Using the iterative latent-class analysis approach to improve attribute accuracy in diagnostic classification models

Zhehan Jiang¹

Published online: 31 January 2019
© The Psychonomic Society, Inc. 2019

Abstract

The focus of diagnostic classification models is on investigating a respondent's mastery status of the attributes required for completing tasks and/or solving problems. Recent advances in model development have produced saturated model variants such as the log-linear cognitive diagnostic model (LCDM), but works focusing on improving the accuracy of their attribute estimates have not been accomplished commensurably. This article proposes an iterative latent class analysis (ILCA) approach to estimating attributes, such that the accuracy can be higher than that of traditional approaches. Particularly, the needs for the ILCA approach are illustrated within a literature review, the detailed procedures of the ILCA are presented via both pseudo-codes and verbal explanations, a simulation study is conducted to demonstrate the estimation accuracy, and finally, a discussion containing limitations and future research directions is provided. The results of this article show that ILCA outperforms its competitors in many conditions. Thus, it can be used to produce score reports.

Keywords Latent class analysis · Cognitive diagnostic model · EM Algorithm · Classification accuracy · Estimation

In social, behavioral, and educational assessments, it is common for researchers and practitioners to adopt measurement tools in order to collect data from respondents sampled from the population of interest. The measurement tools can be in various forms—for example, standardized tests used in academic aptitude measurement, surveys deployed in consumer satisfaction research, or self-rating scales utilized in psychological symptom diagnosis. Among all measurement tools, instruments that comprise multiple dichotomous items are most prevalent, due to this method's long history of use and convention (Crocker & Algina, 1986; Cronbach, 1951). A dichotomous item contains binary responses, such as true–false answers, present–absent symptoms, or endorsed–nonendorsed attitudes. Each item should be carefully created to measure one or multiple latent (unobservable) variables of interest based on particular theories. To illustrate, addition, subtraction, multiplication, and division are four latent variables defined in math ability assessment, where test items such as “ $2 + 4 - 1$ ” measure the first two latent variables, and items such as “ $4 \times 2/3$ ” measure the last two. Table 1 displays a general form, called a *Q-matrix*, which aligns the items with

latent variables in accordance with a theoretical framework or content experts' judgment (it is also called an *expert-matrix*). To obtain latent variable information at the respondent level, granted that a *Q-matrix* has been scientifically verified, one usually chooses statistical modeling strategies such as classical test theory (CTT) and item response theory (IRT; Lord, 1980). Modeling strategies such as CTT and IRT treat latent variables as continua, which are conventionally assumed to follow a (multivariate) normal distribution. Recent advances in analytic methods are collectively referred to as *diagnostic classification models* (DCMs; Rupp & Templin, 2008) or *cognitive diagnostic models* (CDMs; DiBello & Stout, 2007). These assume that the latent variables are also defined on the binary scale; as compared with CTT and IRT, DCMs have higher reliability and can offer rich diagnostic information to aid decision making. Particularly, DCM reliability is uniquely computed as a function of the tetrachoric correlation for the replication contingency table, made of posterior attribute mastery probability estimates (see Templin & Bradshaw, 2014, for computing details). This is derived according to the reliability definition designed for categorical latent traits; the higher reliability yielded by DCMs is a consequence of the smaller range of latent-variable values that examinee estimates can take in DCMs (Liu, Qian, Luo, & Woo, 2017; Rupp & Templin, 2008). In addition, DCMs possess high interpretability for a latent-variable structure that can be used to support or

✉ Zhehan Jiang
zjiang17@ua.edu

¹ The University of Alabama, Tuscaloosa, AL, USA

Table 1 A Q-matrix sample

	Skill 1	Skill 2	Skill 3
Question 1	1	0	0
Question 2	1	0	1
Question 3	0	1	1
Question 4	1	1	0

verify a certain theory. To date, researchers across different fields have adopted DCMs because of their aforementioned advantages. For example, Seixas, Zadrozny, Laks, Conci, and Saade (2014) used DCMs to study dementia and Alzheimer's disease; Dakanalis, Timko, Clerici, Zanetti, and Riva (2014) implemented DCMs to investigate eating disorders; and Clark and Wells (1995) modeled social phobia via a DCM framework. In sum, by analyzing responses via a DCM framework, one can not only obtain information about the latent-variable profile for each respondent, but also about the features of the assessment items. To differentiate from CTT and IRT frameworks, the latent variables on a binary scale are called "attributes" throughout this article.

Although DCMs provide an innovative and useful angle to analyze response data, the attribute estimates can be questionable in four situations. The first situation in which the attribute estimates aren't sufficiently accurate is model misspecification. For instance, when responses were generated from a saturated model¹ and/or a model with a hierarchical attribute structure,² a simpler and/or a misspecified model can yield inappropriate estimates (Liu, 2017; Templin & Bradshaw, 2014). The second occurrence is the inaccurate setting of the Q-matrix. The Q-matrix has strong influences on both the item parameters and the attribute estimates (Rupp & Templin, 2008), and therefore misspecified Q-matrices are detrimental, since the modeling procedures then face the risk of beginning from an incorrect base. The third scenario takes places when the sample sizes are small. Chiu and Douglas (2013) proved that when the sample size is small, traditional estimation approaches do not assign attributes accurately. The fourth occasion, which is the focus of this article, is the inadequacy of the estimation approaches per se. The sample size problem addressed in the third scenario can be regarded as a special case of the present issue; more broadly defined, estimation problems are related to algorithmic robustness. In other words, questions related to estimations that only perform well under limited conditions fall into this category. To summarize, this problem is introduced by the limit of the traditional approaches used in estimation, and therefore, to remedy it,

this article proposes an innovative estimation approach called *iterative latent-class analysis*. The idea is to construct a more robust attribute estimator such that, in many conditions, it can produce more accurate results than the traditional approaches do.

