Chapter 4
Standardized Diagnostic Assessment Design and Analysis: Key Ideas from Modern Measurement Theory

Hye-Jeong Choi, André A. Rupp, and Min Pan

4.1 Introduction

Traditional standardized standards-based assessments created by professional agencies and partially standardized standards-based assessments made by teachers for assessments at the end of a unit, chapter, or term can be reliable indicators of general states of proficiency for groups of students. In short, they serve general monitoring and accountability purposes in selected key domains such as reading, mathematics, and science rather well. However, as Linn (1986) emphasized, they typically have very little or no instructional uses:

> a test that reliably rank orders students in terms of global test scores provides a teacher with relatively little information about the nature of a student's weaknesses, errors, or gaps. For example, the knowledge that a student scores, say, in the 10th percentile on a standardized arithmetic test suggests the student has a general weakness in the area of arithmetic relative to his or her peers. However, such a score does not, by itself, indicate the source of the problem or what should be done to improve the student's level of achievement; that is, it lacks diagnostic information. (p. 1158)

The seemingly increasing dissatisfaction in the field of education with the structure and potential uses of standardized standards-based assessments for guiding and

H-J. Choi
Office of Research and Department of Psychology,
University of Nebraska-Lincoln, Athens, GA, USA
e-mail: hchoi3@unl.edu

A.A. Rupp (✉)
Department of Human Development and Quantitative Methodology (HDQM),
University of Maryland, College Park, MD, USA
e-mail: ruppandr@umd.edu

M. Pan
Department of Measurement, Statistics, and Evaluation,
University of Maryland, College Park, MD, USA
e-mail: minr.l@foxmail.com

M. Mo Ching Mok (ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific*,
Education in the Asia-Pacific Region: Issues, Concerns and Prospects 18,
DOI 10.1007/978-94-007-4507-0_4,

evaluating the students' fine-tuned knowledge state motivated the development of more *diagnostic assessments*. Diagnostic assessments play a key role in establishing an alignment between developmental theories about learning in a domain, curricular objectives as set forth by policy documents, teacher practice in the classroom, and actual learning gains made by students (e.g., Leighton and Gierl 2007, 2011).

4.1.1 Assessment Of, For, and As Learning

The current literature on modern educational measurement for diagnostic assessment purposes makes a distinction between assessments *of*, *for*, and *as* learning, which helps to differentiate the various layers of interpretations drawn from them and the diverse uses to which they are put (e.g., O'Reilly et al. 2008; Mok 2010).

The phrase *assessment of learning* suggests that one purpose of assessments is to identify the achievement of the students at the end of a learning cycle to obtain a rich and sufficiently detailed picture of the degree to which students have met their targeted learning objectives. The information gathered from an assessment can support summative interpretations that allow for overall comparisons of how individual students perform relative to their peers.

The phrase *assessment for learning* suggests that the purpose of an assessment can also be to monitor the continual, ongoing learning process in order to provide directive and supportive feedback in a scaffolding process. The information is collected to seek for answers as to what underlying mechanisms drive the problem-solving strategies enacted by the students so as to make the learning process most efficient, effective, and engaging for the students.

The phrase *assessment as learning* suggests that the purpose of assessment is to make students self-directed by improving their level of metacognition. The process of assessment thus induces the cultivation of a capacity for goal setting, self-monitoring of the learning process, self-assessment of achievement, self-motivation, and self-regulation to enhance further learning.

In terms of assessment for learning in particular, what many teachers seek to guide their day-to-day instructional practice are more fine-grained descriptions of students' proficiency profiles, which are necessary to designing effective instructional interventions that make students efficacious (i.e., efficient and effective) in the targeted domains. Teachers continually collect potentially diagnostic information in informal or partially standardized ways on a daily basis. For instance, teachers may ask questions regarding what concepts or strategies students have mastered and which ones they are still struggling with; they may ask specifically why some students do not understand a particular aspect of what they have taught in class, or they may inquire about whether it is necessary to create certain types of additional opportunities for practice in class. In short, teachers are constantly concerned with how they can construct classroom environments which fit individual student's current learning needs best.

4.1.2 Measurement Models for Diagnostic Assessment Data

Traditional measurement models that can support inferences from summative assessments for quantitative rank-order purposes include predominantly models from the fields of *classical test theory* (CTT) (e.g., Lord and Novick 1968; Crocker and Algina 2006) and *item response theory* (IRT) (e.g., de Ayala 2009; Yen and Fitzpatrick 2006) even though *factor-analytic* (FA) models (e.g., McDonald 1999) can serve these purposes as well. However, the score reports created on the basis of data calibrations with these models are, at best, only partially useful for supporting more formative interpretations for qualitative diagnostic purposes.

Typically, CTT, IRT, and FA models are applied to large-scale standardized standards-based assessments of learning whose operational construct is defined at a rather coarse level of cognitive grain size thus leading to relatively coarse descriptions of students' proficiency levels in the target domain. In contrast, *diagnostic classification models* (DCMs) (e.g., Rupp and Templin 2008; Rupp et al. 2010) are models that are particularly suitable for large-scale standardized assessments for learning whose operational construct is defined at a finer level of cognitive grain size thus supporting more nuanced descriptions about students' proficiency profiles.

In this chapter, we present a few key ideas that are relevant to developing cognitively diagnostic assessments for learning and scaling them with DCMs. Specifically, in the next section, we present a key framework for principled assessment design that can be employed in powerful ways for developing cognitive diagnostic assessments. In the section after that, we introduce a unified specification and estimation framework for DCMs and illustrate its utility for operationalizing different cognitive theories of responding. In the final main section, we present a real-data analysis of a small section of a newly developed diagnostic mathematics assessment to illustrate how DCMs can be used for calibrating the instrument and classifying the students into different proficiency profiles.

4.2 Evidence-Centered Design

Some form of applied cognitive theory (e.g., influenced by information-processing or socio-cognitive perspectives) is necessary to design any test whose items or tasks are supposed to reflect the essential knowledge, skills, and abilities that are to be measured (NRC 2001). Arguably, the explicit focus on fine-grained proficiency profiles for students that can inform learning processes in an assessment for learning sense puts the explication and operationalization of applied cognitive theories at the forefront of diagnostic assessment design. In this chapter, we focus on an important design framework called *evidence-centered design* (ECD).

The ECD (Mislevy et al. 2003, 2004) framework provides a formal structure for *evidence-based reasoning* that provides guidance to interdisciplinary teams of experts who are charged with developing a wide range of assessments for a wide range of purposes. Despite its generality, its power for structuring assessment development, implementation, and score reporting is arguably most evident for assessments that involve *complex performance-based tasks*. The reason for this is that the number of decisions about designing tasks with appropriate constraints, identifying suitable task products, identifying individual pieces of evidence and scoring them, aggregating these scores with the help of modern statistical models, and reporting these scores back to students and stakeholders are much larger and arguably more complex in these contexts than in assessments that employ more selected-response formats.

The core purpose of diagnostic assessment development from an ECD framework perspective is the development of coherent *evidentiary arguments* in an *assessment narrative* about students that can serve as assessment of and assessment for learning, depending on the desired primary purpose of a particular assessment. The structure of the evidentiary arguments that are used in the assessment narrative can be described with the aid of terminology first introduced by Toulmin (1958).

