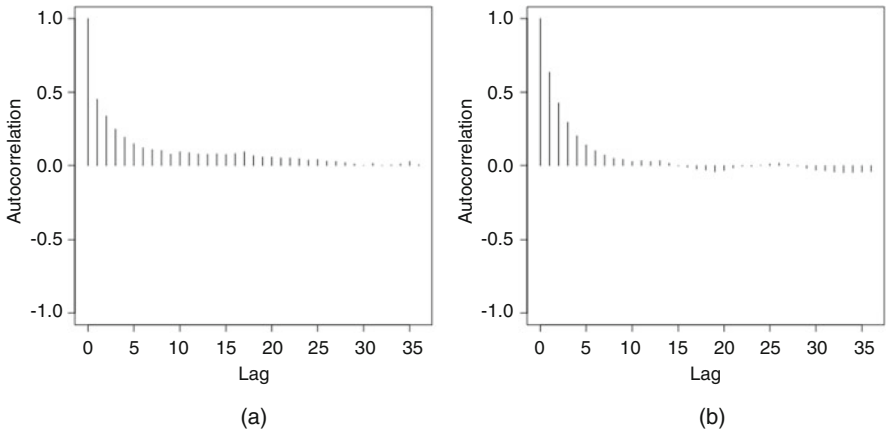


Fig. 31.2 Traceplot of two parameters. (**a**) λ_0 – Item 1. (**b**) λ_{13} – Item 11

We would hope that the autocorrelation between samples quickly shrink to around zero as the lag k increases. From Fig. 31.2, we can see that the autocorrelation decreases very quickly as the lag increases in both cases. It is consistent with the quick convergence and good mixing shown in the trace plots. Based on the convergence diagnostics, our Gibbs sampler seems to perform very well. We decide to treat the first 1000 from each chain as burn-ins and use the rest for the purpose of posterior inference.

Table 31.1 shows the *Expected a Priori* (EAP) estimates and posterior standard deviations of item parameters under the LCDM. Comparing the EAP estimates to the maximum likelihood estimates (MLE) reported in previous literature (see Table 1 in Templin & Bradshaw, 2014), it seems that the EAP estimates are almost identical to the MLE for the items measuring single attributes. However, differences exist between the EAP estimates and the MLE for items measuring two attributes except for the second item where the Bayesian approach yields similar estimates to maximum likelihood.

As pointed out by Templin and Bradshaw (2014), a closer examination of the MLE for two attribute items reveals that many of the ML estimates appeared on the boundary. For example, the main effect of the morphosyntactic rules for the first item is estimated to be zero in Templin and Bradshaw (2014). The standard asymptotic theory does not give any useful approximation to the limiting distribution of the MLE when the ML estimate lies on the boundary. This is reflected by the zero standard error reported in Templin and Bradshaw (2014). The MLE for some of the interaction effects also suffer this problem. They are estimated to be very close to the boundary imposed by the monotonicity constraint. Large standard errors are also observed for many of the estimated effects. These are symptoms of under-identification. Von Davier (2014) also discussed this problem. While an infinitely large sample size would allow the parameters to be estimated precisely and away

Table 31.1 ECPE Bayesian estimates of LCDM item parameters

Item	λ_0	λ_1	λ_2	λ_3	λ_{12}	λ_{13}	λ_{23}
1	0.81(0.08)	0.51(0.4)	0.65(0.23)		0.61(0.53)		
2	1.03(0.08)		1.25(0.15)				
3	−0.34(0.08)	0.76(0.42)		0.35(0.13)		0.52(0.44)	
4	−0.14(0.08)			1.69(0.1)			
5	1.07(0.08)			2.02(0.16)			
6	0.87(0.08)			1.68(0.14)			
7	−0.09(0.08)	1.59(0.67)		0.93(0.13)		0.32(0.7)	
8	1.47(0.09)		1.92(0.24)				
9	0.12(0.07)			1.19(0.1)			
10	0.05(0.06)	2.05(0.15)					
11	−0.05(0.08)	1.19(0.6)		0.96(0.14)		0.39(0.64)	
12	−1.79(0.12)	0.62(0.46)		1.31(0.17)		0.88(0.49)	
13	0.66(0.06)	1.61(0.15)					
14	0.17(0.05)	1.36(0.12)					
15	0.99(0.08)			2.12(0.16)			
16	−0.09(0.08)	1.34(0.57)		0.87(0.13)		0.13(0.59)	
17	1.34(0.09)		0.65(0.41)	0.61(0.27)			0.2(0.52)
18	0.92(0.08)			1.4(0.13)			
19	−0.2(0.08)			1.85(0.11)			
20	−1.43(0.1)	0.97(0.58)		0.94(0.15)		0.67(0.61)	
21	0.16(0.08)	0.98(0.54)		1.13(0.14)		0.12(0.58)	
22	−0.87(0.09)			2.24(0.11)			
23	0.66(0.08)		2.06(0.19)				
24	−0.69(0.09)		1.54(0.12)				
25	0.09(0.05)	1.14(0.11)					
26	0.16(0.08)			1.12(0.1)			
27	−0.89(0.06)	1.7(0.1)					
28	0.56(0.08)			1.75(0.12)			

Note: Attribute 1 – Morphosyntactic rules; Attribute 2 – Cohesive rules; Attribute 3 – Lexical rules.

from boundaries (when the true parameters are away from the boundary), we work with a limited sample size in reality.

