frontiers in Psychology

METHODS
published: 06 November 2018
doi: 10.3389/fpsyg.2018.02142

Integrating Differential Evolution Optimization to Cognitive Diagnostic Model Estimation

Zhehan Jiang[1]* *and Wenchao Ma*[2]

[1] Department of University Libraries, University of Alabama, Tuscaloosa, AL, United States, [2] Department of Educational Studies, University of Alabama, Tuscaloosa, AL, United States

A log-linear cognitive diagnostic model (LCDM) is estimated via a global optimization approach- differential evolution optimization (DEoptim), which can be used when the traditional expectation maximization (EM) fails. The application of the DEoptim to LCDM estimation is introduced, explicated, and evaluated via a Monte Carlo simulation study in this article. The aim of this study is to fill the gap between the field of psychometric modeling and modern machine learning estimation techniques and provide an alternative solution in the model estimation.

OPEN ACCESS

Keywords: differential evolution optimization, cognitive diagnostic model, LCDM, estimation, EM algorithm

Edited by:
Hong Jiao,
University of Maryland, College Park, United States

Reviewed by:
Chun Wang,
University of Minnesota Twin Cities, United States
Peida Zhan,
Zhejiang Normal University, China

***Correspondence:**
Zhehan Jiang
zjiang17@ua.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *15 May 2018*
Accepted: *18 October 2018*
Published: *06 November 2018*

Citation:
Jiang Z and Ma W (2018) Integrating Differential Evolution Optimization to Cognitive Diagnostic Model Estimation. Front. Psychol. 9:2142. doi: 10.3389/fpsyg.2018.02142

Assessments have been widely used in education as a part of a summative program for many purposes, such as evaluating whether students have reached the desired proficiency level and determining whether students should be given a scholarship. However, in the past decades, stakeholders have shown a strong interest in the information of students' strengths and weaknesses of their knowledge and skills. This has led to fruitful exploration in the field of psychometrics of how to extract diagnostic information to enhance classroom instruction and learning. Cognitive diagnostic models (CDMs) are a set of psychometric models developed to identify whether a student masters a set of fine-grained skills, such as addition, subtraction, multiplication, and division in math ability assessments. For example, question "2+4–1" measures addition and subtraction, and "4 × 2/3" measures multiplication and division. Although it seems straightforward to conclude that a student may not master addition or subtraction if s\he fails 2+4–1, it is indeed much more complicated in practice in that students may answer a question correctly by guessing or fail a question due to carelessness. As a result, formal psychometric models such as CDMs should be employed for data analysis to make sure the inferences are valid. In addition to educational testing, CDMs are useful in psychological measurement. For example, the literature indicates that neuro-vegetative symptoms are a general construct that contains three attributes: depression (DEP), fatigue (FAT), and sleeplessness (SLE; Rabinowitz et al., 2011). Using CDMs allows researchers/practitioners to investigate the attributes of a given patient. Among the item data types, a binary scale is the most common one that has been adopted in many surveys and measures.

Prior to the data analysis using CDMs, whether a skill is required for answering a question needs to be determined by content experts and/or cognitive psychologists and specified in a binary matrix (Q-matrix; Tatsuoka, 1983) as illustrated in **Table 1** such that theory-granted structure can be applied to the measurement of interest. Rows of the Q-matrix represent questions and columns represent skills. Element 1 indicates that the skill is measured by the question and 0 indicates that the sill is not measured.

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

Recent advances in modeling development have produced several general CDMs, such as the Log-linear CDM (LCDM; Henson et al., 2009) and, equivalently, the generalized Deterministic Input; Noisy "And" gate model (G-DINA; de la Torre, 2011). The LCDM provides great flexibility such as (1) subsuming most latent variables, (2) enabling both additive and non-additive relationships between skills and questions simultaneously, and (3) syncing with other psychometric models. Rupp et al. (2010, p. 163) proved that LCDM can be constrained to core CDMs such as Deterministic Input; Noisy "And" gate (DINA; Junker and Sijtsma, 2001) model, Noisy Input; Deterministic "And" gate (NIDA; Junker and Sijtsma, 2001) model, and the Reduced Reparameterized Unified Model (RRUM; Hartz, 2002), and Deterministic Input; Noisy "Or" gate (DINO, Templin and Henson, 2006) model.