An evidentiary argument is constructed through a series of logically connected *claims or propositions* that are supported by data through *warrants* and *backing* and can be subjected to *alternative explanations*. In diagnostic assessments, data consist of students' observed responses to particular tasks and the salient features of those tasks, claims concern examinees' proficiency as construed more generally, and warrants posit how responses in situations with the noted features depend on proficiency. Statistical models such as DCMs provide the mechanism for evaluating and synthesizing the evidentiary value in a collection of typically overlapping, often conflicting, and sometimes interdependent observations.

In concrete terms, the ECD framework allows one to distinguish the different structural elements and the required pieces of evidence in narratives such as the following:

> Jamie has most likely mastered basic addition (*claim*), because she has answered correctly a mathematical problem about adding up prices in a supermarket (*data*). It is most likely that she did this because she applied all of the individual addition steps correctly (*backing*) and the task was designed to force her to do that (*backing*). She may have used her background knowledge to estimate the final price of her shopping cart (*alternative explanation*), but that is unlikely given that the final price is exactly correct (*refusal*).

The ECD framework specifies five different assessment design components, which are shown in Fig. 4.1 below.

Guided by the theory-driven process of analyzing and modeling the key facets of expertise in a domain, the core elements in the ECD framework include (1) the *student models*, which formalize the postulated proficiency structures for different tasks, (2) the *task models*, which formalize which aspects of task performance are coded in what manner, and (3) the *evidence models*, which are the psychometric models linking those two elements. These three core components are complemented by (4)

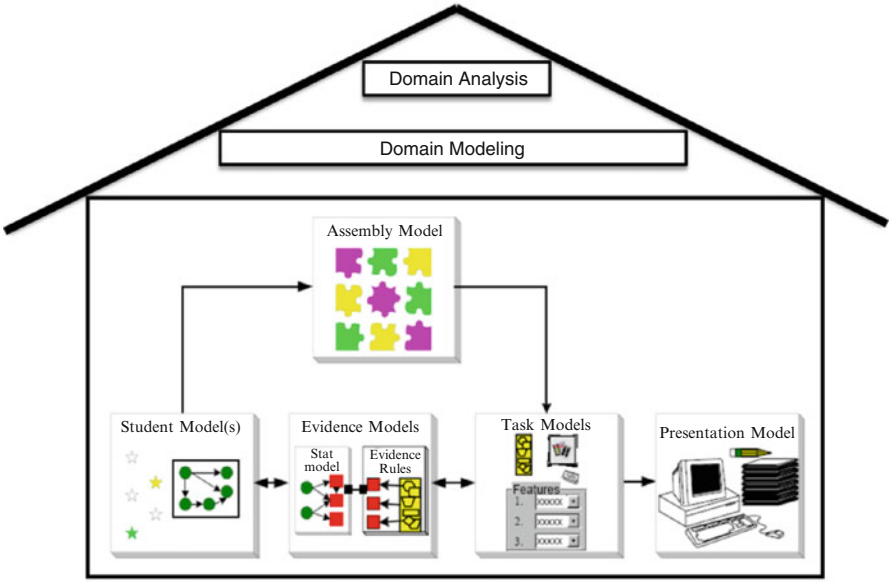

Fig. 4.1 The ECD model (Adapted from Mislevy et al. 2004)

the *assembly model*, which formalizes how these three elements are linked in the assessment, and (5) the *presentation model*, which formalizes how the assessment tasks are being presented.

Specifically, the *student model* is motivated by the learning theory that underlies the diagnostic assessment system. It specifies the relevant variables or aspects of learning that we want to assess at a grain size that suits the purpose of the diagnostic assessment. As many of the characteristics of learning that we want to assess are not directly observable, the student model provides a probabilistic or proxy model for making claims about the state, structure, and development of a more complex underlying system. This might concern a trait or a behavioral disposition in a traditional assessment. In more innovative diagnostic assessments in education such as a game or simulation, it could instead concern the models or strategies a student seems to employ in various situations, or the character or interconnectivity of his or her skills when dealing with certain kinds of situations in a discipline.

To make claims about learning as reflected through changes in the attributes in the student model, we thus have to develop a pair of *evidence models*. The *evaluation component* of the evidence model specifies the salient features of whatever the student says, does, or creates in the task situation, as well as the rules for scoring, rating, or otherwise categorizing the salient features of the assessment. The *probability or statistical component* of the evidence model specifies the rules by which the evidence collected in the evaluation is used to make assertions about the student

model. This means that a suitable statistical model such as a DCM needs to be selected for summarizing observed information contained in indicator variables via statistically created, and typically latent, variables. The statistical model provides the machinery for updating beliefs about student model variables in light of this information. Taken together, evidence models provide a chain of inferential reasoning from observable performance to changes that we believe are significant in a student's cognitive, social, emotional, moral, or other forms of development.

The *task model* provides a set of specifications for the environment in which the student will say, do, or produce something. That is, the task model specifies the conditions and forms under which data are collected, and the variables in a task model are motivated by the nature of the interpretations the assessment is meant to support. Data collected in such models are not restricted to traditional formal, structured, pencil-and-paper assessments and can include information about the context, the student's actions, and the student's past history or particular relation to the setting.

The *assembly model* describes how these different components are combined for answering particular questions about learning in a given assessment situation. Using the analogy of *reusable design templates* within a task bank, the assembly model describes which task model, evidence model, and student model components are linked for a particular assessment or subsections of an assessment. The idea of a reusable design template is similar to the idea of automatic task generation within the general cognitive design system (e.g., Embretson, 1998) framework. However, rather than striving for an automatic generation, the ECD framework strives for principled construction under constraints that will result in tasks that are comparable to one another, both substantively and statistically.

Similarly, the *presentation model* describes whether modes of task and product presentation change across different parts of the assessment and what the expected implications of these changes are. In practice, ECD models for a given assessment are constructed jointly and refined iteratively because the full meaning of any model only emerges from its interrelationship with other components.

ECD has been successfully applied in different fields. *PADI*, *ECDLarge* and *NetPASS* are comprehensive ongoing assessment projects that are based on ECD. Specifically, PADI aims at developing assessments of science inquiry that combine new developments in cognitive psychology, science inquiry, as well as measurement theories and techniques (e.g., Mislevy and Riconscente 2005; see also http://padi.sri.com/index.html). ECDLarge is a successor to the PADI project that focuses on the application of the ECD framework to the development of large-scale assessments (see http://ecd.sri.com/index.html for more information). The NetPASS project is concerned, in part, with developing an authoring tool and simulation-based learning and assessment environment to train network engineers within the context of Cisco Networking Academy Program (e.g., Levy and Mislevy 2004; Mislevy et al. 2003; Rupp et al. in press; West et al. 2009; see also http://cisco.netacad.net/public/index.html). The set of applications cited here, taken together, illustrate the power of the ECD framework for developing a wide range of assessments that can support a wide range of inferences including fine-grained diagnostic feedback for

formative assessment purposes as well as more coarse-grained feedback for summative accountability purposes.