Log-linear cognitive diagnostic models

Within the DCM literature, assumptions about how skills (or cognitive/psychological processes) affect test performance are executed through the selection of a specific DCM. Depending on the specific DCM chosen, the modeling family can be defined as either the noncompensatory models or the compensatory models (Henson, Templin, & Willse, 2009). Within the noncompensatory genre, the models assume that one cannot "make up" for nonmastery of some attributes by mastery of others. On the other hand, compensatory models allow an individual to "make up" for what is lacking in one skill by having mastered another. The assumptions for either compensatory or noncompensatory models can be hard to test. Log-linear CDM (LCDM; Henson et al., 2009) or similar advanced models, however, can avoid such strong assumptions. To introduce LCDM, this section starts with a general description of a DCM.

Given that a DCM is essentially a constrained latent-class model (Day, 1969; Titterton, Smith, & Makov, 1985; Wolfe, 1970), mathematically it can be defined as

$$P(Y_p = y_p) = \sum_{c=1}^C \left(v_c \prod_{i=1}^I \pi_{ci}^{y_{pi}} (1 - \pi_{ci})^{1-y_{pi}} \right), \quad (1)$$

where $y_p = (y_{p1}, y_{p2}, \dots, y_{pI})$ is the binary response vector of person p on a test comprising I items, and element y_{pi} is the response on item i . The variable v_c is the probability of membership in latent class c , and π_{pi} is the probability of a correct response to item i by person p in the class. Converted from Eq. 1, the log-likelihood function for the model of N persons can be expressed as

$$L = \sum_{p=1}^N \log \left\{ \sum_{c=1}^C \left(v_c \prod_{i=1}^I \pi_{ci}^{y_{pi}} (1 - \pi_{ci})^{1-y_{pi}} \right) \right\}. \quad (2)$$

Furthermore, Eq. 2 can be converted to

$$L = \sum_{p=1}^N \log \left\{ \sum_{c=1}^C \left(\exp \left(\log(v_c) + \log \left[\prod_{i=1}^I \pi_{ci}^{y_{pi}} (1 - \pi_{ci})^{1-y_{pi}} \right] \right) \right) \right\} \quad (3)$$

¹ A *saturated* model means the attributes associated with an item all function uniquely, whereas *nonsaturated* models can assume that the expected outcome of mastering one attribute is equal to those from mastering more than one.

² Similar to well-known second-level latent models, DCMs can have hierarchical structures, such that multiple levels of attributes can be modeled.

where $\log \left[\prod_{i=1}^I \pi_{ic}^{y_{pi}} (1 - \pi_{ic})^{1 - y_{pi}} \right]$ is replaced with $\sum_{i=1}^I \log \left[\pi_{ic}^{y_{pi}} (1 - \pi_{ic})^{1 - y_{pi}} \right]$, due to the mathematical properties of the log operation.

Assuming that the number of attributes is D , the mastery status of the attributes for a random person is denoted by $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_D)$, where the elements in α are expressed as binary values (i.e., the a th attribute, α_a , is either 1 or 0). In total, there are 2^D possible attribute profiles (i.e., classes). To illustrate, a person p with a profile $\alpha = (1, 1, 1, 0)$ has mastered all but the last of four dimensions. Recent advances in modeling development have produced general diagnostic model variants—for example, LCDM and, equivalently, the generalized deterministic input, noisy “and” gate model (G-DINA; de la Torre, 2011). Throughout this article, the general diagnostic model is referred to as an LCDM, for the sake of consistency. An LCDM provides great flexibility by subsuming most latent variables, enabling both additive and nonadditive relationships between skills and questions simultaneously, as well as combining with other psychometric models to allow even greater insight. The work of Rupp, Templin, and Henson (2010, p. 163) showed that LCDMs can be converted into core DCMs such as the deterministic input, noisy “and” gate model (DINA; Junker & Sijtsma, 2001); the noisy input, deterministic “and” gate model (NIDA; Junker & Sijtsma, 2001); and the reparameterized unified model (RUM; Hartz, 2002), whereas examples of disjunctive models include the deterministic input, noisy “or” gate model (DINO; Templin & Henson, 2006). Given the aforementioned advantages, the LCDM is introduced here and will be used in the following sections.

As is illustrated in Table 1, a Q-matrix of size $I \times D$ is necessary for an LCDM, where, again, I and D are the numbers of items and attributes, respectively. The (i, a) element of the Q-matrix q_{ia} is 1 when item i measures attribute a , and 0 otherwise. Given that person p 's attribute profile is α_c , the conditional probability of item i can be stated as

$$\pi_{ic} = P(y_{ip} = 1 | \alpha_c) = \frac{\exp(\lambda_{i,0} + \lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i))}{1 + \exp(\lambda_{i,0} + \lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i))}, \quad (4)$$

where $\mathbf{q}_i$ is the set of Q-matrix entries for item i , $\lambda_{i,0}$ is the intercept parameter, λ_i represents a vector of size $(2^D - 1) \times 1$ that contains the main effect and interaction effect parameters for item i , and $\mathbf{h}(\alpha_c, \mathbf{q}_i)$ is a vector of size $(2^D - 1)$ with linear combinations of α_c and $\mathbf{q}_i$. Particularly, $\lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i)$ can be expanded to

