JOINT MAXIMUM LIKELIHOOD ESTIMATION FOR DIAGNOSTIC CLASSIFICATION MODELS

CHIA-YI CHIU

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY

HANS-FRIEDRICH KÖHN

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

YI ZHENG

ARIZONA STATE UNIVERSITY

ROBERT HENSON

UNIVERSITY OF NORTH CAROLINA

Joint maximum likelihood estimation (JMLE) is developed for diagnostic classification models (DCMs). JMLE has been barely used in Psychometrics because JMLE parameter estimators typically lack statistical consistency. The JMLE procedure presented here resolves the consistency issue by incorporating an external, statistically consistent estimator of examinees' proficiency class membership into the joint likelihood function, which subsequently allows for the construction of item parameter estimators that also have the consistency property. Consistency of the JMLE parameter estimators is established within the framework of general DCMs: The JMLE parameter estimators are derived for the Loglinear Cognitive Diagnosis Model (LCDM). Two consistency theorems are proven for the LCDM. Using the framework of general DCMs makes the results and proofs also applicable to DCMs that can be expressed as submodels of the LCDM. Simulation studies are reported for evaluating the performance of JMLE when used with tests of varying length and different numbers of attributes. As a practical application, JMLE is also used with "real world" educational data collected with a language proficiency test.

Key words: cognitive diagnosis, joint maximum likelihood estimation, nonparametric classification, statistical consistency.

1. Introduction

Diagnostic classification models in educational assessment (DCM—often also referred to as "cognitive diagnosis models"; DiBello, Roussos, & Stout, 2007; Haberman & von Davier, 2007; Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010) describe an examinee's ability as a composite of specific discrete (cognitive) skills called attributes, each of which an examinee may or may not have mastered. Distinct profiles of attributes define classes of proficiency. Modeling educational testing data within a DCM-framework seeks to estimate the item parameters and to assign examinees to proficiency classes (i.e., estimate their individual attribute profiles). Current methods for fitting DCMs either use marginal maximum likelihood estimation relying on the Expectation Maximization algorithm (MMLE-EM) or Markov chain Monte Carlo (MCMC) techniques.

Correspondence should be made to Chia-Yi Chiu, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA. Email: chia-yi.chiu@gse.rutgers.edu

These methods work well for simple DCMs (e.g., the implementations of MMLE-EM for the Deterministic Input, Noisy "And" gate [DINA] model; Junker & Sijtsma, 2001; Macready & Dayton, 1977; and the Deterministic Input, Noisy "Or" gate [DINO] model; Templin & Henson, 2006; in the R package CDM; Robitzsch, Kiefer, George, & Uenlue, 2015). However, for more complex DCMs (e.g., the Reduced Reparameterized Unified Model [Reduced RUM]; Hartz, 2002; Hartz & Roussos, 2008) and general DCMs like von Davier's (2005, 2008) General Diagnostic Model (GDM) or the Log-Linear Cognitive Diagnosis Model (LCDM; Henson, Templin, & Willse, 2009; Rupp et al., 2010; Templin & Bradshaw, 2014; Templin & Hoffman, 2013), MMLE-EM and MCMC can be computationally expensive, which may limit their usefulness in research and practice.

This article explores the potential of joint maximum likelihood estimation (JMLE) for fitting DCMs. JMLE has been mostly avoided in psychometrics—despite the mathematical convenience of simple likelihood functions—because the JMLE parameter estimators typically lack statistical consistency (Baker & Kim, 2004; Haberman, 2004; Neyman & Scott, 1948). The JMLE procedure presented here resolves the consistency issue by incorporating an external, statistically consistent estimator of examinees' proficiency class membership into the joint likelihood function, which subsequently allows for the construction of item parameter estimators that also have the consistency property.

The presentation is preceded by a brief review of DCMs. The results and proofs concerning the consistency of the item parameter estimators are derived for the LCDM; using the theoretical framework of general DCMs makes the results and proofs also applicable to DCMs that can be expressed as submodels of the LCDM. The theoretical part is augmented by simulation studies that compare the performance of JMLE with that of MMLE-EM under finite sample conditions using artificial data conforming to the LCDM. In addition, the results of a real-world application of JMLE to the analysis of language assessment data are reported. The paper concludes with a discussion of the findings and directions for future research.

2. Background: Diagnostic Classification Models

DCMs model the functional relation between attribute mastery and the probability of a correct item response. Suppose that K latent binary attributes constitute a certain ability domain; there are then 2^K distinct attribute profiles composed of these K attributes representing $2^K = M$ distinct proficiency classes. Let the K-dimensional vector, $\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \ldots, \alpha_{mK})'$, denote the binary attribute profile of proficiency class $\mathcal{C}_m$, $m = 1, 2, \ldots, M$, where the kth entry indicates whether the respective attribute has been mastered. Y_{ij} is the observed response of examinee i, $i = 1, 2, \ldots, N$, to binary item j, $j = 1, 2, \ldots, J$. The attribute profile of examinee $i \in \mathcal{C}_m$, $\boldsymbol{\alpha}_{i \in \mathcal{C}_m}$, is written as $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$.

Consider a test consisting of J items. Each individual item j is associated with a K-dimensional binary vector $\mathbf{q}_j$ called the item-attribute profile, where $q_{jk} = 1$, $k = 1, 2, \ldots, K$, 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 profiles. The J item-attribute profiles of a test constitute its Q-matrix, $\mathbf{Q} = \{q_{jk}\}_{(J \times K)}$, (Tatsuoka, 1985) that summarizes the constraints specifying the associations between items and attributes.

The distinct parameterization of specific DCMs reflects differences in the underlying theories on how (non-)mastery of attributes affects an examinee's test performance. General DCMs allow for expressing these distinct functional relations in a unified mathematical form and parameterization. The archetypal general DCM is von Davier's (2005, 2008) General Diagnostic Model (GDM). Von Davier defined $h(\mathbf{q}_j, \boldsymbol{\alpha}_i)$ as a general function of the attribute profile of item j and the attribute profile of examinee i to allow for the flexible modeling of examinees' responses to

item j. The item response function (IRF) of presumably the most popular version of von Davier's GDM is formed by the logistic function of the linear combination of all K attribute main effects

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp(\beta_{j0} + \boldsymbol{\beta}'_j h(\mathbf{q}_j, \boldsymbol{\alpha}_i))}{1 + \exp(\beta_{j0} + \boldsymbol{\beta}'_j h(\mathbf{q}_j, \boldsymbol{\alpha}_i))} = \frac{\exp(\beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_{ik})}{1 + \exp(\beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_{ik})},$$

where q_{jk} indicates whether mastery of attribute α_{ik} is required for item j (see Equations 1 and 2; von Davier, 2005). Henson et al. (2009) specified v_j as the linear combination of the K attribute main effects, α_k, and all their two-way, three-way, ..., K-way interactions

$$v_j = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk} \alpha_{ik} + \sum_{k'=k+1}^{K} \sum_{k=1}^{K-1} \beta_{j(kk')} q_{jk} q_{jk'} \alpha_{ik} \alpha_{ik'} + \cdots + \beta_{j12\ldots K} \prod_{k=1}^{K} q_{jk} \alpha_{ik}$$

and defined the IRF of a general DCM termed the Loglinear Cognitive Diagnosis Model (LCDM) as

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp(v_j)}{1 + \exp(v_j)} \tag{1}$$

(see Equation 11 in Henson et al., 2009). By imposing appropriate constraints on the β-coefficients in v_j, the IRFs of specific DCMs can be expressed as submodels of the LCDM. In addition to the logit link, de la Torre (2011) proposed the identity link, $P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = v_j$, and the log link, $P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \exp\{v_j\}$, for constructing the IRF of a general DCM called the Generalized DINA (G-DINA) model (see Equations 1–3 in de la Torre, 2011). (The identity and the log link require additional constraints on the coefficients to guarantee $0 \leq P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) \leq 1$.)

3. Joint Maximum Likelihood Estimation for Diagnostic Classification

Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N)'$ denote the $N \times J$ matrix of observed item responses, where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$ is the vector of observed item responses of examinee i. Conditional independence, given attribute profile $\boldsymbol{\alpha}$, is assumed for the observed item responses. Thus, the joint likelihood is

$$L(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N, \boldsymbol{\Theta}; \mathbf{Y}) = \prod_{i=1}^{N} L_i(\boldsymbol{\alpha}_i, \boldsymbol{\Theta}; \mathbf{y}_i) = \prod_{i=1}^{N} \prod_{j=1}^{J} f(y_{ij} | \boldsymbol{\theta}_j, \boldsymbol{\alpha}_i), \tag{2}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_J)$ denotes the matrix of item parameters.

3.1. If Examinees' Attribute Profiles are Known

Suppose examinees' true attribute profiles $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N$ are known. Then, the joint likelihood in Equation 2 reduces to a function of only a single set of unknowns, the item parameters: $L(\boldsymbol{\Theta}; \mathbf{Y}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N)$ (Baker & Kim, 2004; Birnbaum, 1968; Embretson & Reise, 2000).

The estimators of the elements of the item parameter vector $\boldsymbol{\theta}_j$ are derived by maximizing the logarithm of the item likelihood

$$\ln L_j(\boldsymbol{\theta}_j; \mathbf{y}_j, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N) = \sum_{i=1}^{N} \ln\left(f(y_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\alpha}_i)\right). \tag{3}$$

The resulting estimators of the item parameters are denoted by $\hat{\boldsymbol{\theta}}_j$.