One solution is to impose an attribute hierarchy which effectively reduces the number of parameters to be estimated (Templin & Bradshaw, 2014). The introduced Bayesian method explores another approach. The use of priors provides regularization and enables more parameters to be reasonably estimated (Gelman et al., 2013). The EAP estimates for single attribute items are well-regularized with small posterior standard deviations. While the posterior standard deviations for the two attribute items are larger, they are still reasonable. The largest posterior standard deviation is 0.67 compared to the largest standard error of 1.62 reported in previous research.

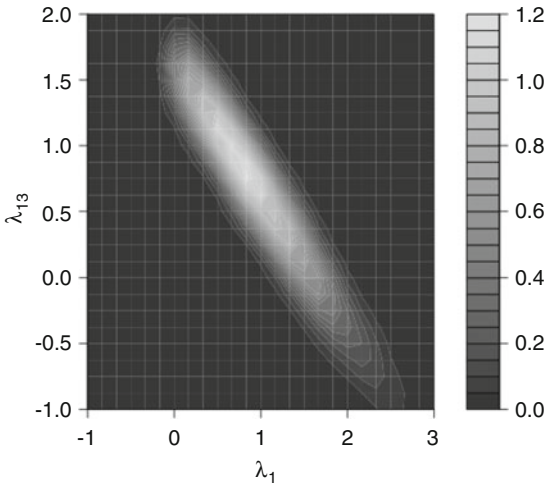


Fig. 31.3 Joint posterior density of λ_1 and λ_{13} for Item 20

The posterior samples can also provide useful information in assessing various aspects of model fit. For example, one source of the misfit is the misspecification of the Q-matrix. Considering the EAP estimates and the associated posterior standard deviations of λ_1 and λ_{13} for item 20 in Table 31.1, one might suspect that morphosyntactic rules are not measured by the item. Both marginal posterior distributions of λ_1 and λ_{13} might have considerable densities around zero. However, if the item doesn't measure this attribute, it would imply that $\lambda_1 = \lambda_{13} = 0$. In other words, we need to inspect the joint posterior distribution of these two effects. Samples from the posterior simulation can achieve this with little effort. Figure 31.3 clearly shows that the origin is away from the region where the joint posterior density is concentrated.

Posterior samples can also be used to check the plausibility of particular CDM models. For example, if the DINA is plausible, it would suggest that the main effects and lower order interactions are all zeros. Since each item measures at most two attributes in the ECPE dataset, we only need to examine the joint posterior distribution of the main effects. Figure 31.4 suggests that DINA is more plausible for Item 1 than Item 11.

31.6 Discussion

MCMC algorithms and Bayesian methods in general will certainly continue to play an important role in the development of various CDM models. In this chapter, we briefly reviewed some of the applications of the MCMC in CDM literature. We also introduced a Gibbs sampler for estimating the saturated LCDM model. With the reparameterization, the sampler is able to take advantage of the standard

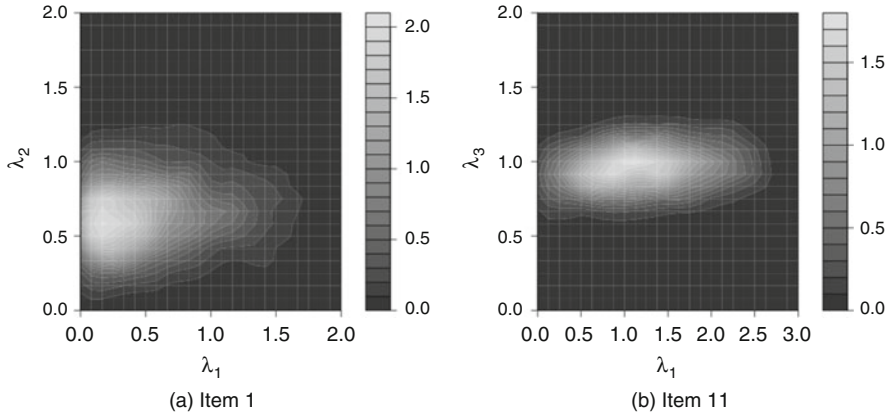


Fig. 31.4 Joint posterior density of the main effects for Items 1 and 11

conjugacy results thus the sampler does not require any tuning. Even though we introduced the sampler for the saturated LCDM, the approach can be modified to fit a wide spectrum of specific CDM models by imposing additional constraints to the saturated LCDM model.