$$\begin{aligned} \lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i) = & \sum_{a=1}^D \lambda_{i,1,(a)} \alpha_{ca} q_{ia} \\ & + \sum_{a=1}^D \sum_{a' > 1}^D \lambda_{i,2,(a,a')} \alpha_{ca} \alpha_{ca'} q_{ia} q_{ia'} + \dots \end{aligned} \quad (5)$$

where $\lambda_{i,1,(a)}$ and $\lambda_{i,2,(a,a')}$ are, respectively, the main effect for the a th attribute α_a and a two-way interaction effect for α_a and $\alpha_{a'}$. Since the elements of α_c and $\mathbf{q}_i$ are binary, $\mathbf{h}(\alpha_c, \mathbf{q}_i)$ contains binary elements that indicate effects that are the estimates of interest. For an item measuring n attributes, n -way interaction effects should be specified in $\mathbf{h}(\alpha_c, \mathbf{q}_i)$. Table 2 provides a sample of a measure with three attributes: the first item, measuring one attribute only (i.e., α_1), has two estimates, whereas the third item, which is associated with all of the given attributes, contains eight estimates.

In practice, an estimating LCDM refers to the expectation maximization (EM) algorithm (Bock & Aitkin, 1981) that maximizes the marginal likelihood; this is the most commonly seen algorithm in the DCM literature. Similarly, Jiang and Ma (2018) deployed the differential evaluation optimizer for LCDMs in the EM framework. In addition to the EM algorithm, Markov chain Monte Carlo (MCMC) techniques can be used to estimate the LCDM, but to date this algorithm has been applied only to simpler DCMs, such as the DINA model (da Silva, de Oliveira, Davier, & Bazán, 2017). Given that their updating procedures are based on random sampling, MCMC approaches tend to be slower than the EM algorithm (Patz & Junker, 1999). This study focuses on the EM algorithm, due to its practicality and popularity. Traditionally, a researcher specifies an LCDM within a likelihood function framework, as Eqs. 1 to 5 demonstrate. For instance, de la Torre (2011) described a marginalized maximum-likelihood estimation for estimating a GDINA model with the EM framework (which can be converted to an LCDM), and Templin and Hoffman (2013) estimated LCDMs using *Mplus* (L. K.

Table 2 A three-item sample of expressions for a log-linear cognitive diagnosis model

Item	α_1	α_2	α_3	Expanded $\lambda_{i,0} + \lambda_i^T \mathbf{h}(\alpha_c, \mathbf{q}_i)$	Shortened Expression
1	1	0	0	$\lambda_{1,0} + \lambda_{1,1}(1) + \lambda_{1,2}(0) + \lambda_{1,3}(0) + \lambda_{1,12}(1 * 0) + \lambda_{1,13}(1 * 0) + \lambda_{1,23}(0 * 0) + \lambda_{1,123}(1 * 0 * 0)$	$\lambda_{1,0} + \lambda_{1,1}(1)$
2	0	1	1	$\lambda_{2,0} + \lambda_{2,1}(0) + \lambda_{2,2}(1) + \lambda_{2,3}(1) + \lambda_{2,12}(0 * 1) + \lambda_{2,13}(0 * 1) + \lambda_{2,23}(1 * 1) + \lambda_{2,123}(0 * 1 * 1)$	$\lambda_{2,0} + \lambda_{2,2}(1) + \lambda_{2,3}(1) + \lambda_{2,23}(1)$
3	1	1	1	$\lambda_{3,0} + \lambda_{3,1}(1) + \lambda_{3,2}(1) + \lambda_{3,3}(1) + \lambda_{3,12}(1 * 1) + \lambda_{3,13}(1 * 1) + \lambda_{3,23}(1 * 1) + \lambda_{3,123}(1 * 1 * 1)$	$\lambda_{3,0} + \lambda_{3,1}(1) + \lambda_{3,2}(1) + \lambda_{3,3}(1) + \lambda_{3,12}(1) + \lambda_{3,13}(1) + \lambda_{3,23}(1) + \lambda_{3,123}(1)$

Muthén & Muthén, 2013), which deploys an accelerated EM algorithm. These two *likelihood-based approaches* offer estimates of both items and respondents' attributes, as well as statistical inferences, and therefore provide strong interpretability that aligns item features and respondent performance in a common scale (i.e., one can find the impacts of attribute changes on the item responses). Alternatively, Jiang and Carter (2018a) have described a Bayesian approach to handle the estimation that yields results similar to those yielded by EM algorithms. However, Chiu and Douglas (2013) proved that when the sample size is small, assigning attributes correctly becomes difficult for likelihood-based approaches; instead, they proposed the nonparametric classification (NPC) method, which was further extended to a general nonparametric classification (GNPC) method by Chiu, Sun, and Bian (2018) in order to handle LCDMs, such that attribute accuracy can be improved when the sample size is small. NPC and GNPC, however, remain less known to date, such that these utilities need further examinations. In this article, only the EM likelihood-based approaches are considered, due to the prevalence of these approaches (see Li, Hunter, & Lei, 2016; Ravand, 2016; Svetina, Dai, & Wang, 2017; and Templin & Bradshaw, 2014, for examples). This article introduces a new EM-based approach for estimating the attributes with a higher accuracy level. The following section depicts the proposed approach.