3.2. If Examinees' Attribute Profiles are Unknown

However, examinees' true attribute profiles $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N$ are never known and can only be estimated from the observed item responses. Suppose that an estimator $\tilde{\boldsymbol{\alpha}}$ of examinees' true attribute profiles is available that is statistically consistent and does not depend on the JMLE procedure. The estimates of examinees' proficiency class membership obtained from the external estimator $\tilde{\boldsymbol{\alpha}}$ can be used in Equation 3, which then becomes

$$\ln L_j(\boldsymbol{\theta}_j; \mathbf{y}_j, \tilde{\boldsymbol{\alpha}}_1, \tilde{\boldsymbol{\alpha}}_2, \ldots, \tilde{\boldsymbol{\alpha}}_N) = \sum_{i=1}^{N} \ln\left(f(y_{ij}|\boldsymbol{\theta}_j, \tilde{\boldsymbol{\alpha}}_i)\right). \tag{4}$$

Maximizing Equation 4 results in the estimators $\tilde{\boldsymbol{\theta}}_j$ of the item parameter estimators. Note that the "tilde-notation" is used to emphasize that these estimators are linked to the consistent estimator $\tilde{\boldsymbol{\alpha}}$, as opposed to the estimators $\hat{\boldsymbol{\theta}}_j$ that require examinees' attribute profiles $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N$ to be known. The item parameter estimators $\tilde{\boldsymbol{\theta}}_j$ for the LCDM are derived in the next section; their consistency is proven in the subsequent section.

4. Joint Maximum Likelihood Estimation for the LCDM

Recall the item response function of the LCDM (see Equation 1)

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i) = \frac{\exp(v_j)}{1 + \exp(v_j)},$$

where

$$v_j = \beta_{j0} + \sum_{k=1}^{K} \beta_{jk} q_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K}\sum_{k=1}^{K-1} \beta_{j(kk')} q_{jk} q_{jk'}\alpha_{ik}\alpha_{ik'} + \cdots + \beta_{j12\ldots K}\prod_{k=1}^{K} q_{jk}\alpha_{ik}$$

Suppose the attribute profiles $\boldsymbol{\alpha}$ are known. The estimators of the elements of the item parameter vector, $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \beta_{j2}, \ldots, \beta_{j12\ldots K})'$, are derived by maximizing the item likelihood

$$L_j(\boldsymbol{\beta}_j; \mathbf{y}_j, \boldsymbol{\alpha}) = \prod_{i=1}^{N} f(y_{ij} \mid \boldsymbol{\beta}_j, \boldsymbol{\alpha}_i) = \prod_{i=1}^{N}\left(\frac{\exp(v_j)}{1 + \exp(v_j)}\right)^{y_{ij}}\left(\frac{1}{1 + \exp(v_j)}\right)^{1-y_{ij}}.$$

Maximizing the logarithm of the item likelihood results in the same estimator as maximizing the likelihood

$$\ln L_j(\boldsymbol{\beta}_j; \mathbf{y}_j, \boldsymbol{\alpha}) = \sum_{i=1}^{N} y_{ij} \ln \left(\frac{\exp(v_j)}{1 + \exp(v_j)} \right) + \sum_{i=1}^{N} (1 - y_{ij}) \ln \left(\frac{1}{1 + \exp(v_j)} \right)$$

$$= \sum_{i=1}^{N} \left(y_{ij} v_j - \ln \left(1 + \exp(v_j) \right) \right). \tag{5}$$

Taking the derivative of Equation 3 with regard to β_{j0} results in

$$\frac{\partial \ln(L_j)}{\partial \beta_{j0}} = \sum_{i=1}^{N} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right).$$

Similarly,

$$\frac{\partial \ln(L_j)}{\partial \beta_{jk}} = q_{jk} \sum_{i=1}^{N} \alpha_{ik} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) \forall k \tag{6}$$

and

$$\frac{\partial \ln(L_j)}{\partial \beta_{jkk'}} = q_{jk} q_{jk'} \sum_{i=1}^{N} \alpha_{ik} \alpha_{ik'} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) \forall k, k' \tag{7}$$

and so on. The derivative of Equation 3 with regard to the coefficient of the highest interaction term $\beta_{j12\ldots K}$ is

$$\frac{\partial \ln(L_j)}{\partial \beta_{j12\ldots K}} = q_{j1} q_{j2} \ldots q_{jK} \sum_{i=1}^{N} \alpha_{i1} \alpha_{i2} \ldots \alpha_{iK} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right). \tag{8}$$

The inspection of the partial derivatives in Equations 6, 7, and 8 shows that they are reduced to zero if item j does not require attribute k because then the corresponding q_{jk} is zero. Assume that 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 $\mathbf{q}_j$. Because $q_{j1} = q_{j2} = \ldots = q_{jK_j^*} = 1$, they become implicit, whereas all α_{jk} corresponding to $q_{jk} = 0$ are eliminated from the expressions of the partial derivatives in Equations 6, 7, and 8:

$$\frac{\partial \ln(L_j)}{\partial \beta_{jk}} = \sum_{i=1}^{N} \alpha_{ik} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) \forall k \in \mathcal{L}_j, \tag{9}$$

where $\mathcal{L}_j = \{1, 2, \ldots, K_j^*\}$ is defined as the collection of indices of the non-zero elements in $\mathbf{q}_j$,

$$\frac{\partial \ln(L_j)}{\partial \beta_{jkk'}} = \sum_{i=1}^{N} \alpha_{ik} \alpha_{ik'} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) \forall k, k' \in \mathcal{L}_j \tag{10}$$

and

$$\frac{\partial \ln(L_j)}{\partial \beta_{j12\ldots K_j^*}} = \sum_{i=1}^{N} \alpha_{i1}\alpha_{i2}\ldots\alpha_{iK_j^*} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right). \tag{11}$$

The expressions of the item parameter estimators are derived in working backwards beginning with Equation 11. Using the indicator function $I[\cdot]$ and setting Equation 11 to zero yields

$$\sum_{i=1}^{N} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) I\left[\bigcap_{k=1}^{K_j^*} \{\alpha_{ik} = 1\} \right] = 0$$

which is equivalent to

$$\sum_{i \in \mathcal{C}(\mathcal{L}_j)} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) = 0, \tag{12}$$

where $\mathcal{C}(\mathcal{L}_j)$ denotes the proficiency class consisting of examinees who master all the K_j^* items required for item j. In general, define the proficiency class $\mathcal{C}(\mathcal{A})$, with $\mathcal{A} \subseteq \mathcal{L}_j = \{1, 2, \ldots, K_j^*\}$, such that $\mathcal{C}(\mathcal{A}) = \{i \mid \alpha_{ik} = 1, \forall k \in \mathcal{A}$ and $\alpha_{ik'} = 0, \forall k' \in \mathcal{A}^c\}$. Equation 12 implies

$$\sum_{i \in \mathcal{C}(\mathcal{L}_j)} y_{ij} - |\mathcal{C}(\mathcal{L}_j)| \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk} + \cdots + \beta_{j12\ldots K_j^*}\right)}{1 + \exp\left(\beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk} + \cdots + \beta_{j12\ldots K_j^*}\right)} = 0$$

$$\Rightarrow \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk} + \cdots + \beta_{j12\ldots K_j^*}\right)}{1 + \exp\left(\beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk} + \cdots + \beta_{j12\ldots K_j^*}\right)} = \frac{\sum_{i \in \mathcal{C}(\mathcal{L}_j)} y_{ij}}{|\mathcal{C}(\mathcal{L}_j)|} = \bar{y}_{j\,\mathcal{C}(\mathcal{L}_j)},$$

where $\bar{y}_{j\,\mathcal{C}(\mathcal{L}_j)}$ is the mean response of proficiency class $\mathcal{C}(\mathcal{L}_j)$ of item j. Therefore,

$$\hat{\beta}_{j0} + \sum_{k=1}^{K_j^*} \hat{\beta}_{jk} + \cdots + \hat{\beta}_{j12\ldots K_j^*} = \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\mathcal{L}_j)}}{1 - \bar{y}_{j\,\mathcal{C}(\mathcal{L}_j)}} \right).$$

There are $\binom{K^*}{K_j^*-1} = K_j^*$ interaction terms of order $(K_j^* - 1)$. Without loss of generality, only the partial derivative of $\ln(L_j)$ with regard to the parameter $\beta_{j12\ldots(K_j^*-1)}$ is analyzed here in detail:

$$\frac{\partial \ln(L_j)}{\partial \beta_{j12\ldots(K_j^*-1)}} = \sum_{i=1}^{N} \alpha_{i1}\alpha_{i2}\ldots\alpha_{i(K_j^*-1)} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right). \tag{13}$$

Setting Equation 13 to zero and using the indicator function leads to

$$
\sum_{i=1}^{N} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) I \left[\bigcap_{k=1}^{K_j^*-1} \{\alpha_{ik} = 1\} \right] = 0
$$

$$
\Rightarrow \sum_{i \in \mathcal{C}(\{1,2,\ldots,K_j^*-1\})} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) + \sum_{i \in \mathcal{C}(\mathcal{L}_j)} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) = 0
$$

$$
\Rightarrow \sum_{i \in \mathcal{C}(\{1,2,\ldots,K_j^*-1\})} \left(y_{ij} - \frac{\exp(v_j)}{1 + \exp(v_j)} \right) = 0. \tag{14}
$$