The previous presentation is not meant to suggest that individual teachers have to think about the ECD framework during their day-to-day practice. However, we believe that teachers may find the language, conceptualization, and key assessment principles embedded within the ECD framework quite accessible and useful for shaping their own professional understanding. The ECD framework can also be very powerful for professional development purposes at the district or state level because it provides a coherent frame for structuring evidentiary arguments about students in a common language. This is essential for developing effective diagnostic assessment systems where experts from different disciplines have to work together efficaciously.

Importantly, the ECD framework underscores, but does not overemphasize, the importance of the statistical models that are used in the evidence model component. Statistical models such as DCMs are tools for reasoning about patterns of behavior of students based on data patterns with differential weighting. However, the choice of how the behavioral patterns are modeled and, thus, which real-life elements are represented in a statistical model, is squarely in the hands of the diagnostic assessment developer. The next section now discusses DCMs as a particular class of modern measurement models that can be useful for analyzing data from standardized diagnostic assessments.

4.3 Diagnostic Classification Models

Before beginning our discussion of DCMs, we want to reiterate that many modeling choices are driven by substantive considerations about the structure of desired evidence-based assessment narratives for students. That is, based on the desired level of precision at which a student characteristic is to be measured and interpretations are to be given as well as the real-life constraints imposed by the informational richness of the available data, diagnostic assessment designers have to decide which characteristics should be represented via variables in the DCM that is chosen for analysis. They need to decide which pieces of information are extracted from the complex performance of students and how these pieces of information are coded so that they can be used as input into the statistical models. The choice or construction of any statistical model thus emerges from a careful consideration of students, learning, situations, and theory; it does not or should not determine what interpretations should be or what observations must be limited to.

In this section, we introduce DCMs as a particular class of statistical models that can be useful for standardized diagnostic assessment processes. Specifically, we first discuss key terminology, then describe a unified specification and estimation framework for DCMs, and finally illustrate, using real data from a newly developed diagnostic assessment of elementary school mathematics, how one can estimate DCMs with a commercially available software program.

Table 4.1 Exemplary Q-matrix

Item		Addition	Subtraction	Multiplication	Division
1	2+3−1	1	1	0	0
2	4/2	0	0	0	1
3	5×3−4	0	1	1	0
4	8+12	1	0	0	0

4.3.1 Attributes, Attribute Profiles, and Q-matrices

The term *attribute* generically refers to unobservable (i.e., latent) characteristics of students. In DCMs, we will operationalize these characteristics using unobservable (i.e., latent) variables. We use the values on these latent variables to reason backwards about students' mastery states on the attributes of interests based on students' observed response patterns to diagnostic assessment items. The resulting pattern of attribute mastery states are known as *attribute profiles* in the literature; under an effective diagnostic assessment design, attribute profiles carry reliable information for making meaningful instructional decisions.

Once the targeted attributes and potential attribute profiles are determined based on an appropriate applied cognitive theory, the next step is to specify which attributes are measured by each individual assessment item (i.e., which attributes are required by the students to obtain a maximum score on an item). The relationship between attributes and items is formally captured in a two-dimensional table known as a *Q-matrix* (Tatsuoka 1990). In general, rows of the table correspond to items, columns of the table correspond to attributes, and entries in the table are typically binary (i.e., "0" or "1"), indicating which attributes are measured by which items.

There are a number of ways of constructing Q-matrices. In educational testing, Q-matrices may be constructed based on theories about learning in the domain triangulated by experts' judgment, empirical research, think-aloud protocols, factor analyses of existing tests, and other means of empirical validation (Buck and Tatsuoka 1998; Gierl et al. 2005). To illustrate the structure of a Q-matrix in practice, we use an example scenario where five items measure four attributes in basic arithmetic ability; this matrix is shown in Table 4.1.

According to this Q-matrix in Table 4.1, item 2 and item 4 only measure one attribute, while item 1 and item 3 measure two attributes. Expressed reversely, only mastery of one attribute is required for item 2 and item 4 to get the maximum score on these items, while mastery of two attributes is required for the other two items.

Consequently, the attribute profile (i.e., the mastery state on all attributes measured by the diagnostic assessment) of each student can be represented in the same way using binary indicators where "1" indicates that a student has mastered an attribute, and "0" indicates that he or she has not. For instance, if a student has mastered only the first two attributes among the four attributes above, his or her attribute profile can be represented as [1,1,0,0].

Given the Q-matrix ($\boldsymbol{Q}$) and a student's attribute profile ($\boldsymbol{\alpha}$), an idealized response pattern (i.e., a response pattern that would be observed if the student responded without error) can be predicted through simple matrix algebra as follows:

$$\mathbf{Q} \times \boldsymbol{\alpha} = \begin{bmatrix} 1100 \\ 0001 \\ 0110 \\ 1000 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0/1 \\ 1 \end{bmatrix}.$$

In this example, the student should respond correctly to item 1 and item 4 but not to item 2. It is not clear, however, whether or not this student would respond correctly to item 3 as he or she has only mastered one out of the two required attributes. Different DCMs are designed to operationalize different relationships between the mastery states on individual attributes and the probabilities of a certain response while allowing for imperfect responding due to random errors.

4.3.2 A Definition of DCMs

DCMs are statistical models that were developed to respond to the desire of diagnostic assessment developers to classify students in terms of their mastery states on individual attributes that constitute of their attribute profiles (for overviews see, e.g., DiBello et al. 2007; Rupp and Templin 2008; Rupp et al. 2010; Templin 2004). Formally,

> Diagnostic classification models (DCMs) are probabilistic confirmatory multidimensional latent variable models. Their loading structure / Q-matrix is typically complex to reflect within-item multidimensionality, but may also be simple. DCMs are suitable for modeling observable reponse variables (i.e., dichotomous, polytomous) and contain unobservable latent categorical predictor variables (i.e., dichotomous, polytomous). The predictor variables are combined in compensatory and non-compensatory ways to generate latent classes. DCMs enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. (Rupp et al. 2010, p. 108)

The literature is replete with DCMs that differ in the number of parameters that they contain for items and attributes and the types and numbers of restrictions they place on these parameters; in other words, the flexibility with which they can handle various data structures. Rather than listing all of these models here, we refer to the overview sources cited earlier for detailed descriptions of these models. More importantly, current theory and practice has evolved to the point where many DCMs can now be parameterized as special cases of more general modeling families.

The three most common families in the literature are the *log-linear cognitive diagnosis model* (LCDM) framework by Henson et al. (2009), the *general diagnostic model* (GDM) framework by von Davier (2005, 2010), and the generalized deterministic inputs, noisy "and" gate (G-DINA) model by de la Torre (2009). For the purposes of this chapter, we will use the LCDM framework and refer to the chapter by de la Torre (Chap. 5, this volume) for an overview of the G-DINA model framework.

4.3.3 The LCDM Framework

As the GDM and G-DINA frameworks, the LDCM framework is a unified framework for the specification and estimation of DCMs. Its development was based on finite mixture models (e.g., McLachlan and Peel 2000), log-linear models (e.g., Agresti 2010), and generalized linear and latent mixed models (e.g., Skrondal and Rabe-Hesketh 2004). In the following, we will focus on the simplest case of an LCDM, which concerns binary item scores (i.e., "1" for a correct response and "0" for an incorrect response), binary attribute mastery states (i.e., "1" for a mastered attribute and "0" for a non-mastered attribute), and binary Q-matrix entries (i.e., "1" for an attribute that is measured by an item and "0" otherwise); extensions are relatively easily specified and estimated.