The iterative latent-class analysis (ILCA) approach

As the name suggests, ILCA is based on sequential latent-class modeling estimations that update attributes in order repeatedly. To define the algorithm, $\mathbf{A}$ is the matrix recording all respondents' attributes; for example, in a simplified two-attribute model with two respondents only, $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ means that the first respondent masters only the second attribute, whereas the second respondent masters both attributes. Let $\mathbf{Y}^a$ be the response matrix comprising items measuring attribute a , and $\mathbf{A}^{-a}$ be the attribute matrix without an a th column. To estimate respondents' a th attributes, this method begins by joining $\mathbf{Y}^a$ and $\mathbf{A}^{-a}$, to form a new response matrix $\mathbf{X}^a$ (i.e., $\mathbf{X}^a = [\mathbf{Y}^a | \mathbf{A}^{-a}]$); that said, given that $\mathbf{Y}^a = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ and $\mathbf{A}^{-a} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, then $\mathbf{X}^a = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$. $\mathbf{X}^a$ is then fed to a latent-class analysis (LCA), which essentially performs unsupervised clustering. The LCA is conducted via the EM algorithm, in which the log-likelihood function is identical to Eq. 2, except that the response matrix is $\mathbf{X}^a$ instead of $\mathbf{Y}$. The derivation details about implementing the EM algorithm in the

LCA can be found in Bartholomew, Knott, and Moustaki (2011) and B. Muthén and Shedden (1999). The summarized steps are provided below.

Assume that all $\hat{\pi}$ s were estimated at the last iteration, $t-1$, and therefore are treated as known parameters, the probability for a respondent $H(C=c | \mathbf{Y}_p = \mathbf{y}_p)$ can be obtained via

$$\frac{v_c^{t-1} \prod_{i=1}^I \pi_{ip}^{y_{ip}} (1-\pi_{ip})^{1-y_{ip}}}{\sum_{c=1}^C v_c^{t-1} \prod_{i=1}^I \pi_{ip}^{y_{ip}} (1-\pi_{ip})^{1-y_{ip}}}, \text{ which leads to a new } \hat{v}_c^t \text{ calculated with}$$

the expression $\frac{\sum_{p=1}^N H(C=c | \mathbf{Y}_p = \mathbf{y}_p)}{N}$. Note that π_{ip} is essentially π_{ic} , since the respondent p is addressed at the condition of the class c . Following from the idea that the estimated parameters from the previous iteration are treated as known, $\hat{\pi}_{ic}$ for the respon-

dent p can be obtained from $\frac{\sum_{p=1}^N y_{ip} H(C=c | \mathbf{Y}_p = \mathbf{y}_p)}{\sum_{p=1}^N H(C=c | \mathbf{Y}_p = \mathbf{y}_p)}$, where $\mathbf{y}_p$ is the

complete responses of respondent p . This step is essentially aggregating the posterior class estimates of the respondents to obtain the class probabilities of the structure model at the current iteration; at this point, an iteration is completed. This intertwined process runs until a certain convergence criterion is met—for example, when a change of the log-likelihood value is lower than .001.

For each attribute vector, ILCA deploys the aforementioned EM algorithm, such that respondents are classified into two classes (i.e., master and nonmaster). Once an LCA is achieved for the a th attribute, the clustering results of attribute a are taken into the sequential LCA for different attributes. This iterative LCA stops either when the estimated attribute matrix $\hat{\mathbf{A}}$ doesn't change from the previous iteration or the iteration number reaches a predefined value. Like other mixture models, the LCA estimation faces the risk of local maxima (Jin, Zhang, Balakrishnan, Wainwright, & Jordan, 2016). To reduce the problem, multiple sets of initial values can be fed to the ILCA, and the set of the results with the largest log-likelihood value can be used as the final estimates. Let $ITER$ be the predefined iteration number; the pseudo-codes are outlined in Table 3,³ where $G(+)$ and $G(-)$ are symbols used to represent two groups/clusters estimated via the LCA. Note that Steps 3 and 4 in the inner *for* loop are implemented to eliminate the label-switching problem. That is, theoretically, the group mastering required attributes should perform better than the other group. The steps listed in the pseudo-code blocks are straightforward: The LCA iterates by using the results from the last iteration to treat both the estimated attributes and the associated items as (known) predictors.

Why are $\mathbf{Y}^a$ and $\mathbf{A}^{-a}$ used to estimate the a th attributes via the LCA? The regular LCA for the a th attribute estimation is based upon $\mathbf{Y}^a$ by its natural definition. In addition, the

³ The code snippet can be found <https://alabama.box.com/s/h78fzxn48eow4swlfzldvsmliwy8sdg>.

Table 3 The iterative latent class analysis approach pseudo-codes

1. Randomly initialize m sets of attribute matrices $\mathbf{A}_s$.
2. For each initial matrix $\mathbf{A}$ do:
For $iter = 1, 2, \dots, ITER$:
For $a = 1, 2, \dots, D$ do:
(1). Form $\mathbf{X}^a$ by joining matrices $\mathbf{Y}^a$ and $\mathbf{A}^{-a}$.
(2). Run the LCA on $\mathbf{X}^a$ to form two groups $G(+)$ and $G(-)$.
(3). Sum $\mathbf{Y}^a$ across items for both $G(+)$ and $G(-)$.
(4). If the mean of sum scores for $G(+)$ is larger than $G(-)$:
Assign 1 to the attribute a for $G(+)$ and 0 for $G(-)$.
Else:
Assign 0 to the attribute a for $G(+)$ and 1 for $G(-)$.
(5). Form a new $\mathbf{A}$ by inserting the new estimates to the a th column of the previous $\mathbf{A}$.
(6). If the new $\mathbf{A} =$ the previous $\mathbf{A}$:
Assign $ITER$ to $iter$.
Assign $iter+1$ to $iter$.
3. Select $\mathbf{A}$ from m sets of results with the highest log-likelihood value.

columns of $\mathbf{A}^{-a}$ (i.e., the other attributes) are presumably informative as to the a th attributes to some degree because, theoretically, the attributes should be correlated. $\mathbf{Y}^a$, on the other hand, may not contain much information useful to the estimation because it doesn't directly associate with the a th attribute, such that it has a high risk of introducing noise. To examine the utility of the ILCA approach and compare it with the likelihood-based approach, a simulation study is described in the coming section.