(The second sum on the left-hand side of Equation 14 is zero due to Equation 12.) The last equation can be written explicitly as

$$
\sum_{i \in \mathcal{C}(\{1,2,\ldots,K_j^*-1\})} y_{ij} - |\mathcal{C}(\{1,2,\ldots,K_j^*-1\})| \frac{\exp\left(\beta_{j0}+\sum_{k=1}^{K_j^*}\beta_{jk}+\cdots+\beta_{j12\ldots(K_j^*-1)}\right)}{1+\exp\left(\beta_{j0}+\sum_{k=1}^{K_j^*}\beta_{jk}+\cdots+\beta_{j12\ldots(K_j^*-1)}\right)} = 0
$$

providing the final result

$$
\hat{\beta}_{j0} + \sum_{k=1}^{K_j^*-1} \hat{\beta}_{jk} + \cdots + \hat{\beta}_{j12\ldots(K_j^*-1)} = \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{1,2,\ldots,K_j^*-1\})}}{1 - \bar{y}_{j\,\mathcal{C}(\{1,2,\ldots,K_j^*-1\})}} \right).
$$

The partial derivatives for the remaining parameters are manipulated in the same manner; suffice it here to present the results for Equations 10 and 9

$$
\hat{\beta}_{j0} + \hat{\beta}_{jk} + \hat{\beta}_{jk'} + \hat{\beta}_{jkk'} = \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k,k'\})}}{1 - \bar{y}_{j\,\mathcal{C}(\{k,k'\})}} \right) \tag{15}
$$

$$
\hat{\beta}_{j0} + \hat{\beta}_{jk} = \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k\})}}{1 - \bar{y}_{j\,\mathcal{C}(\{k\})}} \right) \tag{16}
$$

and finally

$$
\hat{\beta}_{j0} = \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\emptyset)}}{1 - \bar{y}_{j\,\mathcal{C}(\emptyset)}} \right). \tag{17}
$$

The expressions of the estimators of $\hat{\beta}_{jk}$ are then obtained by subtracting $\hat{\beta}_{j0}$ of Equation 17 from Equation 16

$$
\hat{\beta}_{jk} = \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k\})}}{1 - \bar{y}_{j\,\mathcal{C}(\{k\})}} \right) - \hat{\beta}_{j0}
$$

$$
= \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k\})}}{1 - \bar{y}_{j\,\mathcal{C}(\{k\})}} \right) - \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\emptyset)}}{1 - \bar{y}_{j\,\mathcal{C}(\emptyset)}} \right).
$$

The expression of the estimators of $\hat{\beta}_{jkk'}$ is obtained by subtracting the lower-order parameter estimators $\hat{\beta}_{j0}$, $\hat{\beta}_{jk}$, and $\hat{\beta}_{jk'}$ from the left-hand sum of estimators in Equation 15

$$
\begin{aligned}
\hat{\beta}_{jkk'} &= \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k,k'\})}}{1-\bar{y}_{j\,\mathcal{C}(\{k,k'\})}}\right) - \hat{\beta}_{jk} - \hat{\beta}_{jk'} - \hat{\beta}_{j0} \\
&= \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k,k'\})}}{1-\bar{y}_{j\,\mathcal{C}(\{k,k'\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k\})}}{1-\bar{y}_{j\,\mathcal{C}(\{k\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{k'\})}}{1-\bar{y}_{j\,\mathcal{C}(\{k'\})}}\right) + \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\emptyset)}}{1-\bar{y}_{j\,\mathcal{C}(\emptyset)}}\right).
\end{aligned}
$$

In summary, if examinees' attribute profiles are known, then the (closed-form) expressions of the estimators $\hat{\beta}_{j\,\ldots}$ of the coefficients of k-order terms, $k \in \{1, 2, \ldots, K_j^*\}$ are functions of the means of the 2^k proficiency classes characterized by attribute profiles $\boldsymbol{\alpha}$, where the first k attributes are mastered or not, and the remaining attributes are not mastered. For example, let $K^* = 3$; then, $\hat{\beta}_{12}$ is a function of the means of the four proficiency classes with $\boldsymbol{\alpha}_{\mathcal{C}(\emptyset)} = (000)'$, $\boldsymbol{\alpha}_{\mathcal{C}(\{1\})} = (100)'$, $\boldsymbol{\alpha}_{\mathcal{C}(\{2\})} = (010)'$, and $\boldsymbol{\alpha}_{\mathcal{C}(\{1,2\})} = (110)'$

$$
\begin{aligned}
\hat{\beta}_{j12} &= \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{1,2\})}}{1-\bar{y}_{j\,\mathcal{C}(\{1,2\})}}\right) - \hat{\beta}_{j1} - \hat{\beta}_{j2} - \hat{\beta}_{j0} \\
&= \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{1,2\})}}{1-\bar{y}_{j\,\mathcal{C}(\{1,2\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{1\})}}{1-\bar{y}_{j\,\mathcal{C}(\{1\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\{2\})}}{1-\bar{y}_{j\,\mathcal{C}(\{2\})}}\right) + \ln\left(\frac{\bar{y}_{j\,\mathcal{C}(\emptyset)}}{1-\bar{y}_{j\,\mathcal{C}(\emptyset)}}\right).
\end{aligned}
$$

The estimators of the $\beta_{j\,\ldots}$ under the proposed JMLE framework are obtained by solving Equation 12 with α_{ik} replaced by $\tilde{\alpha}_{ik}$. To be specific, define the proficiency class $\tilde{\mathcal{C}}(\mathcal{A}) = \{i \mid \tilde{\alpha}_{ik} = 1, \forall k \in \mathcal{A} \text{ and } \tilde{\alpha}_{ik'} = 0, \forall k' \in \mathcal{A}^c\}$. The expressions of the estimators of the item parameters $\tilde{\beta}_{j0}$, $\tilde{\beta}_{jk}$, and $\tilde{\beta}_{jkk'}$, are derived as

$$
\begin{aligned}
\tilde{\beta}_{j0} &= \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\emptyset)}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\emptyset)}}\right) \\
\tilde{\beta}_{jk} &= \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\{k\})}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\{k\})}}\right) - \tilde{\beta}_{j0} \\
&= \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\{k\})}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\{k\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\emptyset)}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\emptyset)}}\right) \\
\tilde{\beta}_{jkk'} &= \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\{k,k'\})}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\{k,k'\})}}\right) - \tilde{\beta}_{jk} - \tilde{\beta}_{jk'} - \tilde{\beta}_{j0} \\
&= \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\{k,k'\})}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\{k,k'\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\{k\})}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\{k\})}}\right) - \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\{k'\})}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\{k'\})}}\right) + \ln\left(\frac{\bar{y}_{j\,\tilde{\mathcal{C}}(\emptyset)}}{1-\bar{y}_{j\,\tilde{\mathcal{C}}(\emptyset)}}\right).
\end{aligned}
\tag{18}
$$

The expressions of the estimators of the remaining parameters can be readily deduced from the profile emerging from the equations of $\hat{\beta}_{j0}$, $\tilde{\beta}_{jk}$, and $\tilde{\beta}_{jkk'}$.

5. Consistency of the Item Parameter Estimators

It was stated earlier that the estimators of the item parameters of the LCDM, $\tilde{\beta}_{j\ldots}$, derived by differentiating the joint likelihood function are statistically consistent provided the estimation procedure uses an (external) estimator of examinees' proficiency class membership that is itself statistically consistent. In the previous section, the conceptual distinction was made between the item parameter estimators $\hat{\beta}_{j\ldots}$ and $\tilde{\beta}_{j\ldots}$ depending on whether examinees' attribute profiles are known or must be estimated using the consistent estimator $\tilde{\alpha}$. In this section, two consistency theorems are presented that establish asymptotic consistency of $\tilde{\beta}_{j\ldots}$ (Theorem 1) and a stronger form of consistency called *uniform* consistency of $\tilde{\beta}_{j\ldots}$ (Theorem 2). To avoid any redundancy, only the proofs concerning $\tilde{\beta}_{j0}$ are presented because the proofs for $\tilde{\beta}_{jk}$, $\tilde{\beta}_{jkk'}$, and so on, can be readily constructed using the same argument. The proofs of Theorems 1 and 2 require two lemmas that establish (a) the normality of $\hat{\beta}_{j0}$ and (b) the convergence of $\tilde{\beta}_{j0}$ to $\hat{\beta}_{j0}$. These lemmas are presented first.