4.3.3.1 Model Specification

In the LCDM, the probability of a correct response as a function of attribute mastery states is defined as

$$P\left(Y_{ij}=1\left|\boldsymbol{\alpha}_i,\mathbf{q}_j\right.\right)=\frac{\exp\left[\lambda_{0j}+\boldsymbol{\lambda}'_j\,h\left(\boldsymbol{\alpha}_i,\mathbf{q}_j\right)\right]}{1+\exp\left[\lambda_{0j}+\boldsymbol{\lambda}'_j\,h\left(\boldsymbol{\alpha}_i,\mathbf{q}_j\right)\right]}, \tag{4.1}$$

where i and j denote student and item, respectively; λ_{0j} is an intercept and $\boldsymbol{\lambda}_j$ represents a vector of coefficient indicating the effects of attribute mastery on the response probability for item j, and $h(\boldsymbol{\alpha}_i,\mathbf{q}_i)$ is a set of linear combinations of the attribute mastery indicators $\boldsymbol{\alpha}_i$ and the Q-matrix entries $\mathbf{q}_j$. Specifically, the kernel of the above expression has the following general form:

$$\lambda_{0j}+\boldsymbol{\lambda}_j h\left(\boldsymbol{\alpha}_i,\mathbf{q}_j\right)=\lambda_{0j}+\sum_{u=1}^{k}\lambda_{ju}\left(\alpha_k q_{ju}\right)+\sum_{u=1}^{k}\sum_{v>u}\lambda_{juv}\left(\alpha_u\alpha_v q_{ju}q_{jv}\right)+\cdots \tag{4.2}$$

which is similar to the structure of factorial analysis of variance (ANOVA) models.

The intercept can be interpreted as a *guessing parameter* because it reflects the probability of providing a correct response for those students who have not mastered any attributes – this is the lowest probability for any attribute profile. The λ_{ju} parameters represent the main effects of each attribute on the response probability for item j, and the λ_{juv} parameters represent the two-way interaction effects of the combination of the mastery states of attributes u and v on the response probability for item j; higher-order parameters are defined likewise with aligned meanings. In other words, the specification of the kernel follows the specification of factorial ANOVA models with intercept, main-effect, and interaction-effect parameters.

Depending on how many attributes are included in the item, the LCDM can include main effects for each attribute, two-way and three-way interactions among

attributes, and so forth. Simulation studies (Kunina-Habenicht, Rupp, and Wilhelm, 2012; Choi et al. 2010) have shown that interaction-effect parameters require very large sample sizes for reliable estimation, however, so that main-effect parameter specifications are probably most appropriate for most practical contexts.

To illustrate the general expression for the LCDM with a concrete example, consider the Q-matrix from Table 4.1. Since item 1 measures attribute 1 and attribute 2, $q_{11}=q_{12}=1$, while $q_{13}=q_{14}=0$. Consequently, the probability of a correct response for item 1 takes the form

$$P\left(Y_{i1}=1\,|\,\boldsymbol{\alpha}_i,\,\mathbf{q}=(1,1,0,0)\right)=\frac{\exp\left(\lambda_{10}+\lambda_{11}\alpha_1+\lambda_{12}\alpha_2+\lambda_{112}\alpha_1\alpha_2\right)}{1+\exp\left(\lambda_{10}+\lambda_{11}\alpha_1+\lambda_{12}\alpha_2+\lambda_{112}\alpha_1\alpha_2\right)}, \quad (4.3)$$

with the exact probability values for each attribute profile (i.e., each combination of attribute mastery states for attribute 1 and attribute 2) depending on the values of the item parameters λ_{10}, λ_{11}, λ_{12}, and λ_{112}, which need to be estimated in practice from the student response data.

4.3.3.2 Illustrative Special Cases

As the response probability for this item is influenced by the mastery states on two attributes, we can ask several questions: What is the response probability for students who have mastered only one attribute out of two? Does mastering attribute 1 have a bigger impact on the response probability than mastering attribute 2? Is there an additional effect on the response probability for mastering both attributes once one of them has already been mastered?

These questions can be answered empirically either by specifying the most general DCM in Eq. 4.2 and inspecting the values of the resulting parameter estimates a posteriori or by specifying specific DCMs that reflect different hypotheses in alignment with these three questions a priori. To illustrate the flexibility of the LCDM framework, we discuss particular DCMs that would result from such a priori specifications in the following.

For the first scenario, if the DCM is supposed to reflect the assumption that both attributes need to be mastered to provide a correct response, then Eq. 4.3 can be modified as follows:

$$P\left(Y_{i1}=1\,|\,\boldsymbol{\alpha}_i,\mathbf{q}=(1,1,0,0)\right)=\frac{\exp\left(\lambda_{10}+(0)\alpha_1+(0)\alpha_2+\lambda_{112}\alpha_1\alpha_2\right)}{1+\exp\left(\lambda_{10}+(0)\alpha_1+(0)\alpha_2+\lambda_{112}\alpha_1\alpha_2\right)}, \quad (4.4)$$

Here, the main effects for attribute 1 and attribute 2 are set to 0, and only the intercept and interaction effect take on non-zero values. Thus, the response probabilities for this item are identical for students who have not mastered any of the two or only one of the two measured attributes. This model is referred to as the *deterministic input*, *noisy "and" gate* (DINA) model in the literature and substantively reflects a situation where the mastery of a subset of attributes cannot

compensate for the lack of mastery of any other attribute(s) that is not mastered by a student but measured by an item (e.g., Junker and Sijtsma 2001; de la Torre 2009). In substantive terms for our simple example, this model reflects the assumption that students are not likely to solve item 1 if they have not mastered both addition and subtraction.

For the second scenario, consider the case where an item can be solved when only one of several attributes has been mastered. For example, suppose that students are asked to determine the interior angle of a regular pentagon. Some students may draw a picture to determine how many triangles there are in a pentagon. Once they figure out that there are three triangles inside the pentagon, the answer becomes $180 * 3 = 540$ because the interior angle of a triangle is 180. Others may solve the same question using the analytic knowledge that for any regular polygon, the sum of the interior angles $= 180(n-2)$ where n is the number of sides. Since a pentagon has five sides, $180(5-2) = 540$. If both strategies were coded as attributes that this item measured, then mastering both attributes does not increase the probability of a correct response.

For this situation, Eq. 4.3 can be modified as follows:

$$P\left(Y_{i1}=1 \mid \boldsymbol{\alpha}_i, \mathbf{q}=(1,1,0,0)\right)=\frac{\exp\left(\lambda_{10}+\lambda_1\alpha_1+\lambda_1\alpha_2+(-\lambda_1)\alpha_1\alpha_2\right)}{1+\exp\left(\lambda_{10}+\lambda_1\alpha_1+\lambda_1\alpha_2+(-\lambda_1)\alpha_1\alpha_2\right)}, \quad (4.5)$$

where the probability of getting a correct answer for those who possess the knowledge about triangles, those who have mastered analytic knowledge, or those who know both is the exactly same. This model is referred to as the *deterministic input*, *noisy* "or" *gate* (DINO) model in the literature and reflects the assumption that mastery of subset of attribute can compensate for the lack of mastery of other attribute(s) (e.g., Templin and Henson 2006).