The LCDM is essentially a restricted latent class model (Day, 1969; Wolfe, 1970; Titterington et al., 1985), and mathematically, it can be defined as:

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

where $\boldsymbol{y}_p = (y_{p1}, y_{p2}, \ldots, y_{pI})$ is the binary response vector of person p on a test comprised of I items, and element y_{pi} is the response on item i. v_c is the probability of membership in latent class c, and π_{ci} is the probability of correct response to item i by person p from latent class c. The log-likelihood of observing item responses of N persons can be expressed as

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

Further, Equation 2 can also be converted to:

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

where $\log(\prod_{i=1}^{I} \pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}})$ can be replaced by $\sum_{i=1}^{I} \log(\pi_{ci}^{y_{pi}} \left(1-\pi_{ci}\right)^{1-y_{pi}})$ due to the mathematical property of *log* operation.

Assume the number of attributes is A, the mastery profile of the attributes for a random person is denoted by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_A)$, where element α_a is either 1 or 0. In total, there are 2^A possible attribute profiles and correspondingly 2^A latent classes. For example, when A=4, a person with attribute profile $\boldsymbol{\alpha}$= (1, 1, 1, 0) has mastered the first three attributes except the last one. As illustrated in **Table 1**, a Q-matrix of size I^*A is necessary for a LCDM, where the (i, a) element q_{ia} is 1 when item i measures attribute a and 0 otherwise. The conditional probability of person p with attribute profile $\boldsymbol{\alpha}_c$ answering item i correctly can be written by

$$\pi_{ci} = P\left(y_{pi}=1 \middle| \boldsymbol{\alpha}_c\right) = \frac{\exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)\right)}{1+\exp\left(\lambda_{i,0} + \boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)\right)}, \quad (4)$$

where $\boldsymbol{q}_i$ is the set of Q-matrix entries for item i, $\lambda_{i,0}$ is the intercept parameter, where $\boldsymbol{\lambda}_i$ represents a vector of length 2^A-1 that contains main effect and interaction effect parameters of item i, and $\boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$ is a vector of the same length with linear combinations of the $\boldsymbol{\alpha}_c$ and $\boldsymbol{q}_i$. Particularly, $\boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right)$ can be expanded to:

$$\boldsymbol{\lambda}_i^T \boldsymbol{h}\left(\boldsymbol{\alpha}_c, \boldsymbol{q}_i\right) = \sum_{a=1}^{A} \lambda_{i,1,(a)} \alpha_{ca} q_{ia} + \sum_{a=1}^{A-1} \sum_{a'>a}^{A} \lambda_{i,2,(a,a')} \alpha_{ca} \alpha_{ca'} q_{ia} q_{ia'} + \ldots, \quad (5)$$

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

LCDM ESTIMATION

Traditionally, estimating LCDMs refers to the expectation maximization (EM) algorithm (Bock and Aitkin, 1981) that maximizes the marginal likelihood; this is the most commonly-seen algorithm in the CDM literature. In addition to the EM algorithm, Markov chain Monte Carlo (MCMC) techniques can be, theoretically, used to estimate the LCDM, but to date its application remains upon simpler CDMs such as the DINA model (da Silva et al., 2017; Jiang and Carter, 2018a). This study focuses on the EM algorithm due to its practicality and popularity. The EM algorithm is an intertwined updating mechanism consisting of E- and M-steps. With the provisional item parameter and probability of membership estimates from iteration t-1 (i.e., $\boldsymbol{\lambda}$s and $\boldsymbol{v}$s), the posterior class probability for person p can be obtained in the E-step by

$$H\left(C=c \middle| \boldsymbol{Y_p}=\boldsymbol{y_p}\right) = \frac{v_c \prod_{i=1}^{I} \pi_{pi}^{y_{pi}} \left(1-\pi_{pi}\right)^{1-y_{pi}}}{\sum_{c=1}^{C} v_c \prod_{i=1}^{I} \pi_{pi}^{y_{pi}} \left(1-\pi_{pi}\right)^{1-y_{pi}}} \quad (6)$$

Based on Equation (6), the expected number of persons in latent class c and the expected number of persons in latent class c who

TABLE 2 | A 3-item sample of expressions of a log-linear cognitive diagnostic model.