Lemma 1. *Let $\hat{\beta}_{j0}$ be the estimator of the parameter β_{j0}, $j = 1, \ldots, J$, of the LCDM when examinees' attribute profiles $\boldsymbol{\alpha}_i$ are known (see Equation 17). Let $\mathcal{A}$ be a subset of $\mathcal{L}_j = \{1, 2, \ldots, K_j^*\}$ for item j, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ and $\mathcal{C}(\mathcal{A}) = \{i \mid \alpha_{ik} = 1, \forall k \in \mathcal{A} \text{ and } \alpha_{ik'} = 0, \forall k' \in \mathcal{A}^c\}$. Then, $\sqrt{|\mathcal{C}(\emptyset)|}\,(\hat{\beta}_{j0} - \beta_{j0}) \xrightarrow{\mathcal{D}} N\left(0, \frac{(1+\exp(\beta_{j0}))^2}{\exp(\beta_{j0})}\right)$ for all j.*

Proof. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ})$ be the item response vector of examinee i. Based on the item response function of the LCDM in Equation 1, the conditional expectation of Y_{ij} reduces to

$$E(Y_{ij} \mid \boldsymbol{\alpha}_i)$$
$$= \frac{\exp\left(\beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \beta_{jkk'}\alpha_{ik}\alpha_{ik'} + \cdots + \beta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}\right)}{1 + \exp\left(\beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \beta_{jkk'}\alpha_{ik}\alpha_{ik'} + \cdots + \beta_{j12\ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}\right)}.$$

If $\boldsymbol{\alpha}_i = (0, 0, \ldots, 0)'$, then the conditional expectation of Y_{ij} is

$$E\left(Y_{ij} \mid \boldsymbol{\alpha}_i = (0, 0, \ldots, 0)'\right) = \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})}.$$

Because Y_{ij} is binary, the variance of Y_{ij} is

$$\mathrm{Var}\left(Y_{ij} \mid \boldsymbol{\alpha}_i = (0, 0, \ldots, 0)'\right) = \frac{\exp(\beta_{j0})}{\left(1 + \exp(\beta_{j0})\right)^2}$$

with $0 < \frac{\exp(\beta_{j0})}{\left(1+\exp(\beta_{j0})\right)^2} < \infty$. Due to the Central Limit Theorem

$$\sqrt{|\mathcal{C}(\emptyset)|}\left(\bar{Y}_{j\,\mathcal{C}(\emptyset)} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})}\right) \xrightarrow{\mathcal{D}} N\left(0, \frac{\exp(\beta_{j0})}{\left(1 + \exp(\beta_{j0})\right)^2}\right) \forall j \qquad (19)$$

where $\bar{Y}_{j\,\mathcal{C}(\emptyset)}$ is the mean of the responses Y_{ij} in proficiency class $\mathcal{C}(\emptyset)$. Define the function $g : (0, 1) \to \mathbb{R}$ as $g(x) = \ln(\frac{x}{1-x})$. Then g has a continuous derivative at $\frac{\exp(\beta_{j0})}{1+\exp(\beta_{j0})}$ because $0 < \frac{\exp(\beta_{j0})}{1+\exp(\beta_{j0})} < 1$. Applying the Delta Method to Equation 19 results in

$$\sqrt{|\mathcal{C}(\emptyset)|} \left(g(\bar{Y}_{j\,\mathcal{C}(\emptyset)}) - g\left(\frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right) \right) \xrightarrow{\mathcal{D}}$$
$$N\left(0, \frac{\exp(\beta_{j0})}{\left(1 + \exp(\beta_{j0})\right)^2} \left(g'\left(\frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right) \right)^2 \right)$$

which can be simplified to

$$\sqrt{|\mathcal{C}(\emptyset)|} \left(\ln\left(\frac{\bar{Y}_{j\,\mathcal{C}(\emptyset)}}{1 - \bar{Y}_{j\,\mathcal{C}(\emptyset)}} \right) - \beta_{j0} \right) \xrightarrow{\mathcal{D}} N\left(0, \frac{\left(1 + \exp(\beta_{j0})\right)^2}{\exp(\beta_{j0})} \right).$$

Recall that the estimator $\hat{\beta}_{j0} = \ln \frac{\bar{Y}_{j\,\mathcal{C}(\emptyset)}}{1 - \bar{Y}_{j\,\mathcal{C}(\emptyset)}}$ (see Equation 17); hence,

$$\sqrt{|\mathcal{C}(\emptyset)|} \left(\hat{\beta}_{j0} - \beta_{j0} \right) \xrightarrow{\mathcal{D}} N\left(0, \frac{\left(1 + \exp(\beta_{j0})\right)^2}{\exp(\beta_{j0})} \right)$$

for all j.   □

Lemma 2. *Let $\hat{\beta}_{j0}$ and $\tilde{\beta}_{j0}$ be the parameter estimators as defined in Equations 17 and 18, respectively. Assume $J \to \infty$ and $J < N$. Then $\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \hat{\beta}_{j0}$ for all j.*

Proof. Because $\tilde{\alpha}$ is a consistent estimator of α

$$P\left(\bigcup_{i=1}^{N} \{ |\tilde{\alpha}_i - \alpha_i| > \varepsilon \} \right) \to 0 \tag{20}$$

as $J \to \infty$. Equations 17 and 18 show that $\tilde{\beta}_{j0} = \hat{\beta}_{j0}$ if $\tilde{\alpha}_i = \alpha_i$ for all i, which can be expressed as

$$P\left(\{ \tilde{\alpha}_i = \alpha_i \} \right) \leq P\left(\{ \tilde{\beta}_{j0} = \hat{\beta}_{j0} \} \right)$$

or equivalently as

$$P\left(\{ \tilde{\beta}_{j0} \neq \hat{\beta}_{j0} \} \right) \leq P\left(\{ \tilde{\alpha}_i \neq \alpha_i \} \right) \forall i$$

Hence, for every $\varepsilon > 0$,

$$P\left(|\tilde{\beta}_{j0} - \hat{\beta}_{j0}| > \varepsilon \right) \leq P\left(\bigcup_{i=1}^{N} \{ |\tilde{\alpha}_i - \alpha_i| > \varepsilon \} \right) \tag{21}$$

Due to Equation 20, Equation 21 also implies

$$P\big(|\tilde{\beta}_{j0} - \hat{\beta}_{j0}| > \varepsilon\big) \to 0$$

as $J \to \infty$. Note that $J \to \infty$ implies $N \to \infty$ because $N > J$. Therefore,

$$\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \hat{\beta}_{j0}$$

for all j. □

In reiterating, Lemma 1 establishes that the estimator $\hat{\beta}_{j0}$ is normally distributed if examinees' attribute profiles $\boldsymbol{\alpha}_i$ are known—a useful property if hypothesis testing or interval estimation is desired. Lemma 2 establishes the convergence of $\tilde{\beta}_{j0}$ to $\hat{\beta}_{j0}$, and thus, connects $\tilde{\beta}_{j0}$ with the true β_{j0}. Lemmas 1 and 2 are instrumental for the proofs of Theorems 1 and 2.

Theorem 1. *Suppose that there exist $0 < \varepsilon_1, \varepsilon_2 < 1$ such that $\varepsilon_1 < \frac{|\mathcal{C}(\emptyset)|}{N} < 1 - \varepsilon_2$. Let $\hat{\beta}_{j0}$ and $\tilde{\beta}_{j0}$ be the estimators of β_{j0} with properties as specified in Lemmas 1 and 2. Assume $N > J$ and $J \to \infty$. Also, assume $\tilde{\boldsymbol{\alpha}}_i$ is a consistent estimator of $\boldsymbol{\alpha}_i$. Then $\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \beta_{j0}$ for all j.*

Proof. Lemma 1 stated

$$\sqrt{|\mathcal{C}(\emptyset)|}(\hat{\beta}_{j0} - \beta_{j0}) \xrightarrow{\mathcal{D}} N\left(0, \frac{\big(1 + \exp(\beta_{j0})\big)^2}{\exp(\beta_{j0})}\right). \tag{22}$$

Lemma 2 stated

$$\tilde{\beta}_{j0} - \hat{\beta}_{j0} \xrightarrow{\mathcal{P}} 0 \,\forall j \tag{23}$$

Because $\varepsilon_1 < \frac{|\mathcal{C}(\emptyset)|}{N} < 1 - \varepsilon_2$, the condition $|\mathcal{C}(\emptyset)| \to \infty$ as $N \to \infty$ is guaranteed. Then, because $\frac{1}{\sqrt{|\mathcal{C}(\emptyset)|}} \xrightarrow{\mathcal{P}} 0$, due to Slutsky's theorem, Equation 22 can be written as

$$\hat{\beta}_{j0} - \beta_{j0} \xrightarrow{\mathcal{D}} \left(\frac{1}{\sqrt{|\mathcal{C}(\emptyset)|}}\right) N\left(0, \frac{\big(1 + \exp(\beta_{j0})\big)^2}{\exp(\beta_{j0})}\right) = 0 \,\forall j \tag{24}$$

Note that because $\hat{\beta}_{j0} - \beta_{j0}$ converges to 0 in distribution, and 0 is a constant, $\hat{\beta}_{j0} - \beta_{j0}$ converges to 0 in probability as well—that is,

$$\hat{\beta}_{j0} - \beta_{j0} \xrightarrow{\mathcal{P}} 0 \,\forall j$$

Now, applying Slutsky's theorem again to Equations 23 and 24 yields

$$\tilde{\beta}_{j0} - \hat{\beta}_{j0} + \hat{\beta}_{j0} - \beta_{j0} = \tilde{\beta}_{j0} - \beta_{j0} \xrightarrow{\mathcal{D}} 0$$

which implies

$$\tilde{\beta}_{j0} - \beta_{j0} \xrightarrow{\mathcal{P}} 0 \,\forall j$$

because 0 is a constant—or equivalently,

$$\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \beta_{j0}$$

for all j.   □

Theorem 1 states that $\tilde{\beta}_{j0}$ converges in probability to β_{j0}; hence, $\tilde{\beta}_{j0}$ is a statistically consistent estimator of β_{j0}. Theorem 2 states that $\tilde{\beta}_{j0}$ is also a uniformly consistent estimator of β_{j0}. The proof of Theorem 2 requires a modification of Lemma 2 that is given here as

Proposition 1. *If $\tilde{\boldsymbol{\alpha}}_i \xrightarrow{\mathcal{P}} \boldsymbol{\alpha}_i$ uniformly, $\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \hat{\beta}_{j0}$ uniformly.*