For the third scenario, consider the case where the probability of getting a correct response to an item increases as the number of mastered attributes increases. For example, suppose that a reading comprehension item with a passage regarding physics is presented to students. The impact of understanding the meaning of a certain vocabulary in the text and knowledge of syntactic structure may be additive on the probability of students' correct answer.

In this case, Eq. 4.3 can be modified as follows:

$$P\left(Y_{i1}=1 \mid \boldsymbol{\alpha}_i, \mathbf{q}_j\right)=\frac{\exp\left(\lambda_{10}+\lambda_{11}\alpha_1+\lambda_{12}\alpha_2+(0)\alpha_1\alpha_2\right)}{1+\exp\left(\lambda_{10}+\lambda_{11}\alpha_1+\lambda_{12}\alpha_2+(0)\alpha_1\alpha_2\right)}, \quad (4.6)$$

where the interaction effect sets to zero, indicating no additional effect of mastering both attributes. This model is referred to in the literature as the compensatory reparameterized unified model (C-RUM) (e.g., Hartz 2002; Roussos et al. 2007) and also reflects the assumption that mastery of a particular attribute can compensate for the lack of mastery of any other attribute, albeit not as strongly as in the DINO model for scenario two above.

4.3.4 Estimating DCMs via the LCDM Framework

To date, there exist no specific software programs that are designed to specify and estimate DCMs within a user-friendly GUI environment. In the past, researchers have typically written their own estimation codes. For instance, the commercially available *Arpeggio* program (www.assess.com) was originally developed specifically for the RUM/Fusion model and requires sophisticated knowledge of Bayesian estimation for reliable use, the code for the G-DINA model was written in the programming language Ox (http://www.doornik.com/) and is still under development, and the program MDLTM for the GDM (von Davier 2006) originally relied on a syntax interface and is available as a research license only.

However, since DCMs are special cases of restricted latent class models, they can be estimated within any commercial program for latent class models that allows for the imposition of parameter constraints if a unified framework like the LCDM is used. For example, Choi et al. (2010), Templin et al. (2011), Kunina-Habenicht et al. (2010), and Rupp et al. (2010) have demonstrated how DCMs can be specified and estimated in M*plus*. In the following section, we present an additional example based on the data from Kunina et al. (2010).

4.4 Illustrative Extended Example

4.4.1 Data Description and Q-matrix

The *diagnostic mathematics assessment* (DMA) that is the focus of this example was developed to provide information on basic arithmetic ability for students in the 3rd and 4th grades in Germany (Kunina-Habenicht et al. 2009; 2010). Test items were constructed to measure several basic arithmetic skills such as addition, subtraction, multiplication, division, executing inverse operation, executing carry over, solving word problems, and converting measurement units. The original item pool consisted of 70 items and was administered to a sample of 2,032 4th grade students in different schools in Germany in 2008 using a complex booklet design (Frey et al. 2009). For illustration purposes, we analyzed only a subset of 20 items, which reflected the structure of the Q-matrix of the original item pool.

Even though several fine-grained skills were originally defined and used in the item development process, Kunina-Habenicht et al. (2010) found that a Q-matrix with four attributes was most strongly supported when various FA models and DCMs were used for data analysis. The four resulting attributes were addition/subtraction (A/S), multiplication/division (M/D), modeling (model), and converting units (units); Table 4.2 shows the Q-matrix for our example using the same attribute definitions. As shown in Table 4.2, items 1–10 measure one attribute, while items 11–20 measure two attributes.

Table 4.2 Q-matrix of diagnostic mathematics assessment (DMA)

Item	A/S	M/D	Model	Units
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	0	1	0	0
6	0	1	0	0
7	0	1	0	0
8	0	1	0	0
9	0	0	1	0
10	0	0	1	0
11	0	1	1	0
12	0	1	1	0
13	0	1	1	0
14	0	1	1	0
15	1	0	0	1
16	1	0	0	1
17	1	0	0	1
18	1	0	0	1
19	1	0	0	1
20	1	0	0	1

4.4.2 Model Selection and Item Parameter Estimation

For illustration purposes, we fit the four different DCMs to this data set that we discussed in the previous section, namely, the full LCDM, the DINA, the DINO, and the C-RUM. Recall that, for items that measure two attributes, the full LCDM model includes both main-effect parameters and the interaction-effect parameter; the DINA model contains only the two-way interaction-effect parameter; the DINO model contains both main-effect parameters and a negative two-way interaction-effect parameter, all constrained to equality; and the C-RUM contains only main-effect parameters. Thus, the full LCDM is the most flexible model, while the DINA model is the most restrictive model with the remaining two models representing special intermediate cases. All models were estimated in Mplus 6.0 (Muthén and Muthén 1998–2010).

After fitting the four competing models, relative model fit indices were used to determine the best-fitting model. We used *Akaike's information criterion* (AIC) (Akaike 1974) and Schwarz's (1978) *Bayesian information criterion* (BIC) that were provided in the output files. As is typical in practical applications, AIC and BIC did not always agree about the best-fitting model because they penalize differentially strong for the parametric complexity of the fitted models and sample size. As shown in Table 4.3, the AIC suggested that the C-RUM was the best-fitting model, while the BIC suggested that the DINA was the best-fitting model; according to the AIC, the full model is a close competitor to the C-RUM.

Table 4.3 Results of fit indices for model selection

	DINA	DINO	C-RUM	FULL
AIC	19352.16	19359.94	**19314.87**	19316.85
BIC	**19649.06**	19656.84	19665.75	19721.71
Number of parameters	55	55	65	75

Boldfaced entries indicate model with the smallest information criterion value

Table 4.4 Item parameter estimate from two models

	DINA						C-RUM				
		Main effect			Interaction effect			Main effect			
Item	Intercept	A/S	M/D	Model	(M/D)× (Model)	(A/S)× (Units)	Intercept	A/S	M/D	Model	Units
1	−0.74	2.28					−0.60	2.33			
2	−1.09	2.40					−0.99	2.53			
3	0.09	1.87					0.17	1.98			
4	0.42	1.98					0.51	2.07			
5	−2.08		2.67				−1.95		2.69		
6	0.37		1.92				0.42		2.04		
7	−0.26		2.85				−0.19		3.18		
8	−0.94		3.00				−0.81		3.14		
9	−1.85			2.04			−1.73			1.97	
10	−1.46			2.01			−1.40			2.05	
11	**−1.69**				**1.91**		**−1.90**		**0.43**	**1.80**	
12	−2.18				2.72		−2.46		0.98	2.16	
13	−1.18				1.79		−1.39		0.34	1.81	
14	−2.76				2.63		−3.00		1.74	1.25	
15	−0.67					1.92	−1.05	1.19			1.70
16	−1.92					2.73	−2.62	1.81			2.52
17	−0.70					2.09	−1.11	1.40			1.65
18	−0.32					1.82	−0.62	0.85			1.94
19	1.16					1.12	0.98	0.99			0.46
20	−2.17					2.17	−2.63	1.84			1.15

The item used for illustrative computations is shown in boldface

Which models one chooses does not matter for items 1–10 because those items measure only one attribute, but it matters for items 11–20 because they measure two attributes. To see the impact of choosing either the DINA or the C-RUM models for those items, we present the estimated model parameters for all 20 items in Table 4.4.