Item	α_1	α_2	α_3	Expanded $\lambda_{i,0}+\lambda_i^T h(\alpha_c, q_i)$ expression	Shortened expression
1	1	0	0	$\lambda_{1,0}+\lambda_{1,1}(1)+\lambda_{1,2}(0)+\lambda_{1,3}(0)+\lambda_{1,12}(1\times0)+\lambda_{1,13}(1\times0)+\lambda_{1,23}(0\times0)+\lambda_{1,123}(1\times0\times0)$	$\lambda_{1,0}+\lambda_{1,1}(1)$
2	0	1	1	$\lambda_{2,0}+\lambda_{2,1}(0)+\lambda_{2,2}(1)+\lambda_{2,3}(1)+\lambda_{2,12}(0\times1)+\lambda_{2,13}(0\times1)+\lambda_{2,23}(1\times1)+\lambda_{2,123}(0\times1\times1)$	$\lambda_{2,0}+\lambda_{2,2}(1)+\lambda_{2,3}(1)+\lambda_{2,23}(1)$
3	1	1	1	$\lambda_{3,0}+\lambda_{3,1}(1)+\lambda_{3,2}(1)+\lambda_{3,3}(1)+\lambda_{3,12}(1\times1)+\lambda_{3,13}(1\times1)+\lambda_{3,23}(1\times1)+\lambda_{3,123}(1\times1\times1)$	$\lambda_{3,0}+\lambda_{3,1}(1)+\lambda_{3,2}(1)+\lambda_{3,3}(1)+\lambda_{3,12}(1)+\lambda_{3,13}(1)+\lambda_{3,23}(1)+\lambda_{3,123}(1)$

answer item i correctly can be obtained by:

$$n_c = \sum_{p=1}^{N} H\left(C=c \,\middle|\, \boldsymbol{Y_p}=\boldsymbol{y_p}\right), \text{ and}$$

$$r_{ci} = \sum_{p=1}^{N} y_{pi} H\left(C=c \,\middle|\, \boldsymbol{Y_p}=\boldsymbol{y_p}\right),$$

respectively. In the M-step, the following function is maximized with respect to item parameters $\boldsymbol{\lambda}$:

$$\ell = \sum_{i=1}^{I}\sum_{c=1}^{2^A}\left[r_{ci}\log\pi_{ci} + (n_c - r_{ci})\log(1-\pi_{ci})\right],$$

and the probability of membership is updated by

$$\nu_c = \frac{\sum_{p=1}^{N} H\left(C=c \,\middle|\, \boldsymbol{Y_p}=\boldsymbol{y_p}\right)}{N}.$$

Maximizing objective function ℓ usually requires Newton or Fisher scoring methods, where first- and second-order derivatives i.e., $\frac{\partial L}{\partial \lambda}\cdot\left(\frac{\partial^2 L}{\partial \lambda^2}\right)^{-1}$ where the first component is a vector and the second component is a matrix) of the objective function are needed. If $\frac{\partial^2 L}{\partial \lambda^2}$ becomes 0, the iteration process will stop and therefore fail to converge.