Simulation study

Simulation design

Simulations were conducted to evaluate the performance of the ILCA approach by comparing it with the EM algorithm (referred to as the *likelihood-based approach* from now on) when the data conformed to the LCDM. The simulation conditions were varied by manipulating (1) the number of attributes $D = [3, 5]$, (2) the number of items $I = [30, 50]$, (3) the number of respondents $N = [30, 50, 100, 500]$, (4) the item quality level $\omega = [\text{"Low"}, \text{"Medium"}, \text{"High"}]$, and (5) the shape of the latent-variable profile $S = [\text{"Flat"}, \text{"Nonflat"}]$. Similar simulation settings can be found in Chiu, Sun, and Bian (2018), as well as Templin and Bradshaw (2014). In total, there were $2 \times 2 \times 4 \times 3 \times 2 = 96$ conditions. The levels of D , I , and N are straightforward and therefore need no further clarifications. When there were three attributes ($D = 3$), a balanced Q-matrix was utilized in which each item measured either one or two attributes; similarly, in the condition with five attributes, each item measured two to three attributes. To avoid situations in which the attributes were not associated with any item, a constraint that each attribute should be measured at least three times was imposed when generating the Q-matrix. The levels of ω reflected the responding power of

items: how the changes in the required attributes influenced respondents' responses. When $\omega = \text{"Low"}$, the main effects were all set to a value of 0.2, and the interactions were all set to a value of 0.1. In the same order, these two parameters were set to 1 and -0.5 when $\omega = \text{"Medium"}$, and 3 and -1.5 when $\omega = \text{"High"}$. These settings were similar to Chiu et al.'s. Given known main and interaction effects, the intercepts were set to $-0.5 \cdot \lambda_i^T h(\alpha_{2D'}, q_i)$, with $\alpha_{2D'} = 1$. S described the relations between attributes. The "Flat" shape indicated that the probabilities of each latent class were equal (e.g., the probability for any one of the eight classes when $D = 3$ was .125). That said, the attributes were independent. The "Nonflat" shape, meaning that attributes were correlated with each other, was achieved through more steps: Continuous values were initially generated from a multinormal distribution $MV(\mathbf{0}, \Sigma)$, of which the diagonal elements of Σ were constrained to be 1 and the off-diagonal values were randomly drawn from a uniform distribution ranging from .6 to .9, and these continuous values were further converted onto the binary scale by comparing the values with zero (i.e., 1 if the value was larger than zero, and 0 otherwise).

The attribute accuracy was recorded in two ways: (1) the average classification rate of all individual attributes (this result is labeled as γ_1) and (2) the attribute profile classification rate (this result is labeled as γ_2). That is, the first measure examined the accuracy of attribute a for $a = [1, \dots, D]$, whereas the latter measure focused on the latent-class level.

Software and hardware

The simulation study was conducted in the R environment (R Core Team, 2018). Specifically, the likelihood-based estimation approach was implemented through the CDM package (George, Robitzsch, Kiefer, Groß, & Ünlü, 2016); alternatively, one could choose the package GDINA (Ma & de la Torre, 2018) or Latent Gold (Vermunt & Magidson, 2005; see DeCarlo, 2018, for details), which offer functionalities similar to those of CDM. On the other hand, the proposed approach deployed the polCA package (Linzer & Lewis, 2011) to execute the LCA. Each condition was replicated 50 times on a cloud computing facility with 32(64) [i.e., four 8(16)-core processors] AMD Opteron "bulldozer" processors and 64 GB RAM.

Results

The complete results are presented in Table 4. Overall, the ILCA approach outperformed the likelihood-based one substantially. The minimum, the mean, and the maximum of γ_1 for the likelihood-based approach were .52, .65, and .95, respectively, where the values were .52, .73, and .98 for the ILCA approach. In nine out of the 96 conditions, the ILCA