Proof. The claim $\tilde{\beta}_{j0} = \hat{\beta}_{j0}$ if $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i$ for all i (used already in the proof of Lemma 2) also implies

$$P\left(\bigcap_{i=1}^{N} \{\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i\}\right) \leq P\left(\bigcap_{j=1}^{J} \{\tilde{\beta}_{j0} = \hat{\beta}_{j0}\}\right)$$

$$\Rightarrow P\left(\bigcup_{j=1}^{J} \{\tilde{\beta}_{j0} \neq \hat{\beta}_{j0}\}\right) \leq P\left(\bigcup_{i=1}^{N} \{\tilde{\boldsymbol{\alpha}}_i \neq \boldsymbol{\alpha}_i\}\right).$$

Hence, for every $\varepsilon > 0$,

$$P\left(\bigcup_{j=1}^{J} \{|\tilde{\beta}_{j0} - \hat{\beta}_{j0}| > \varepsilon\}\right) \leq P\left(\bigcup_{i=1}^{N} \{|\tilde{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\}\right). \tag{25}$$

Because $\tilde{\boldsymbol{\alpha}}_i \xrightarrow{\mathcal{P}} \boldsymbol{\alpha}_i$ uniformly

$$P\left(\bigcup_{i=1}^{N} \{|\tilde{\boldsymbol{\alpha}}_i - \boldsymbol{\alpha}_i| > \varepsilon\}\right) \xrightarrow{\mathcal{P}} 0. \tag{26}$$

Combining Equations 25 and 26 results in

$$P\left(\bigcup_{j=1}^{J} \{|\tilde{\beta}_{j0} - \hat{\beta}_{j0}| > \varepsilon\}\right) \to 0$$

as $J \to \infty$. Note that $J \to \infty$ implies $N \to \infty$ because $N > J$. Therefore,

$$\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \hat{\beta}_{j0}$$

uniformly.   □

Theorem 2. *Assume $\tilde{\alpha}_i$ is a uniformly consistent estimator of α_i. Also, assume $J < N$ and $\varepsilon_1 < \frac{|\mathcal{C}(\emptyset)|}{N} < 1 - \varepsilon_2$, where $0 < \varepsilon_1, \varepsilon_2 < 1$. Then $\tilde{\beta}_{j0}$ is a uniformly consistent estimator of β_{j0} for all j if $N \exp(-J) \to 0$ as $J \to \infty$.*

Proof. The proof of Theorem 2 uses Hoeffding's theorem (Hoeffding, 1963), which states that $P(|\frac{1}{N} \sum_{i=1}^{N} X_i - E(X_i)| > \epsilon) < 2 \exp(-2N\epsilon^2)$ where $X_1, \ldots, X_N$ are iid random variables and $0 \le X_i \le 1$ for all i. In formal agreement with Hoeffding's theorem, the argument is developed for $\frac{\exp(\tilde{\beta}_{j0})}{1+\exp(\tilde{\beta}_{j0})}$ (and not $\tilde{\beta}_{j0}$)—the final convergence result, however, is transformed into an expression in terms of $\tilde{\beta}_{j0}$. Consider

$$\max_j \left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| \le \max_j \left(\left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} \right| \right.$$
$$\left. + \left| \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| \right).$$

Hence,

$$P\left(\max_j \left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| > \varepsilon \right)$$
$$\le P\left(\max_j \left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} \right| \right.$$
$$\left. + \max_j \left| \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| > \varepsilon \right)$$
$$\le P\left(\left\{ \max_j \left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} \right| > \frac{\varepsilon}{2} \right\} \right.$$
$$\left. \bigcup \left\{ \max_j \left| \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| > \frac{\varepsilon}{2} \right\} \right)$$
$$\le P\left(\max_j \left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} \right| > \frac{\varepsilon}{2} \right)$$
$$+ P\left(\max_j \left| \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| > \frac{\varepsilon}{2} \right).$$

Due to Hoeffding's theorem and the assumption $\varepsilon_1 < \frac{|\mathcal{C}(\emptyset)|}{N} < 1 - \varepsilon_2$, the first term in the last line can be written as

$$P\left(\max_j \left| \frac{\exp(\tilde{\beta}_{j0})}{1 + \exp(\tilde{\beta}_{j0})} - \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} \right| > \frac{\varepsilon}{2} \right)$$
$$= P\left(\bigcup_j \left\{ \left| \frac{\exp(\hat{\beta}_{j0})}{1 + \exp(\hat{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1 + \exp(\beta_{j0})} \right| > \frac{\varepsilon}{2} \right\} \right)$$

$$\leq \sum_{j=1}^{J} P\left(\left|\frac{\exp(\hat{\beta}_{j0})}{1+\exp(\hat{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1+\exp(\beta_{j0})}\right| > \frac{\varepsilon}{2}\right)$$

$$\leq J\left(2\exp\left(-2|\mathcal{C}(\emptyset)|\left(\frac{\varepsilon}{2}\right)^2\right)\right)$$

$$\leq 2N\exp\left(-N\varepsilon_1\frac{\varepsilon^2}{2}\right)$$

$$= 2N\exp\left(-N\frac{(\sqrt{\varepsilon_1}\varepsilon)^2}{2}\right)$$

$$\leq 2N\exp\left(-J\frac{(\sqrt{\varepsilon_1}\varepsilon)^2}{2}\right) \longrightarrow 0 \qquad (27)$$

as $J \to \infty$, $N\exp(-J) \to 0$ for all ε. Proposition 1 states that $\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \hat{\beta}_{j0}$ uniformly. Consider the continuous function $t : \mathbb{R} \to (0, 1)$ defined as $t(x) = \frac{\exp(x)}{1+\exp(x)}$. Applying the Continuous Mapping Theorem to Proposition 1 results in

$$\frac{\exp(\tilde{\beta}_{j0})}{1+\exp(\tilde{\beta}_{j0})} \xrightarrow{\mathcal{P}} \frac{\exp(\hat{\beta}_{j0})}{1+\exp(\hat{\beta}_{j0})}$$

uniformly. Thus,

$$P\left(\max_j\left|\frac{\exp(\tilde{\beta}_{j0})}{1+\exp(\tilde{\beta}_{j0})} - \frac{\exp(\hat{\beta}_{j0})}{1+\exp(\hat{\beta}_{j0})}\right| > \frac{\varepsilon}{2}\right) \to 0 \qquad (28)$$

as $J \to \infty$, $N\exp(-J) \to 0$ for all ε. Then, Equations 27 and 28 imply that

$$P\left(\max_j\left|\frac{\exp(\tilde{\beta}_{j0})}{1+\exp(\tilde{\beta}_{j0})} - \frac{\exp(\beta_{j0})}{1+\exp(\beta_{j0})}\right| > \varepsilon\right) \to 0$$

as $J \to \infty$ and $N\exp(-J) \to 0$. Equivalently,

$$\frac{\exp(\tilde{\beta}_{j0})}{1+\exp(\tilde{\beta}_{j0})} \xrightarrow{\mathcal{P}} \frac{\exp(\beta_{j0})}{1+\exp(\beta_{j0})}$$

uniformly. Notice that the assumption $J < N$ implies that $N \to \infty$ as $J \to \infty$. Hence, $\frac{\exp(\tilde{\beta}_{j0})}{1+\exp(\tilde{\beta}_{j0})} \xrightarrow{\mathcal{P}} \frac{\exp(\beta_{j0})}{1+\exp(\beta_{j0})}$ as $N \to \infty$ and $N\exp(-J) \to 0$. Define the continuous function $g : (0, 1) \to \mathbb{R}$ as $g(x) = \ln(\frac{x}{1-x})$. Thus, based on the Continuous Mapping Theorem

$$\tilde{\beta}_{j0} \xrightarrow{\mathcal{P}} \beta_{j0}$$

uniformly. $\square$

In summary, Theorems 1 and 2 address two different kinds of convergences. Theorem 1 establishes the pointwise convergence in probability, whereas Theorem 2 establishes uniform convergence, the guarantee that all parameter estimators converge, which is a much stronger form of statistical consistency. (Note that uniform convergence implies pointwise convergence. However, Theorem 1 has been presented here too because its proof uses a different technical argument and has its own theoretical merit.) Finally, it should be noted that the assumptions $J < N$ (i.e., the test length must not exceed the number of examinees) and $\varepsilon_1 < \frac{N_j}{N} < 1 - \varepsilon_2$ are often made automatically in testing without explicit mentioning.

6. Simulation Studies

Two simulation studies were conducted to evaluate the performance of the JMLE estimators on artificial data sets under finite data conditions for selected test settings. All data sets conformed to the LCDM. The item parameter estimates and the classification of examinees obtained from JMLE were compared to those obtained from MMLE-EM (in using the implementation of the EM algorithm in Mplus; Muthén & Muthén, 1998–2012; see also, Templin & Hoffman, 2013). In addition, all data sets were also fitted by a procedure called conditional maximum likelihood estimation (CMLE). CMLE uses examinees' (known) true attribute profiles as input when estimating the item parameters, and the (known) true item parameters as input when estimating examinees' attribute profiles. Thus, CMLE provided a most conservative benchmark for the results obtained from JMLE and MMLE-EM. (CMLE is implemented in the R package NPCD, Zheng & Chiu, 2014).