As a restricted latent class models, LCDM estimation faces the risk of local maxima (Jin et al., 2016). Theoretically, to obtain valid and accurate estimates, the model estimation should converge at a global maximum of the likelihood function, however, the mixture component of a mixture model is likely to trap the aforementioned EM updates to local maxima. In addition, label switching can occur and therefore lead to a misinterpretation of an estimation. For instance, a person mastering all attributes of interest can be mistakenly labeled as one with zero-mastery. Basing on the traditional EM approach, Rupp et al. (2010) add constraints to the parameter estimates (e.g., ensuring main effects are non-negative); this constraint approach substantially reduces the risks of local maxima and label switching (Lao and Templin, 2016). Using *Mplus* (Muthén and Muthén, 2013), a commercial software designed for latent variable modeling that by default deploys the traditional EM approach, Templin and Hoffman (2013) outline the procedures to specify syntax with parameter constraints for the LCDM estimation. Note that in the LCDM estimation, the EM approach in *Mplus* is turned into an accelerated version, meaning its updating steps are replaced with Quasi-Newton and Fisher scoring, this, however, still falls under the family of the traditional EM algorithm. Although Templin and Hoffman's *Mplus* practice has been implemented in many published works and is proved to be efficient (see Bradshaw and Templin, 2014; Li et al., 2016; Ravand, 2016 for example), it is still not avoiding the convergence failure issue: Templin and Bradshaw (2014) conduct a simulation study with vast conditions each of which was replicated 500 times, where the result shows the numbers of converged replications range from 330 to 447. To avoid the convergence issue while maintaining the properties of the EM approach, we introduce a machine-learning technique named Differential Evolution to estimate LCDMs.

DIFFERENTIAL EVOLUTION

Global optimization under machine-learning umbrella has gained tremendous attention from researchers, mathematicians as well as professionals in the field of engineering, finance, and scientific areas (Mohamed et al., 2012). Many applications of this kind impose complex optimization problems such traditional estimation techniques based upon derivatives become cumbersome or even impossible. To avoid the mathematical deriving procedures yet provide reliable solutions to complex models, Differential Evolution (DE) is invented (Storn and Price, 1997), developed, and applied to practice in different fields (e.g., Paterlini and Krink, 2006; Das et al., 2008; Rocca et al., 2011). Inspired by Darwinian evolution that entails the idea of mutation, crossover, and selection, DE is an enhanced version of derivative-free evolutionary algorithms and has been recognized as a simple yet efficient optimization approach in solving a variety of benchmark problems. The complete DE algorithm cycle can be found in **Figure 1**; in particular, the algorithm starts by sampling D candidate solutions to the problem of interest, where each candidate solution can be either a scalar or a vector (if there are more than one estimate). The mutation procedure takes place by performing simple arithmetic operations (i.e., addition, subtraction, and multiplication) among the existing solutions (namely parent solutions). The resultant mutation outcomes are then crossed over with the parent solutions to produce new candidate (offspring) solutions. Finally, in a one-to-one selection process of each pair of offspring and parent vectors, candidate solutions that fit the model better are passes into the next evolutionary cycle. This cycle iterates until the estimation

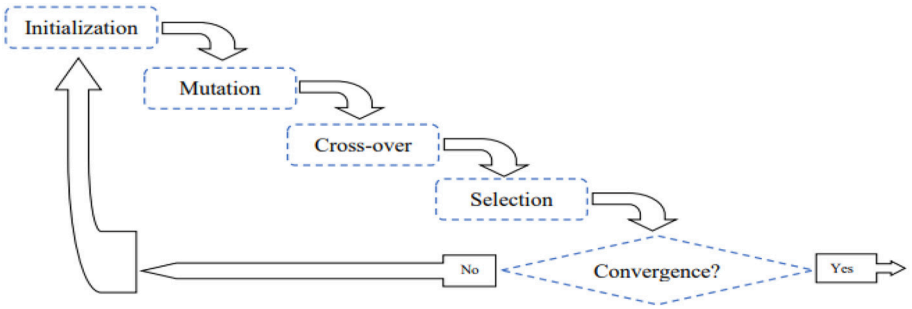


FIGURE 1 | The iterative cycle of differential evolution algorithm.

converge. Mathematical and algorithmic details can be found in the following paragraph.