Since parameter estimates are on the logit scale and it is easier to think in terms of response probabilities, it is insightful to look at the difference in response probabilities for students with different attribute profiles under different models. Due to space limitations, we present here the corresponding response probabilities for item 11 in Table 4.5. As only the two attributes M/D and model were required for this item, only the mastery states for these two attributes influence the resulting response probabilities.

Table 4.5 Probability of a correct answer for item 11

Attribute				Model	
A/S	**M/D**	**Model**	Units	DINA	C-RUM
0	**0**	**0**	0	0.16	0.13
0	**0**	**0**	1	0.16	0.13
0	**0**	**1**	0	0.16	0.48
0	**0**	**1**	1	0.16	0.48
0	**1**	**0**	0	0.16	0.19
0	**1**	**0**	1	0.16	0.19
0	1	1	0	0.55	0.58
0	1	1	1	0.55	0.58
1	0	0	0	0.16	0.13
1	0	0	1	0.16	0.13
1	0	1	0	0.16	0.48
1	0	1	1	0.16	0.48
1	1	0	0	0.16	0.19
1	1	0	1	0.16	0.19
1	1	1	0	0.55	0.58
1	1	1	1	0.55	0.58

Latent classes with identical probabilities are shown in identical shades of grey

As can be shown in Fig. 4.2, these response probabilities were computed as follows. The response probability for students with different attribute profiles under the DINA model for item 11 is

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.69\right)}{1+\exp\left(-1.69\right)}=0.16,$$

for those who have not mastered any or only one of the two measured attributes, while the response probability for those who have mastered both measured attributes is

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.69+1.91\right)}{1+\exp\left(-1.69+1.91\right)}=0.55.$$

The response probability for students with different attribute profiles under the C-RUM model for item 11 is

$$P(Y_{11}=1)=\frac{\exp(-1.90)}{1+\exp(-1.90)}=0.13.,$$

for those who have not mastered either measured attribute,

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.90+0.43\right)}{1+\exp\left(-1.90+0.43\right)}=0.19,$$

for those who have mastered only one M/D,

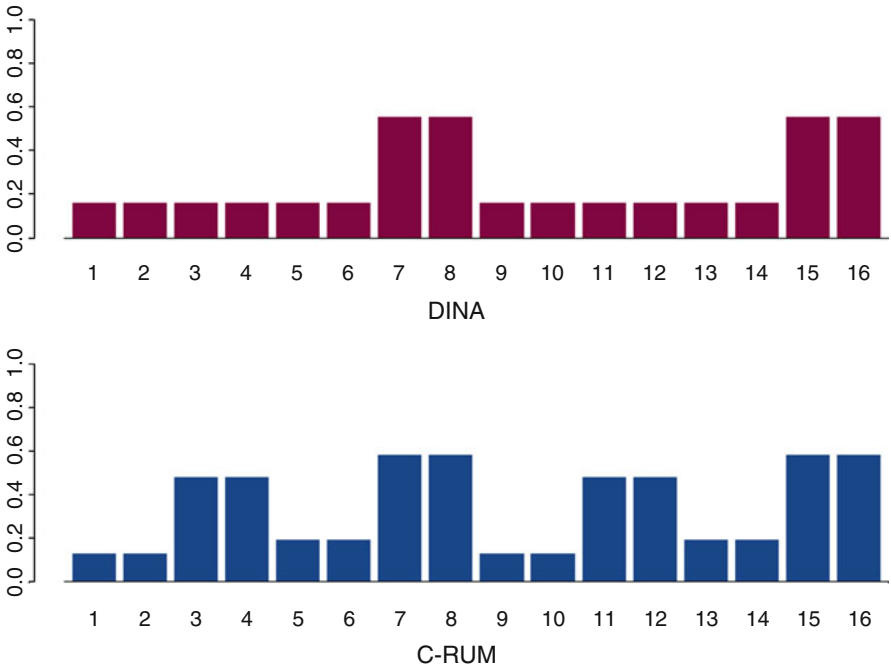

Fig. 4.2 Probability of a correct answer from each attribute profile for item 11

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.90+1.8\right)}{1+\exp\left(-1.90+1.8\right)}=0.48,$$

for those who have mastered only model and for those who have mastered both measured attributes,

$$P\left(Y_{11}=1\right)=\frac{\exp\left(-1.90+0.43+1.8\right)}{1+\exp\left(-1.90+0.43+1.8\right)}=0.58.$$

These probability computations illustrate nicely how the C-RUM allows for a finer differentiation between students with different attribute profiles than the DINA model in terms of their resulting response probabilities. It is also worth noting that, for both models, the response probabilities for students who have not mastered any attributes are non-zero because the estimates of the intercept parameters are non-zero.

4.4.3 Reporting Attribute Profiles for Groups of Students

The primary purpose of DCMs is to classify students into one of a number of prespecified attribute profiles that correspond to sequences of mastery states on the attributes measured by the diagnostic assessment. Table 4.6 and Fig. 4.3 illustrate

Table 4.6 Distribution of attribute profiles

A/S	M/D	Model	Units	Proportion (%)
0	0	0	0	30.4
0	0	0	1	6.2
0	0	1	1	3.4
0	1	1	1	2.7
1	0	0	0	5.7
1	0	1	0	4.4
1	1	0	0	11.2
1	1	1	0	6.8
1	1	1	1	27.7