6.1. Implementation of Joint Maximum Likelihood Estimation

6.1.1. The Consistent Estimator $\tilde{\alpha}$ of Examinees' Proficiency Class Membership   The Nonparametric Classification (NPC) method (Chiu & Douglas, 2013) was used here to obtain estimates of examinees' attribute profiles α (i.e., their proficiency class membership) that are needed as input to the JMLE algorithm. Wang and Douglas (2015) proved that under certain regularity conditions $\tilde{\alpha}$ obtained by the NPC method is a statistically consistent estimator of an examinee's attribute profile 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 classication will be consistent as the test length increases." (Wang & Douglas, 2015, p. 99)

The NPC method estimates the proficiency class membership of examinees by comparing their observed item response profiles with each of the ideal response profiles of the possible proficiency classes. The ideal response is a function of the q-vector of item j, $\mathbf{q}_j$, and the attribute profile $\boldsymbol{\alpha}_m$ of proficiency class $\mathcal{C}_m$. The ideal response to item j is the score that would be realized by an examinee in proficiency class $\mathcal{C}_m$ (having attribute profile $\boldsymbol{\alpha}_m$) if no perturbation occurred. The NPC estimator $\tilde{\alpha}$ of an examinee's attribute profile is defined as the attribute profile associated with that ideal item response profile, which minimizes the distance between all ideal item response profiles and an examinee's observed item response profile. Said differently, the estimator $\tilde{\alpha}$ identifies the attribute profile underlying that ideal item response profile, which among all possible ideal item response profiles is closest—or most similar—to the observed item response profile (for further technical details, consult Chiu & Douglas, 2013). An implementation of the NPC method is available in the R package NPCD (Zheng & Chiu, 2014).

6.1.2. JMLE Algorithm   The algorithm used here is an adaptation of Birnbaum's paradigm (Birnbaum, 1968), a two-stage procedure for JMLE (Baker & Kim, 2004; Embretson & Reise, 2000). Examinees' attribute profiles and the item parameters are treated as two sets where one is assumed to consist of known parameters, whereas those in the second set are to be estimated. The algorithm is initialized with the estimates of examinees' attribute profiles as input, which are obtained by the consistent estimators $\tilde{\alpha}_1, \tilde{\alpha}_2, \ldots, \tilde{\alpha}_N$ of the NPC method. The joint likelihood in Equation 2 then reduces to a function of only the item parameters. The estimator of β_j is derived by maximizing the logarithm of the item likelihood L_j (see Equation 5) for all j.

6.2. Simulation Study I

The purpose of Study I was to assess the accuracy of the JMLE estimates (i.e., item parameters and examinees' attribute profiles) with replicated data sets.

6.2.1. Design   Item responses conforming to the LCDM of $N = 3000$ examinees were generated. $K = 3$ attributes were used; the number of items was $J = 30, 40$. Examinees' attribute profiles were generated based on the multivariate normal threshold model, with variances and covariances set to 1 and 0.3, respectively (for further details, consult Chiu, Daouglas, & Li, 2009). The Q-matrices were designed such that (a) the maximum number of attributes required per item was two, (b) each attribute was used by 36% to 40% of the items (a proportion that corresponds to Q-matrix compositions used in other simulation studies of general CDMs; e.g., de la Torre, 2011; de la Torre & Chiu, 2010), and (c) they were complete (Chiu et al., 2009; Chiu & Köhn, 2015). Based on the result by Wang and Douglas (2015) mentioned earlier, the item parameters of the underlying LCDM were chosen such that $P(Y_{ij} = 1 \mid \xi_{ij} = 0)$ was less than 0.5 and $P(Y_{ij} = 1 \mid \xi_{ij} = 1)$ was about 0.8 (ξ is generic notation for the ideal item response). Examinees' manifest item responses were sampled from a Bernoulli distribution, with $P(Y_{ij} = 1 \mid \alpha)$ determined by the IRF of the LCDM. For each condition, 25 replicated data sets were generated. For each replicated data set, examinees' attribute profiles and responses were re-sampled, whereas the Q-matrix and the item parameters were held constant.

6.2.2. Results   The data were analyzed using JMLE, MMLE-EM, and CMLE. The multivariate normal threshold model that was used for generating examinees' attribute profiles imposes a higher-order structure on the attributes, which determines the distribution of the attribute profiles α that characterize the different proficiency classes. Recall that fitting educational data by a DCM, in addition to estimating the item parameters, requires estimating the distribution of the 2^K proficiency classes, which involves $2^K - 1$ parameters. Explicitly modeling the distribution of the α allows for a reduction of the number of model parameters, which might be useful especially if K, the number of attributes, is large. Mplus, for example, offers the option to construct a loglinear model for the distribution of α, which can range from a parsimonious main effects model to more complicated models with any order of interactions (for details, consult Rupp et al., 2010, Ch. 8). As a by-product, estimates of the correlations between the attributes can be derived from the parameter estimates of the loglinear model. The NPC method used for estimating examinees' α-profiles—and thus, the proportions of the different proficiency classes—relies on a nonparametric algorithm that, by definition, does not incorporate a (parametric) higher-order structure among the attributes. Chiu and Douglas (2013) established in a series of simulation studies the accuracy of the NPC method in estimating examinees' α-profiles for data without as well as for data with an underlying higher-order attribute structure. As JMLE relies on the input from NPC, JMLE does not—unlike Mplus—provide the means for explicitly modeling the distribution of the α. These conceptual and algorithmic differences between JMLE and Mplus-MMLE-EM raised the question how their performance could be compared in a fair and meaningful way. Specifically, when fitting the data with Mplus-MMLE-EM, should the distribution of α be explicitly

modeled, or not? Using `Mplus`-MMLE-EM without explicitly modeling the distribution of $\boldsymbol{\alpha}$ appeared to be the closest match to JMLE. But, on the other hand, this choice could lead to inferior `Mplus`-MMLE-EM parameter estimates; especially, when the number of attributes is large (as is shown in Simulation Study II below). Thus, as a reasonable compromise, the data were fitted by `Mplus`-MMLE-EM without and with explicitly modeling the distribution of $\boldsymbol{\alpha}$. (In the latter case, following Rupp et al. [2010], a logistic model including all main effects and all two-way interactions was used.)

The performance of JMLE, MMLE-EM, and CMLE in assigning examinees to their true proficiency classes was assessed by computing the classification-correct rate (CCR) (i.e., the proportion of correctly identified examinee attribute profiles). The accuracy of the item parameter estimates was evaluated by computing the root mean squared errors (RMSE). The CPU times were recorded for JMLE and MMLE-EM. The results are reported in Table 1. JMLE, MMLE-EM, and CMLE obtained (more or less) identical CCR values. Not too surprising, CMLE realized the lowest RMSE, whereas the RMSE of JMLE was slightly inferior to MMLE-EM. Table 1 also reports the 95% coverage intervals computed for CCR and RMSE; all three methods produce stable results, as can be concluded from the narrow confidence intervals. Surprisingly, modeling the distribution of $\boldsymbol{\alpha}$ explicitly with `Mplus` did neither appear to affect the CCR nor the accuracy of the item parameter estimates (assessed via the RMSE). The effect on CPU time was inconclusive: for $J = 30$ items, an average increase of the CPU time by about 30 min was observed, whereas for $J = 40$ the average CPU time decreased by almost 8 min. Perhaps the most remarkable result

TABLE 1.

Simulation Study I: average CCR and RMSE of JMLE, `Mplus` MMLE-EM, and CMLE plus average CPU times of JMLE and `Mplus` MMLE-EM when the data conformed to the LCDM; $N = 3000$, $K = 3$, 25 replications.

CCR

	JMLE	MMLE-EM	MMLE-EM-HO	CMLE
$J = 30$	0.929	0.929	0.929	0.930
LL	0.921	0.921	0.922	0.922
UL	0.938	0.937	0.937	0.938
$J = 40$	0.963	0.963	0.963	0.963
LL	0.957	0.956	0.957	0.957
UL	0.969	0.970	0.969	0.970

RMSE

	JMLE	MMLE-EM	MMLE-EM-HO	CMLE
$J = 30$	0.141	0.127	0.127	0.117
LL	0.106	0.095	0.096	0.087
UL	0.176	0.159	0.158	0.146
$J = 40$	0.129	0.126	0.126	0.120
LL	0.105	0.103	0.103	0.097
UL	0.153	0.149	0.148	0.144

CPU

	JMLE	MMLE-EM	MMLE-EM-HO	
$J = 30$	00:00:08	00:05:23	00:05:55	
$J = 40$	00:00:07	00:50:02	00:42:32	

LL and UL denote the lower and upper limit, respectively, of the 95% coverage interval; columns MMLE-EM/MMLE-EM-HO contain the results for fitting the data without/with explicitly modeling the distribution of $\boldsymbol{\alpha}$.

was the difference in average CPU times between JMLE and MMLE. The former required on average 7 and 8 s per data set, whereas the average CPU times used by MMLE-EM ranged from about 5–50 min per data set depending on the number of items, and whether the distribution of $\boldsymbol{\alpha}$ was explicitly modeled.

6.3. Simulation Study II

The purpose of Study 2 was to check the performance of JMLE for a relatively large number of attributes and items. Like in Study 1, the results of CMLE served as a benchmark.

6.3.1. Design   Compared to Study 1, the design of Study 2 needed to be stripped down so as to meet the CPU time requirements of `Mplus`. Hence, no replicated data sets were used; instead, four data sets were generated each containing the responses of $N = 3000$ examinees to $J = 30, 40, 50,$ and 60 items conforming to the LCDM. $K = 5$ attributes were used. The Q-matrices were again designed such that (a) the maximum number of attributes required per item was three, (b) each attribute was used by 36% to 40% of the items and (c) they were complete. Examinees' attribute profiles were generated based on the multivariate normal threshold model, with variances and covariances set to 1 and 0.3, respectively.