References

Brooks, S., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). Handbook of Markov chain Monte Carlo. *Handbook of Markov chain Monte Carlo*. https://doi.org/10.1201/b10905

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika, 75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Chen, Y., Culpepper, S. A., Chen, Y., & Douglas, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika, 83*(1), 89–108. Retrieved from https://doi.org/10.1007/s11336-017-9579-4

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*(510), 850–866. Retrieved from https://doi.org/10.1080/01621459.2014.934827

Chung, M.-t. (2014). *Estimating the Q-matrix for cognitive diagnosis models in a Bayesian framework*. Doctoral dissertation, Columbia University. Retrieved from https://search.proquest.com/docview/1548332406

Culpepper, S. A. (2015) Bayesian estimation of the DINA model with Gibbs sampling. *Journal of Educational and Behavioral Statistics, 40*(5), 454–476. Retrieved from https://doi.org/10.3102/1076998615595403

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement, 36*(6), 447–468. Retrieved from https://doi.org/10.1177/0146621612449069

DeCarlo, L. T., & Kinghorn, B. R. (2016). *An Exploratory Approach to the Q-Matrix Via Bayesian Estimation.* (Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC)

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B, 39*(1), 1–38. Retrieved from http://www.jstor.org/stable/2984875, https://doi.org/10.2307/2984875

de la Torre, J. (2008). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130. Retrieved from https://doi.org/10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. Retrieved from https://doi.org/10.1007/BF02295640

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*(410), 398–409. Retrieved from https://doi.org/10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3 ed.). Boca Raton: Chapman and Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*. https://doi.org/10.1214/ss/1177011136

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6*(6), 721–741. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22499653, https://doi.org/10.1109/TPAMI.1984.4767596

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software, 74*(2). Retrieved from http://www.jstatsoft.org/v74/i02/ https://doi.org/10.18637/jss.v074.i02

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4*. Oxford: Oxford University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82*(4), 711–732. Retrieved from https://doi.org/10.1093/biomet/82.4.711

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Doctoral dissertation, University of Illinois at Urbana-Champaign. Retrieved from http://hdl.handle.net/2142/87393

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*(1), 97–109. Retrieved from http://www.jstor.org/stable/2334940

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210. Retrieved from https://doi.org/10.1007/s11336-008-9089-5

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272. Retrieved from https://doi.org/10.1177/01466210122032064

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics, 2009*, 1–18. Retrieved from http://www.hindawi.com/journals/jps/2009/537139/, https://doi.org/10.1155/2009/537139

Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement, 76*(2), 181–204. Retrieved from https://doi.org/10.1177/0013164415588946

Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36*(7), 548–564. Retrieved from https://doi.org/10.1177/0146621612456591

Liu, J., Xu, G., & Ying, Z. (2013). Theory of self-learning Q-matrix. *Bernoulli, 19*(5), 1790–1817. Retrieved from https://doi.org/10.3150/12-BEJ430

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009) The BUGS project: Evolution, critique and future directions. *Statistics in Medicine, 28*(25), 3049–3067. Retrieved from https://doi.org/10.1002/sim.3680

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics*. https://doi.org/10.1063/1.1699114

Neal, R. M. (1998). Probabilistic inference using Markov chain Monte Carlo methods. *Technical Report, 1*, 1–144. Retrieved from papers2://publication/uuid/0C88167E-5379-4E4E-A9E4-007ABA4F716D, https://doi.org/10.1021/np100920q

Park, J. Y., Johnson, M. S., & Lee, Y. S. (2015). Posterior predictive model checks for cognitive diagnostic models. *International Journal of Quantitative Research in Education, 2*(3/4), 244. Retrieved from http://www.inderscience.com/link.php?id=71738, https://doi.org/10.1504/IJQRE.2015.071738

Plummer, M. (2005). JAGS: Just another Gibbs sampler. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (dsc 2003).*

Roberts, G. O., Gelman, A., & Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability, 7*(1), 110–120. https://doi.org/10.1214/aoap/1034625254

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12*(4), 1151–1172. Retrieved from http://projecteuclid.org/euclid.aos/1176346785, https://doi.org/10.1214/aos/1176346785

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*(4), 375–394. Retrieved from https://doi.org/10.1111/j.1745-3984.2005.00021.x

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, 64*(4), 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2014). *OpenBUGS User Manual* (Vol. 164). Retrieved from http://www.openbugs.net/Manuals/Manual.html

Tatsuoka, K. K. (1983) Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354. Retrieved from https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317–339. Retrieved from https://doi.org/10.1007/s11336-013-9362-0

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50. https://doi.org/10.1111/emip.12010

Thomas, A., Spiegelhalter, D. J., & Gilks, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4*. Oxford: Oxford University Press.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *The British Journal of Mathematical and Statistical Psychology, 61*(Pt 2), 287–307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series, 2014*(2), 1–13. Retrieved from https://doi.org/10.1002/ets2.12043

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*(3), 625–649. https://doi.org/10.1007/s11336-015-9471-z

Index

M. von Davier, Y.-S. Lee (eds.), *Handbook of Diagnostic Classification Models*,
Methodology of Educational Measurement and Assessment,
https://doi.org/10.1007/978-3-030-05584-4

N

O

S