Table 4 Simulation results

D	I	N	ω	Likelihood (Flat)		ILCA (Flat)		Likelihood (NonFlat)		ILCA (NonFlat)	
				γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
3	30	30	lo	0.56	0.16	0.59	0.18	0.55	0.19	0.56	0.18
3	30	30	me	0.67	0.33	0.84	0.44	0.62	0.21	0.68	0.26
3	30	30	hi	0.81	0.50	0.88	0.56	0.74	0.31	0.82	0.38
3	30	50	lo	0.55	0.15	0.59	0.18	0.54	0.17	0.57	0.15
3	30	50	me	0.68	0.35	0.83	0.44	0.63	0.22	0.68	0.27
3	30	50	hi	0.81	0.51	0.88	0.57	0.75	0.31	0.82	0.37
3	30	100	lo	0.55	0.13	0.58	0.16	0.54	0.16	0.54	0.16
3	30	100	me	0.72	0.39	0.82	0.48	0.64	0.25	0.66	0.24
3	30	100	hi	0.84	0.61	0.91	0.67	0.76	0.35	0.81	0.39
3	30	500	lo	0.54	0.14	0.61	0.21	0.52	0.14	0.55	0.17
3	30	500	me	0.82	0.56	0.84	0.55	0.76	0.47	0.67	0.56
3	30	500	hi	0.95	0.87	0.99	0.90	0.92	0.78	0.82	0.86
3	50	30	lo	0.57	0.16	0.57	0.15	0.58	0.21	0.54	0.25
3	50	30	me	0.66	0.34	0.85	0.45	0.63	0.20	0.71	0.27
3	50	30	hi	0.83	0.54	0.91	0.60	0.74	0.28	0.85	0.38
3	50	50	lo	0.58	0.16	0.60	0.15	0.55	0.18	0.54	0.18
3	50	50	me	0.69	0.37	0.86	0.47	0.63	0.21	0.70	0.27
3	50	50	hi	0.83	0.54	0.92	0.61	0.75	0.30	0.82	0.36
3	50	100	lo	0.55	0.13	0.60	0.17	0.54	0.17	0.54	0.17
3	50	100	me	0.73	0.40	0.84	0.49	0.64	0.24	0.70	0.29
3	50	100	hi	0.78	0.50	0.90	0.63	0.77	0.34	0.85	0.41
3	50	500	lo	0.54	0.16	0.61	0.22	0.52	0.15	0.52	0.14
3	50	500	me	0.82	0.56	0.80	0.63	0.79	0.54	0.72	0.65
3	50	500	hi	0.95	0.85	0.97	0.84	0.84	0.52	0.85	0.51
5	30	30	lo	0.57	0.05	0.69	0.15	0.55	0.05	0.61	0.10
5	30	30	me	0.60	0.11	0.81	0.43	0.55	0.04	0.63	0.11
5	30	30	hi	0.72	0.25	0.86	0.35	0.61	0.05	0.65	0.09
5	30	50	lo	0.56	0.04	0.74	0.33	0.52	0.04	0.58	0.09
5	30	50	me	0.60	0.12	0.83	0.47	0.55	0.04	0.64	0.11
5	30	50	hi	0.68	0.19	0.84	0.45	0.61	0.07	0.69	0.13
5	30	100	lo	0.57	0.04	0.79	0.25	0.52	0.04	0.59	0.09
5	30	100	me	0.62	0.12	0.83	0.42	0.55	0.04	0.65	0.13
5	30	100	hi	0.66	0.18	0.89	0.52	0.62	0.06	0.70	0.13
5	30	500	lo	0.58	0.04	0.75	0.25	0.53	0.04	0.60	0.10
5	30	500	me	0.71	0.21	0.83	0.51	0.59	0.07	0.68	0.15
5	30	500	hi	0.75	0.25	0.95	0.67	0.66	0.09	0.74	0.16
5	50	30	lo	0.58	0.05	0.74	0.28	0.55	0.05	0.57	0.04
5	50	30	me	0.60	0.10	0.80	0.29	0.56	0.04	0.64	0.10
5	50	30	hi	0.73	0.27	0.83	0.55	0.61	0.05	0.69	0.11
5	50	50	lo	0.56	0.05	0.76	0.25	0.54	0.06	0.58	0.09
5	50	50	me	0.58	0.09	0.81	0.22	0.56	0.04	0.69	0.16
5	50	50	hi	0.71	0.19	0.88	0.59	0.62	0.05	0.73	0.14
5	50	100	lo	0.55	0.04	0.80	0.28	0.53	0.04	0.59	0.09
5	50	100	me	0.62	0.09	0.87	0.35	0.55	0.04	0.65	0.13
5	50	100	hi	0.65	0.15	0.78	0.36	0.60	0.05	0.73	0.17
5	50	500	lo	0.57	0.03	0.73	0.43	0.53	0.04	0.61	0.11
5	50	500	me	0.70	0.17	0.83	0.49	0.60	0.09	0.74	0.20
5	50	500	hi	0.76	0.27	0.89	0.61	0.66	0.08	0.84	0.17

approach yielded higher γ_1 than the likelihood-based approach. As compared with γ_1 , γ_2 had larger variability, since matching the entire profile is generally harder than matching a single attribute. Following the same order, the three aforementioned statistics were .03, .22, and .87 for the likelihood-based approach, and .04, .28, and .90 for the ILCA approach. There was only a 9.6% chance that a larger γ_2 would be yielded by the likelihood-based approach.

When D was small (i.e., 3), both γ_1 and γ_2 were more accurate than in the conditions with five attributes: the means of γ_1 were .69 and .74 for the likelihood-based and ILCA approaches, respectively, which became .60 and .73 when $D = 5$. This pattern was more obvious for γ_2 , in that, when $D = 3$, the means were .34 and .39 for the likelihood-based and ILCA approaches, respectively, and dropped to .09 and .26 when D increased to 5. Across all conditions, when D increased from 3 to 5, the standard deviations (SD s) of γ_1 and γ_2 decreased for both approaches; the mean sizes of the SD decrements were .10 and .16, meaning that the accuracy had less variability when D was larger. When D varied, ILCA was more robust when the five quantiles of the *accuracy changes* were below .05 for γ_1 and below .23 for γ_2 , where the same statistics were .19 and .60 for the likelihood-based approach.

The sample size N influenced the simulation results with a tempering trend when $N < 500$. The means of γ_1 for the likelihood-based approach were .64, .64, .63, and .69 for the four sample-size levels, where the values obtained via the ILCA approach were .73, .73, .73, and .76. This pattern, again, could be found in the results for γ_2 . However, the leaps of γ_2

from $N = 100$ to $N = 500$ were relatively large: (1) the means were boosted from .19 to .30 for the likelihood-based approach and from .30 to .42 for the ILCA approach, and (2) the standard deviations increased from .16 to .27 for the likelihood-based approach and from .17 to .26 for the ILCA approach. These findings reaffirmed the robustness of the ILCA approach when the sample size varied.