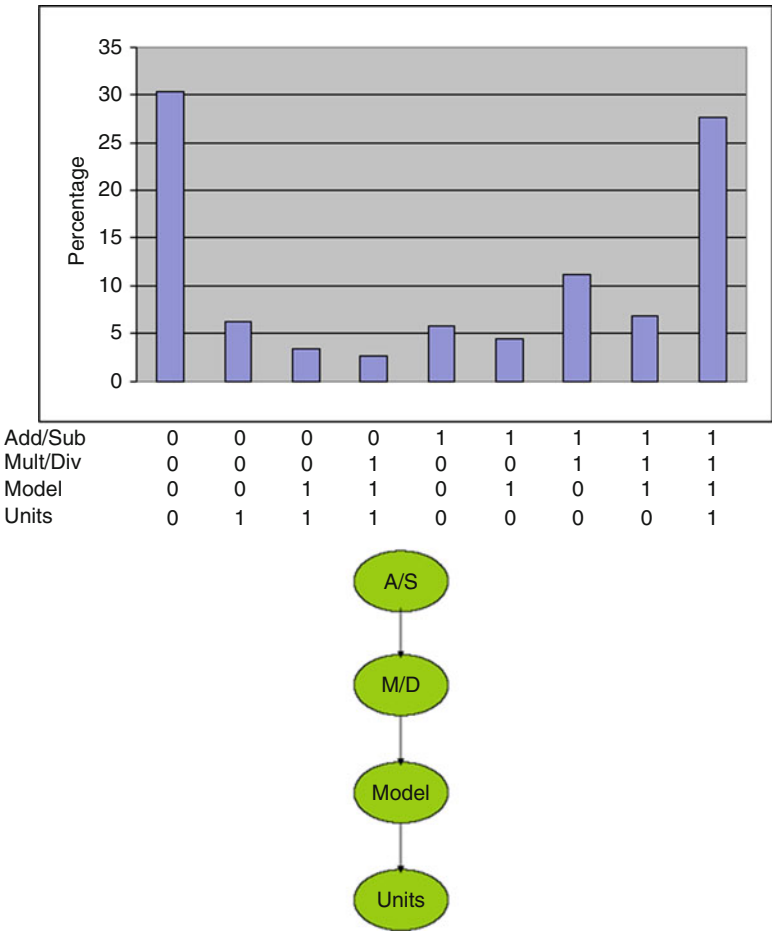

Fig. 4.3 Attribute profiles in sample (*left*) and inferred relationship among attributes (*right*)

how one could display the distribution of attribute profiles for the DMA in our example. Note that with four attributes that are defined in terms of mastery and non-mastery, there exist a total of 16 possible attribute profiles; however, empirically, only nine attribute profiles were populated for these data. Figure 4.2 clearly shows that students predominantly belonged to the two attribute profiles that reflected the lack of mastery of all attributes (30%) and the mastery of all attributes (28%). Moreover, 11% of students were classified as having mastered the first two attributes (A/S and M/D), and about 7% of students were classified as having mastered the first three attributes.

These results gently suggest what is known in the literature as a *linear attribute hierarchy* where the basic arithmetic skills (addition, subtraction, multiplication, division) seem to be mastered before the modeling and unit knowledge skills. However, it needs to be remembered that such inferences are tentative at best because (a) the current data are cross-sectional and not longitudinal in nature, making developmental claims inappropriate, (b) several attribute patterns have similarly low membership probabilities associated with them, and (c) no additional validation results are presented here.

The item parameters and distribution of attribute profiles can be interesting for those who are in charge of test development and require summative statements of students' proficiencies in the assessment of learning sense, while reporting about each student's attribute profile may be more useful for teachers, students, and parents to support assessment for learning.

4.4.4 Reporting Attribute Profiles for Individual Students

To illustrate how report cards for individual students could be constructed, we show here the attribute profiles for selected students in Table 4.7. First, for each student, each column indicates the probability that a student should be classified as having each of the nine empirically observed attribute profiles, while the last four columns show the probabilities that each student possesses each of the four attributes that are measured by the test separately. For example, the first student is classified as having mastered attributes A/S and M/D but neither model nor units. This can be seen in the high probabilities of mastery for the first two attributes, which are 0.92 and 0.85, respectively, and the low probabilities of mastery for the last two attributes, which are.15 and.01, respectively. It can also be seen in the fact that his or her probability for the attribute profile [1,1,0,0] is considerably higher at.72 than the probability for any of the other eight attribute profiles.

At the same time, note how there can be challenges in reliably classifying individual students. The second student has a probability of mastery of .55 for the first attribute but is nevertheless classified as having mastered none of the attributes in the profile with a probability of .33. This probability is rather low, however, compared to the highest probability for the first, third, and fourth students and is relatively close to the probability for the attribute profile where only the first attribute

Table 4.7 Sample probabilities for attribute profiles and individual attributes

	Attribute profiles									Attributes			
ID	[0,0,0,0]	[0,0,0,1]	[0,0,1,1]	[0,1,1,1]	[1,0,0,0]	[1,0,1,0]	[1,1,0,0]	[1,1,1,0]	[1,1,1,1]	A/S	M/D	Model	Units
1	.07	.00	.00	.00	.06	.02	**.72**	.12	.00	.92	.85	.15	.01
2	**.33**	.08	.04	.01	.20	.13	.12	.05	.05	.55	.22	.28	.18
3	.00	.00	.01	.00	.02	.04	.05	.07	**.78**	.98	.91	.93	.82
4	.02	.00	.00	.00	.06	.02	**.75**	.13	.02	.98	.90	.17	.02

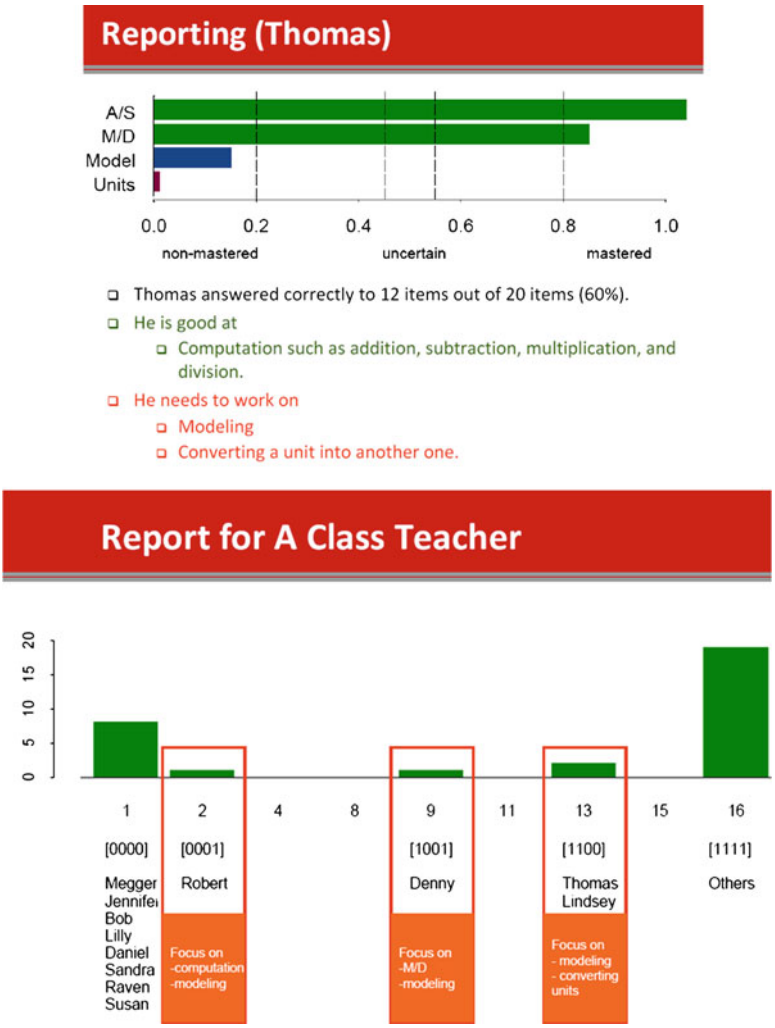

Fig. 4.4 Exemplary report card for each student (*top*) and for each class (*bottom*)

is mastered. In practice, it would not be advisable to use this student's classification for high-stakes decision-making, but it may still be useful to suggest to the student additional practice on all attributes with a particular emphasis on the last three.