Let R be the number of estimates, $\boldsymbol{\lambda}_{LO}$ and $\boldsymbol{\lambda}_{HI}$ be the lower and the upper limits (vectors) of the parameters of the estimates, and $G(\cdot)$ be the objective function. Initial candidate solutions $\boldsymbol{\lambda}_d = (\lambda_{d1}, \lambda_{d2}, \ldots, \lambda_{dR})$ for $d = 1, 2, \ldots, D$ can be generated by (1) randomly drawing samples from certain distribution(s) or (2) specifying values with educated guesses, where D is the number of candidate solutions. Mutation procedure can be achieved via different strategies: (1) DE /rand/ 1, (2) DE/current-to-rest/1, and (3) DE/best/1. In particular, for a given set of candidate solutions $\boldsymbol{\lambda}_d$ for $d = 1, 2, \ldots, D$, the mutation outcomes $\boldsymbol{m}_d$ can be calculated as:

$$
\begin{aligned}
DE/rand/1 \qquad & \boldsymbol{m}_d = \boldsymbol{\lambda}_{\delta_o} + F_d{}^*(\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2}) \\
DE/current-to-rest/1 \qquad & \boldsymbol{m}_d = \boldsymbol{\lambda}_{\delta_o} + F_d{}^*(\boldsymbol{\lambda}_{best} - \boldsymbol{\lambda}_{\delta_2}) \\
& \qquad\quad + F_d{}^*(\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2}) \\
DE/best/1 \qquad & \boldsymbol{m}_d = \boldsymbol{\lambda}_{best} + F_d{}^*(\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2})
\end{aligned}
$$

Where δ_o, δ_1, and δ_2 are distinct integers uniformly sampled from 1 to D, $\boldsymbol{\lambda}_{\delta_1} - \boldsymbol{\lambda}_{\delta_2}$ is the difference vector that would be used to mutate two selected parent candidates (e.g., DE/rand/1), $\boldsymbol{\lambda}_{best}$ is the best candidate solution at the current iteration, and finally F_d is the mutation scaling factor that is randomly drawn from a uniform distribution on the interval (0, 1). Some $\boldsymbol{m}_d$ may be produced beyond the constraints set by $\boldsymbol{\lambda}_{lo}$ and $\boldsymbol{\lambda}_{hi}$; some effective solutions to the violation include (1) re-generating a candidate solution until it is valid and (2) setting penalty to the objective function. If an element r in a candidate solution encounter the boundary issue, a quick fix by setting the violating elements to be the middle between boundaries and the that of its parent solution. That is, $m_{dr} = \frac{\lambda_{LOr} + \lambda_{dr}}{2}$ for $m_{dr} < \boldsymbol{\lambda}_{LOr}$ and $m_{dr} = \frac{\lambda_{HIr} + \lambda_{dr}}{2}$ for $m_{dr} > \boldsymbol{\lambda}_{HIr}$. After obtaining $\boldsymbol{m}_d$ from the mutation procedure, a "binomial" crossover operation forms the offspring candidate solutions: let CR be a crossover probability that controls the fraction of the elements that are copied from the parent candidate solution and u_{dr} be a candidate solution, if a random number z_r sampled from a uniform distribution (0, 1) is smaller than CR, the element r of the offspring of u_{dr} is m_{dr}, and λ_{dr} otherwise. The default CR is usually set to 0.5 for a balanced stochastic move. Finally, if $G(\boldsymbol{u_d})$ is better than $G(\boldsymbol{\lambda_d})$, $\boldsymbol{u_d}$ would replace $\boldsymbol{\lambda_d}$ to serve as a parent solution for the next iteration. The DE algorithm can be tailored to a parallel computing platform; technically each candidate solution can be calculated in an independent computational unit such that queuing time can be shortened. That said, instead of sequentially updating the candidate solutions, a parallel DE algorithm can perform simultaneous updates.