The item number I didn't show substantial main effects in either accuracy measure. On the other hand, the item quality ω did make a great difference, as Fig. 1 shows (see Jiang & Carter, 2018b, for visualization details). When $\omega = \text{"Low,"}$ the variability of the accuracy measures for the likelihood-based approach was much lower: The maximums for γ_1 and γ_2 were only .58 and .21, respectively. The ILCA approach, on the other hand, could still yield some accurate results in conditions of low item quality. As can be seen in the figure, the item quality was an important driving factor in attribute accuracy. When $\omega = \text{"High,"}$ the ILCA approach had a 75% chance for $\gamma_1 > .8$ and $\gamma_2 > .3$, where the chances for the likelihood-based approach were around 25% and 50%, respectively. Finally, I describe the attribute shape S . When $S = \text{"Flat,"}$ both approaches produced more accurate results than in the "Nonflat" conditions.

Discussion

This article has proposed a straightforward approach—ILCA—to estimate attributes within an LCDM framework.

Fig. 1 Influence of item quality

As compared with traditional likelihood-based estimation, the ILCA approach yields substantially higher accuracy, such that it is more appropriate to use it for producing performance/score reports. Performing LCA via the EM algorithm is generally fast, and therefore ILCA, which essentially is chaining LCA for each attribute, does not demand substantial computing capacity. However, it may slow down the speed if a large number of initial value sets are fed into the ILCA approach; a solution to this problem would be to use the package *doParallel* (Calaway, Microsoft Corp., Weston, & Tenenbaum, 2017), which executes parallel computing such that simultaneous operations can be substituted for unnecessary loops, and therefore the computation time can be reduced. Although this was not the primary concern of this article, ILCA is slower than the EM algorithm, especially when the number of attributes is small and the modeling complexity is low. For instance, the average computing times for the conditions with $D = 3$ were 11 and 34 s for the EM and ILCA algorithms, respectively. However, these times became 232 and 313 s when $D = 5$.

The simulation study showed that when attributes are independent (i.e., the shape is flat), attribute accuracy is higher than with correlated attributes. The subscore literature has shown that in practice, however, the attribute correlation is rarely zero; instead, in many situations the correlation level can be higher than .90 (see Haberman, 2008, and Sinharay, Haberman, & Puhan, 2007, for details). If the true attribute correlation level is above .9, the attribute accuracy is expected to be lower, according to these simulation results. However, if this situation indeed happens, the concern becomes whether it is necessary to distinguish the attributes, given that theoretically they can be collapsed into one attribute. Discussion of the theoretical definitions of subscore and attribute/latent variable relations can be found in Sinharay, Puhan, and Haberman (2011).

Although ILCA can be used to estimate the attributes, it does not provide item-level information as the traditional approaches do. In some situations, especially for large-scale (and standardized) assessments, item parameters are important, since they can be used for linking different forms, anchoring different groups, and other similar functions that allocate subjects into a common scale (Vale, 1986). In addition, item parameters are useful indicators for item bank construction, item feature verification, computerized adaptive-testing design, and theoretical framework investigations. Without providing item-level information, the ILCA approach cannot offer the aforementioned rich features that the traditional approaches do. However, if item parameters are not the primary concern, which might be the case in some informal and/or small-sample-size settings, the ILCA approach can be adopted, since it produces more accurate assessment results. Essentially, these results are often treated as the core component of score reports: Classifying respondents into the right groups is

critical, so that further treatments and/or interventions can be made accordingly and precisely. To emphasize, ILCA is not recommended in every possible CDM application, especially with large-scale assessments; whether to use it depends on the purpose of the work.

Although ILCA is based on sequential parametric approaches, as a whole it is not necessarily a parametric approach. Therefore, it can't provide inferential statistics to inform one as to the uncertainty of the estimates. To address this drawback, future research should focus on adopting bootstrapping or other nonparametric techniques to estimate the uncertainty. In addition, this article has only examined binary-type attributes and binary responses. Recently the DCM literature has developed beyond binary-scale studies; for example, Chen and de la Torre (2013) described models for polytomous attributes, and Ma and de la Torre (2016) proposed a sequential cognitive diagnosis model for polytomous item designs. Therefore, future ILCA works can focus on extending the technique to polytomous designs and investigate how it performs when it comes to an ordinal scale (Liu & Jiang, 2018). In addition, features such as response times and population-level variables in multiple contexts are not considered here (Jiao, Zhan, Liao, & Man, 2018). Future studies should integrate ILCA into those settings of rich information. Furthermore, how ILCA can be incorporated into a longitudinal setting (especially ones with intervention effects; Madison & Bradshaw, 2018) should be investigated in follow-up research. Last but not least, ILCA is likely to be used to conduct group comparisons, in which the invariance of the proposed approach can be tested (Shi, Song, Liao, Terry, & Snyder, 2017; Shi, DiStefano, McDaniel, & Jiang, 2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Hoboken, NJ: Wiley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.
- Calaway, R., Microsoft Corp., Weston, S., & Tenenbaum, D. (2017). *doParallel: Foreach parallel adaptor for the “parallel” package* [Computer software version 1.0.14]. Retrieved from <https://cran.r-project.org/web/packages/doParallel>
- Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, 37, 419–437.
- Chiu, C. Y., & Douglas, J. (2013). A nonparametric approach to cognitive diagnosis by proximity to ideal response patterns. *Journal of Classification*, 30, 225–250.