6.3.2. Results   Like in Simulation Study I, the data were fitted by `Mplus`-MMLE-EM without and with explicitly modeling the distribution of $\boldsymbol{\alpha}$ (using a logistic model including all main effects and all two-way interactions; see Rupp et al., 2010). Table 2 reports the results observed on the four LCDM data sets. CMLE realized the highest CCR and the lowest RMSE, whereas JMLE and MMLE-EM attained about the same CCR and RMSE. Similar to Simulation Study I, the findings concerning the effect of explicitly modeling the distribution of $\boldsymbol{\alpha}$ on the CPU times for `Mplus`-MMLE-EM were inconclusive. For $J = 30$ items, the CPU time increased when the distribution of $\boldsymbol{\alpha}$ was explicitly modeled, whereas the CPU times decreased for $J = 40, 50$. Remarkably, for $J = 60$, `Mplus`-MMLE-EM did not provide any results when used without explicitly modeling the distribution of $\boldsymbol{\alpha}$. (Perhaps, without imposing this additional structure, the number of parameters to be estimated is simply too large when the test involves 60 items— however, an observation made on a single data set should not be over-interpreted.) The differences in CPU time between JMLE and MMLE-EM, however, were quite impressive: JMLE required between 22 and 34 s per data set, as opposed to MMLE-EM that used at least about three and a half hours ($J = 30$, without explicitly modeling the distribution of $\boldsymbol{\alpha}$) and more than 37 h if $J = 60$ (with explicitly modeling the distribution of $\boldsymbol{\alpha}$).

7. Practical Application: Analysis of Language Testing Data

As a real-world application of JMLE , data from a retired version of the Examination for the Certificate of Proficiency in English (ECPE) were fitted with the LCDM. The ECPE is a test of advanced English language proficiency, developed, administered, and scored by Cambridge Michigan Language Assessments (CaMLA). The test is given annually around the world to approximately 41,000 non-native speakers of English (ECPE 2013 Report, 2013). The items used in this study are a subset from the grammar section of the ECPE. They have been previously analyzed by Buck and Tatsuoka (1998), Feng, Habing, and Huebner (2014), Henson and Templin (2007), Liu, Douglas, and Henson (2009), Templin and Hoffman (2013), and Templin and Bradshaw (2014). Responses to $J = 28$ items were collected from $N = 2922$ examinees. The test involved $K = 3$ attributes (α_1 = lexical skills; α_2 = morphosyntactic skills; and α_3 = cohesive skills); the individual items required at most the mastery of $K = 2$ attributes (the complete Q-matrix is given in Templin & Hoffman, 2013).

TABLE 2.
Simulation Study II: CCR and RMSE of JMLE, `Mplus` MMLE-EM, and CMLE plus CPU times of JMLE and MMLE-EM when the data conformed to the LCDM; $N = 3000$, $K = 5$.

CCR

J	JMLE	MMLE-EM	MMLE-EM-HO	CMLE
30	0.679	0.676	0.681	0.691
40	0.763	0.763	0.765	0.767
50	0.849	0.851	0.853	0.860
60	0.883	NA	0.882	0.885

RMSE

J	JMLE	MMLE-EM	MMLE-EM-HO	CMLE
30	0.294	0.242	0.239	0.191
40	0.243	0.227	0.225	0.170
50	0.169	0.151	0.152	0.149
60	0.231	NA	0.199	0.194

CPU

J	JMLE	MMLE-EM	MMLE-EM-HO
30	00:00:34	03:35:07	07:23:01
40	00:00:42	07:44:06	06:32:31
50	00:00:22	06:33:58	05:59:14
60	00:00:30	NA	37:48:29

Columns MMLE-EM/MMLE-EM-HO contain the results for fitting the data without/with explicitly modeling the distribution of $\boldsymbol{\alpha}$.

7.1. Results

The ECPE data were recently analyzed with the LCDM by Templin and Hoffman (2013). (They used MMLE-EM relying on the implementation of the EM algorithm in `Mplus`; Muthén & Muthén, 1998–2012.) The Templin-Hoffman findings (see also Templin & Bradshaw, 2014) are presented here for comparison with the JMLE results. Table 3 reports the item parameter estimates obtained from JMLE and MMLE-EM.

The estimates obtained by JMLE and MMLE-EM were quite different. Overall, MMLE-EM yielded higher guessing probabilities than JMLE (i.e., $\hat{\beta}_0$; in general, values of $\hat{\beta}_0 > 0$ correspond to a guessing probability higher than 0.5) and lower coefficient estimates for the interaction terms than JMLE. In an attempt to resolve the discrepancies between MMLE-EM and JMLE, the ECPE data were also fitted using MCMC. However, the MCMC results (not reported here), in turn, were different from the MMLE-EM and JMLE results.

To further probe these discrepant findings, the estimated attribute profiles (i.e., proficiency class membership) of a subset of five examinees were inspected. These five examinees had already been used by Templin and Hoffman (2013, p. 47) as exemplary cases for a deeper analysis of their findings. Table 4 presents the observed item response profiles; the estimated attribute profiles of these examinees, as they were obtained from JMLE and MMLE-EM are reported in Table 4 (a). Note that the two methods produced identical attribute profile estimates for examinees 1 and 29. For the other three examinees, the estimates disagree: JMLE resulted in higher probabilities of attribute mastery than MMLE-EM.

As a descriptive measure for evaluating the attribute profile estimates, for each of the five examinees, the proportion of correct answers to all items requiring mastery of the kth attribute was computed for $k = 1, 2, 3$. For example, examinees 1 and 33 both had an 83% chance (i.e.,

TABLE 3.
Estimates of the item parameters obtained by JMLE and MMLE-EM (Templin & Hoffman, 2013).

Item	JMLE					MMLE-EM				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_{kk'}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_{kk'}$
1	0.47	−0.75	−0.05	–	2.47	0.84	0.00	0.60	–	1.22
2	−0.54	–	2.49	–	–	1.04	–	1.25	–	–
3	−0.68	0.48	–	−0.03	0.97	−0.34	0.75	–	0.35	0.54
4	−0.97	–	–	2.05	–	−0.14	–	–	1.69	–
5	0.18	–	–	2.28	–	1.08	–	–	2.02	–
6	0.08	–	–	1.97	–	0.87	–	–	1.69	–
7	−0.65	0.12	–	0.42	1.85	−0.11	2.86	–	0.95	−0.95
8	0.08	–	2.58	–	–	1.48	–	1.92	–	–
9	−0.76	–	–	1.81	–	0.12	–	–	1.20	–
10	−0.67	1.86	–	–	–	0.06	2.05	–	–	–
11	−0.36	0.09	–	0.20	1.69	−0.04	0.82	–	0.96	0.78
12	−2.23	0.29	–	0.31	1.89	−1.77	0.00	–	1.29	1.52
13	−0.02	1.70	–	–	–	0.66	1.63	–	–	–
14	−0.41	1.44	–	–	–	0.18	1.37	–	–	–
15	0.07	–	–	2.34	–	1.00	–	–	2.11	–
16	−0.55	0.15	–	0.38	1.52	−0.11	2.34	–	0.89	−0.87
17	0.62	–	0.26	−0.27	1.94	1.35	–	0.77	0.60	0.08
18	−0.05	–	–	2.05	–	0.93	–	–	1.39	–
19	−1.25	–	–	2.40	–	−0.20	–	–	1.85	–
20	−2.07	0.13	–	0.38	1.94	−1.39	0.24	–	0.91	1.41
21	−0.42	0.31	–	0.64	1.26	0.16	1.05	–	1.13	0.04
22	−2.04	–	–	2.80	–	−0.87	–	–	2.25	–
23	−0.63	–	2.41	–	–	0.67	–	2.07	–	–
24	−2.22	–	2.53	–	–	−0.67	–	1.52	–	–
25	−0.47	1.31	–	–	–	0.09	1.14	–	–	–
26	−0.43	–	–	1.44	–	0.16	–	–	1.12	–
27	−1.64	1.81	–	–	–	−0.89	1.71	–	–	–
28	−0.54	–	–	2.39	–	0.57	–	–	1.75	–

5 out of 6) to answer items requiring attribute 2 correctly. The attribute estimates obtained from MMLE-EM indicated that examinee 1 did master attribute 2, but examinee 33 did not. Closer inspection, however, showed that examinees 1 and 33 had identical responses to all items that required attribute 2. Said differently, it is presumably hard to explain to parents and students—independent of what particular DCM was used—why a student—here: examinee 33—was flagged as failing attribute 2 if, in fact, he or she got 5 out of 6 items correct that required mastery of attribute 2.

Templin and Bradshaw (2014) re-analyzed the ECPE data, but used, different from Templin and Hoffman (2013), the Hierarchical Diagnostic Classification Model (HDCM), a variant of the LCDM modified to accommodate a hierarchical structure supposed to underlie the attributes. A hierarchy implies an order among attributes such that mastery of certain attributes requires mastery of other attributes as a prerequisite. If the attributes have a hierarchical structure, then, typically, several proficiency classes are empty because the corresponding combination of attributes cannot occur, given the specific attribute hierarchy. Templin and Bradshaw (2014) postulated a linear hierarchy among attributes (i.e., the attributes are ordered along a line, with precedence implying hierarchy; see also Leighton, Gierl, & Hunka, 2004) such that α_3: lexical rules $\preccurlyeq$ α_2: cohesive

TABLE 4.

Analysis of the ECPE data: Observed item response profiles and estimated attribute profiles of examinees 1, 10, 14, 29, and 33 obtained from JMLE and MMLE-EM.