Based on the classification probabilities shown in Table 4.7, one can create *diagnostic report cards* for each individual student; Fig. 4.4 shows a sample report card for the first student and a class, respectively.

This card shows a total score that expresses how well a student, fictitiously named Thomas, did on the assessment overall and also his mastery states for each attribute that can inform him of his strengths and weaknesses in particular areas if he is taught how to read this information well.

4.5 Conclusions

Developments in the areas of diagnostic assessment design, from a procedural perspective, and DCMs, from a statistical perspective, have the potential to lead to well-aligned large-scale diagnostic assessment systems that can yield more fine-tuned and more instructionally relevant information about students' strengths and weaknesses. In particular, this can be useful as assessment for learning as well as assessment as learning. Nevertheless, it is important to note a variety of caveats.

Substantively, what is crucially needed is a focus on long-term investigations of student progress similar to innovative work in performance-based science assessment (e.g., Thadani et al. 2009). Since education is an ongoing process in class and monitoring students' growth is one of the primary tasks of teachers, diagnostic assessment needs to be carried out with a longitudinal perspective of an assessment-intervention cycle.

Statistically, because of the complexity of the desired diagnostic inferences and the resulting parametric complexity of DCMs, the design requirements for diagnostic assessments are high. On the one hand, it is crucial that every effort be put into place to ensure that calibrations of resulting response data yield reliable profiles on multiple attributes (i.e., separable statistical dimensions). This requires longer assessments in general because sufficient information is needed for each attribute to achieve a reliable statistical classification with DCMs. However, the amount of required statistical information is somewhat smaller than when traditional models from multidimensional IRT or FA are used due to the discrete nature of classifications. On the other hand, this requires data from hundreds or thousands of students per assessment item because item parameters need to be estimated reliably in preoperational settings. Once diagnostic assessments have been calibrated with DCMs, however, it is much easier to score future generations of students with these assessments.

In the end, DCMs are just statistical tools that serve a larger purpose of creating a defensible evidence-based assessment narrative about students. Since the specification of DCMs is still relatively tedious, a wider implementation of these models will probably also not take place unless more user-friendly software is made available. We also want to underscore that they are also not the only models that can be used for diagnostic assessment purposes as the special issue of the *Journal of Educational Measurement* in 2007 demonstrated. For example, multidimensional models from IRT (e.g., Reckase 2009) or FA (e.g., McDonald 2009), as well as cluster analysis methods (e.g., Gan et al. 2007; Steinley 2006), may provide reasonable alternatives even though they result in multiple continuous scales rather than discrete attribute profiles. IRT and FA models in particular have been in use much longer than DCMs and are, thus, generally more strongly trusted by interdisciplinary specialists. Cluster analysis models have a similarly long history in the social and behavioral sciences and are computationally more efficient than DCMs. Thus, they represent attractive modeling alternatives for day-to-day implementations of diagnostic assessments (see Nugent et al. 2009, 2010).

Acknowledgements We would like to thank Olga Kunina-Habenicht for giving us access to the data that was used for the example in the last section in this chapter. The work of Dr. Kunina-Habenicht, including the design, implementation, and analysis, was funded, in part, by grant number RU-424/3–1 from the *German Research Foundation* (*Deutsche Forschungsgemeinschaft, DFG*) in the *Priority Research Program* entitled "Models of Competencies."

References

Agresti, A. (2010). *Categorical data analysis* (2nd ed.). New York: Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*(2), 119–157.

Choi, H-. J., Templin, J. L., Cohen, A. S., & Atwood, C. H., (2010, April). *The impact of model misspecification on estimation accuracy in diagnostic classification models (DCMs)*. Paper presented at the annual meeting of the National Council for Measurement and Education, Denver, CO.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Pacific Grove: Wadsworth.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

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

DiBello, L., Roussos, L., & 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, pp. 979–1030). Amsterdam: Elsevier.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 380–396.

Frey, A., Hartig, J., & Rupp, 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*, 39–53.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications*. Alexandria: American Statistical Association.

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT* (Research Rep. No. 2005–2011). New York: College Examination Board.

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 at 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*(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.

Kunina-Habenicht, O., Rupp, A., & Wilhelm, O. (2010, May). *Modelling the latent structure of a diagnostic mathematics assessment within a general log-linear modelling framework*. Presented at the annual meeting of the National Council for Measurement in Education (NCME), Denver, Colorado.

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.

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*, 64–70.

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

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. New York: Cambridge University Press.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4*, 333–369.

Linn, R. L. (1986). Testing and assessment in education: Policy issues. *American Psychologist, 41*, 1153–1160.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah: Erlbaum.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Rep. 9). Menlo Park: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Technical Rep. 632). Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mok, M. M. C. (2010). *Self-directed learning oriented assessment: Assessment that informs learning & empowers the learner*. Hong Kong: Pace Publications Ltd.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus* (Version 6) [Computer software]. Los Angeles: Muthén & Muthén.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Nugent, R., Ayers, E., & Dean, N. (2009). Conditional subspace clustering of skill mastery: Identifying skills that separate students. In *Proceedings from the 2nd international conference on educational data mining* (pp. 101–110). Retrieved July 19, 2010, from www.educational-datamining.org/EDM2009/

Nugent, R., Dean, N., & Ayers, E. (2010). Skill set profile clustering: The empty K-means algorithm with automatic specification of starting cluster centers. In *Proceedings from the 3rd international conference on educational data mining* (pp. 151–160). Retrieved July 19, 2010, from http://educationaldatamining.org/EDM2010/

O'Reilly, T. P., Sheehan, K. M., & Bauer, M. I. (2008, March). *Cognitively based assessments of, for, and as learning: Bridging the gap between research and practice*. Presented at the annual meeting of the American Educational Research Association, New York.

Reckase, M. (2009). *Multidimensional item response theory*. New York: 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. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275–318). Cambridge: Cambridge University Press.

Rupp, 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.

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

Rupp, A. A., Levy, R., DiCerbo, K., Sweet, S., et al. (in press). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. Journal of Educational Data Mining.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461–464.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Steinley, D. (2006). *K*-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology, 59*, 1–34.

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: Erlbaum.

Templin, J. L. (2004). *Generalized linear mixed proficiency models*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychology Methods, 11*(3), 287–305.

Templin, J. L., Henson, R. A., & Douglas, J. (2011). *General theory and estimation of cognitive diagnosis models: Using Mplus to derive model estimates*.

Thadani, V., Stevens, R. H., & Tao, A. (2009). Measuring complex features of science instruction: Developing tools to investigate the link between teaching and learning. *The Journal of the Learning Sciences, 18*, 285–322.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge: Cambridge University Press.

von Davier, M. (2005). *A general diagnostic model applied to language testing data.* (ETS Research Rep. No. RR-05–16). Princeton: Educational Testing Service.

von Davier, M. (2006). Multidimensional latent trait modelling (MDLTM) [Software program]. Princeton: Educational Testing Service.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., Crawford, A., Choi, Y., & Behrens, J. (2009, June). *A Bayes net approach to modeling learning progressions and task performances*. Paper presented at the Learning Progressions in Science (LeaPS) conference, Iowa City, IA.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport: American Council on Education.