- Chiu, C. Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, 83, 355–375.
- Clark, D. M., & Wells, A. (1995). A cognitive model of social phobia. In R. G. Heimberg, M. R. Liebowitz, D. A. Hope, & F. R. Schneier (Eds.), *Social phobia: Diagnosis, assessment, and treatment* (pp. 69–93). New York, NY: Guilford Press.
- Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Orlando, FL: Holt, Rinehart & Winston.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:<https://doi.org/10.1007/BF02310555>
- da Silva, M. A., de Oliveira, E. S., Davier, A. A., & Bazán, J. L. (2017). Estimating the DINA model parameters using the No-U-Turn Sampler. *Biometrical Journal*, 60, 352–368.
- Dakanalis, A., Timko, C. A., Clerici, M., Zanetti, M. A., & Riva, G. (2014). Comprehensive examination of the trans-diagnostic cognitive behavioral model of eating disorders in males. *Eating Behaviors*, 15, 63–67.
- Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56, 463–474.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179–199.
- DeCarlo, L. T. (2018, April). *Insights from reparameterized DCMs: Implementation, monotonicity, and duality*. paper presented at the annual meeting of the National Council on Measurement in Education. New York City, New York.
- DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, 44, 285–291.
- George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R package CDM for cognitive diagnosis models. *Journal of Statistical Software*, 74, 1–24.
- Haberman, S. J. (2008). Subscores and validity (ETS Research Report RR-08-64). Princeton, NJ: Educational Testing Service.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation, ProQuest Information & Learning).
- 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. doi:<https://doi.org/10.1007/s11336-008-9089-5>
- Jiang, Z., & Carter, R. (2018a). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behavior Research Methods*. Advance online publication. doi:<https://doi.org/10.3758/s13428-018-1069-9>
- Jiang, Z., & Carter, R. (2018b). Visualizing library data interactively: Two demonstrations using R language. *Library Hi Tech News*, 35, 14–17.
- Jiang, Z., & Ma, W. (2018). Integrating differential evolution optimization to cognitive diagnostic model estimation. *Frontiers in Psychology*, 9, 2142. doi:<https://doi.org/10.3389/fpsyg.2018.02142>
- Jiao, H., Zhan, P., Liao, D., & Man, K. (2018). *Cognitive diagnostic modeling of responses and time for items in multiple contexts*. Paper presented at annual meeting of the National Council on Measurement in Education, New York City, New York.
- Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., & Jordan, M. I. (2016). Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences. In *Advances in neural information processing systems* (pp. 4116–4124). Cambridge, MA: MIT Press.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258–272.
- Li, H., Hunter, V., & Lei, P. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33, 391–409.
- Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42, 1–29.
- Liu, R. (2017). Misspecification of attribute structure in diagnostic measurement. *Educational and Psychological Measurement*, 78, 605–634. doi:<https://doi.org/10.1177/0013164417702458>
- Liu, R., & Jiang, Z. (2018). Diagnostic classification models for ordinal item responses. *Frontiers in Psychology*, 9, 2512. doi:<https://doi.org/10.3389/fpsyg.2018.02512>
- Liu, R., Qian, H., Luo, X., & Woo, A. (2017). Relative diagnostic profile: A subscore reporting framework. *Educational and Psychological Measurement*, 78, 1072–1088.
- Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.
- Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, 69, 253–275.
- Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework [Computer software version 2.1]. Retrieved from <https://CRAN.R-project.org/package=GDINA>
- Madison, M. J., & Bradshaw, L. (2018). Evaluating intervention effects in a diagnostic classification model framework. *Journal of Educational Measurement*, 55, 32–51.
- Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463–469.
- Muthén, L. K., & Muthén, B. O. (2013). *Mplus user's guide* (Version 6.1) [Computer software and manual]. Los Angeles, CA: Muthén & Muthén.
- Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, 24, 342–366.
- R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <http://www.Rproject.org/>
- Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment*, 34, 782–799.
- Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, 68, 78–96.
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Methodology in the social sciences. Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Seixas, F. L., Zadrozny, B., Laks, J., Conci, A., & Saade, D. C. M. (2014). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Computers in Biology and Medicine*, 51, 140–158.
- Sinharay, S., Haberman, S., & Puhán, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, 26, 21–28.
- Sinharay, S., Puhán, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, 30, 29–40.
- Shi, D., DiStefano, C., McDaniel, H. L., & Jiang, Z. (2018). Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling*, 25, 924–945. doi:<https://doi.org/10.1080/10705511.2018.1449653>
- Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research*, 52, 430–444. doi:<https://doi.org/10.1080/00273171.2017.1306432>
- Svetina, D., Dai, S., & Wang, X. (2017). Use of cognitive diagnostic model to study differential item functioning in accommodations. *Behaviormetrika*, 44, 313–349.

- 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. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, 11, 287–305. doi:<https://doi.org/10.1037/1082-989X.11.3.287>
- Templin, J., & Hoffinan, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, 32, 37–50.
- Titterton, D. M., Smith, A. F., & Makov, U. E. (1985). Statistical analysis of finite mixture distributions, Chichester, UK: Wiley.
- Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, 10, 333–344.
- Vermunt, J. K., & Magidson, J. (2005). Latent GOLD® choice 4.0 user's manual. Belmont, MA: Statistical Innovations Inc.
- Wolfe, J. H. (1970). Profile clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5, 329–350.