ID	Data
1	1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1
10	1 1 1 1 0 0 1 1 1 1 1 0 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1
14	0 1 1 0 1 1 1 1 1 1 1 0 0 0 1 0 1 1 0 1 1 1 0 0 0 0 0 1
29	0 0 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1 1 1
33	1 1 1 0 0 1 1 1 1 1 1 0 0 1 1 0 1 0 0 1 1 0 1 0 1 0 1 1

(a): Results from Templin and Hoffman (2013)

	JMLE							MMLE-EM			
	Proportion Correct			Probability of Mastery				Probability of Mastery			
ID	α_1	α_2	α_3	α_1	α_2	α_3	$\hat{\alpha}$	α_1	α_2	α_3	$\hat{\alpha}$
1	1.00	**0.83**	0.94	1.00	1.00	1.00	(111)	1.00	**0.96**	1.00	**(111)**
10	0.77	0.67	0.67	0.99	0.96	0.99	(111)	0.62	0.15	0.10	(100)
14	0.46	0.50	0.72	0.40	0.78	1.00	(011)	0.00	0.05	0.54	(001)
29	0.77	0.33	0.83	1.00	0.18	1.00	(101)	0.56	0.08	0.95	(101)
33	0.77	**0.83**	0.56	0.99	0.99	0.71	(111)	0.21	**0.18**	0.02	**(000)**

(b): Results from Templin and Bradshaw (2014)

	JMLE							MMLE-EM			
	Proportion Correct			Probability of Mastery				Probability of Mastery			
ID	α_1	α_2	α_3	α_1	α_2	α_3	$\hat{\alpha}$	α_1	α_2	α_3	$\hat{\alpha}$
1	1.00	**0.83**	0.94	1.00	1.00	1.00	(111)	1.00	**1.00**	1.00	**(111)**
10	0.77	0.67	0.67	0.99	0.96	0.99	(111)	0.10	0.12	0.20	(000)
14	0.46	0.50	0.72	0.40	0.78	1.00	(011)	0.00	0.06	0.56	(001)
29	0.77	0.33	0.83	1.00	0.18	1.00	(101)	0.17	0.20	0.96	(001)
33	0.77	**0.83**	0.56	0.99	0.99	0.71	(111)	0.02	**0.03**	0.03	**(000)**

rules $\preccurlyeq \alpha_1$: morphosyntactic rules, where $a \preccurlyeq b$ denotes that a precedes (is a prerequisite for) b. This hierarchy implies that of the eight proficiency classes from the previous analysis, only four are meaningfully defined: $\boldsymbol{\alpha}_1 = (000)$, $\boldsymbol{\alpha}_2 = (001)$, $\boldsymbol{\alpha}_3 = (011)$, and $\boldsymbol{\alpha}_4 = (111)$. The interesting question, of course, is whether this important modification of the LCDM might have resolved the inconsistencies described earlier for the subset of examinees. Hence, the probabilities of attribute mastery were re-computed based on the new item parameter estimates obtained for the HDCM by Templin and Bradshaw (2014, Table 5) and the estimates of the "structural parameters" reported in Figure 1. The revised probabilities of attribute mastery are reported in Table 4(b). Unfortunately, using the HDCM did not resolve the issues discussed earlier. In summary, further research seems warranted to determine, which method might be most appropriate for analyzing the ECPE data.

8. Discussion

In this article, JMLE was developed for fitting DCMs—that is, for estimating the item parameters and examinees' proficiency class membership. JMLE has been barely used in Psychometrics

because JMLE-parameter estimators typically lack statistical consistency. The JMLE procedure presented here resolves the consistency issue by incorporating an external, statistically consistent estimator of examinees' proficiency class membership into the joint likelihood function, which subsequently allows for the construction of item parameter estimators that also have the consistency property. This claim was developed and proven using the framework of general DCMs and the LCDM in particular. Two consistency theorems established (a) pointwise convergence in probability (Theorem 1) and (b) uniform convergence (Theorem 2) of the JMLE item parameter estimators to the true item parameters. Two simulation studies were conducted for evaluating the performance of JMLE when used with tests of varying length and numbers of attributes. The results showed that the JMLE-based item parameter estimates and examinee classification were essentially as accurate as those obtained from MMLE using the EM algorithm. However, the computational efficiency of JMLE outperformed that of Mplus MMLE-EM—occasionally, reducing the CPU times by a factor of 1000 or even larger. In light of these results, three questions remain to be addressed.

First, $\tilde{\boldsymbol{\alpha}}$ provided by the NPC method was used as (external) statistically consistent estimator of examinees' attribute profiles $\boldsymbol{\alpha}$ (i.e., their proficiency class memberships) for initializing JMLE in the computational experiments reported here. At present, NPC appears to be the only method available for obtaining $\tilde{\boldsymbol{\alpha}}$. But what are the specific conditions under which $\tilde{\boldsymbol{\alpha}}$ from NPC is guaranteed to be a statistically consistent estimator of $\boldsymbol{\alpha}$? Wang and Douglas (2015) proved that the NPC method guarantees the statistical consistency of $\tilde{\boldsymbol{\alpha}}$ for any CDM provided the probability of a correct response is greater than 0.5 for examinees who master the required attributes and less than 0.5 for examinees who do not master all the required attributes.

Second, to increase the numerical accuracy of the estimates, the JMLE estimation procedure can be iterated (an option that was used in the simulations reported earlier). The item parameter estimates obtained initially are used in a second stage for re-estimating examinees' attribute profiles by maximizing $\ln L\big(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_N; \mathbf{Y}, (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \ldots, \tilde{\boldsymbol{\beta}}_J)\big)$. The updated examinee attribute profiles $\tilde{\boldsymbol{\alpha}}$ are then used as input for re-estimating the item parameter estimates $\tilde{\boldsymbol{\beta}}_j$, and so on. These steps can be repeated until the estimates do not change much. The convergence of the estimation can be monitored by the relative likelihood change

$$\frac{\ln L^t - \ln L^{t-1}}{\ln L^{t-1}},$$

where t and $t-1$ refer to consecutive iterations. For the estimates of examinee's attribute profiles, a viable criterion is

$$\frac{1}{N} \sum_{i=1}^{N} I\big[\tilde{\boldsymbol{\alpha}}_i^{(t)} = \tilde{\boldsymbol{\alpha}}_i^{(t-1)}\big] \geq 0.99$$

($I[\cdot]$ denotes the indicator function). For the item parameter estimates, consider (in generic θ-notation)

$$\max_{j=1}^{J} \left\{ \big|\tilde{\boldsymbol{\theta}}_j^{(t)} - \tilde{\boldsymbol{\theta}}_j^{(t-1)}\big| \right\} \leq 0.001.$$

One should note that iterations are not required to guarantee consistency of the JMLE estimators; the sole purpose of iterating the algorithm is to improve the numerical accuracy of the estimates. An interesting aside within this context is the question whether the consistency property of the

parameter estimators is preserved during iterations. Theorem 4.2 in Junker (1991) suggests that this is the case.

Third, what are the current computational options for educational researchers and practitioners who wish to use the LCDM in their testing programs and empirical research? Of course, writing code from scratch is always an option. If a researcher wants to use MCMC then he or she can refer, for example, to `OpenBUGS` (Lunn, Spiegelhalter, Thomas, & Best, 2009). Alternatively, a user can opt for a commercial package that offers an implementation of the EM algorithm for fitting (constrained) latent class models, for example, `Latent GOLD` (Vermunt & Magidson, 2000) and `Mplus` (Muthen & Muthen, 1998–2012). (For details on how to use `Mplus` for fitting the LCDM, consult the tutorial by Templin & Hoffman, 2013.) In light of the considerable amounts of CPU time often encountered with MCMC and MMLE-EM, researchers and educational practitioners might indeed consider JMLE as a viable computational alternative for estimating the item parameters of DCMs and examinees' attribute profiles.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

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.

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*, 119–157.

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., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika*, *74*, 633–665.

Chiu, C.-Y., & Köhn, H.-F. (2015). Consistency of cluster analysis for cognitive diagnosis: The DINO model and the DINA model revisited. *Applied Psychological Measurement*, *39*, 465–479.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Chiu, C.-Y. (2010, April). *A general empirical method of Q-matrix validation.* Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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*. Psychometrics (Vol. 26, pp. 979–1030). Amsterdam: Elsevier.

ECPE 2013 report. (2013). Retrieved March 10, 2013, from http://www.cambridgemichigan.org/resources/ecpe/reports

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

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*. Psychometrics (Vol. 26, pp. 1031–1038). Amsterdam: Elsevier.

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.

Henson, R. A., & Templin, J. (2007, April). *Large-scale language assessment using cognitive diagnosis models*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13–30.

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.

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

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*, 205–236.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement*, *33*, 579–598.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *33*, 379–416.

Muthén, L. K., & Muthén, B. O. (1998–2012). Mplus *user's guide* (7th edn.) Los Angeles: Muthén & Muthén.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). *CDM: Cognitive diagnosis modeling. R package version 3.1-14*. Retrieved from the Comprehensive R Archive Network [CRAN] website

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

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55–73.

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

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 model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*, 37–50.

Vermunt, J. K., & Magidson, J. (2000). Latent GOLD *users's guide*. Boston: Statistical Innovations Inc.

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.

Wang, S., & Douglas, J. (2015). Consistency of nonparametric classification in cognitive diagnosis. *Psychometrika*, *80*, 85–100.

Zheng, Y., & Chiu, C.-Y. (2014). NPCD: Nonparametric methods for cognitive diagnosis. *R package version 1.0-5*. http://CRAN.R-project.org/package=NPCD