To illustrate how the DE algorithm functions, an example of a simple regression estimation is provided here. Let independent variable $\boldsymbol{x} = [22, 14, 15, 12, 10, 26, 11, 28]$ and dependent variable $\boldsymbol{y} = [44, 29, 30, 27, 24, 51, 25, 56]$ resulting in $\hat{\boldsymbol{\beta}} = [4.98, 1.78]$ with the ordinary least squares (OLS) estimator. Using the DE algorithm in this case sets the objective function $G(\lambda)$ to $-1^* \sum (\boldsymbol{y} - \hat{\boldsymbol{y}})^2$, which ideally should be maximized to -7.19 according to the OLS result. To keep the demonstration simple, let the number of candidate solutions $D = 3$ and initial values for $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, and $\boldsymbol{\lambda}_3$ were arbitrarily set to [2, 1], [−3, 5], and [1, 2]. At the initial iteration, the best solution was [1, 2] as $G(\boldsymbol{\lambda_3}) = -24$ where $G(\boldsymbol{\lambda_1})$ and $G(\boldsymbol{\lambda_2})$ were −2400 and −21672. Therefore, $\boldsymbol{\lambda}_{best}$ at this stage became $\boldsymbol{\lambda_3}$. With certain random draws for a given mutation calculation (e.g., DE/best/1), $\boldsymbol{m}_1$, $\boldsymbol{m}_2$, and $\boldsymbol{m}_3$ happened to be [3.5, 0.8], [−1, 3.5], and [2, 1.8]. Let $CR = 0.5$, if a random generation produced $z_1 = 0.7$ and $z_2 = 0.4$ for example, the first offspring $\boldsymbol{u_1}$ became [3.5, 1] by taking elements from $\boldsymbol{\lambda_1}$ and $\boldsymbol{m}_1$. This resulted in $G(\boldsymbol{u_1}) = -2022$, which is larger than $G(\boldsymbol{\lambda_1})$, and therefore the new $\boldsymbol{\lambda_1}$ would be replaced by $\boldsymbol{u_1}$. On the other hand, if $\boldsymbol{u_3}$ became [1, 1.8] which produced $G(\boldsymbol{u_1}) = -116.8$, then the $\boldsymbol{\lambda_3}$ remained still. This process continues until $G(\boldsymbol{\lambda})$ converges to −7.19.

In this paper, we integrated the DE into the EM algorithm to estimate LCDMs[1] To make the proposed approach easy to follow, we name it EM-DEoptim algorithm from here. Especially, the method for updating item parameters within the M-step is replaced by the DE algorithm, while the rest of the EM procedures remain identical. To be concrete, the objective function that the EM-DEoptim maximizes is Equation 3, given v_c for each latent class is known. As the DE is a

[1]The snippet code can be found https://alabama.box.com/s/cbuxetk19b1pk1invi1gnbqxd8hw5fij.

stochastic and global optimization technique, the EM-DEoptim is expected to encounter fewer occurrences of the local maxima problem than the traditional EM algorithm (Celeux et al., 1996). In addition, as addressed above, the EM-DEoptim is based upon derivative-free framework such that it can be easily fitted to arbitrarily customized LCDMs without re-deriving the gradient functions nor re-approximating information matrix. For example, if constraining the main effects of Item i and Item i' to be equal while still allowing others to be estimated freely is needed, the EM-DEoptim algorithm can handle the situation by simply assigning the same labels to the constrained parts in the likelihood function expression, where the traditional EM algorithm needs altering the derivatives. This advantage can effectively prevent the aforementioned un-differentiable situations. Last but not least, the computational speed of the EM-DEoptim algorithm, although not outperform the traditional EM algorithm in a singular operation environment, can be substantially improved via parallel computing facilities that are naturally suited to modern machine-learning-based techniques.

SIMULATION STUDY

We conducted a simulation study to demonstrate the utility of the EM-DEoptim algorithm. Specifically, the study involved two investigations: the number of times that the traditional EM algorithm fails and the comparison between the EM-DEoptim algorithm and the traditional EM algorithm in terms of the parameter recovery. In the simulation study, the numbers of attributes A were set to 3, 4, 5. The Q-matrix was randomly generated: when there were 3 attributes ($A = 3$), a balanced Q-matrix in which each item measures either one or two attributes was utilized; similarly, at the condition of 4 and 5 attributes, each item measures two to three attributes. The number of items I was set to 30 and the number of persons N was set to 300. The attributes were generated via two steps: continuous values were initially generated from a multinormal distribution $MV(\mathbf{0}, \Sigma)$ of which the diagonal elements of Σ were constrained to 1 and the off-diagonal values (i.e., correlations between attributes) were randomly drew from a uniform distribution ranging from 0.7 to 0.9, and these continuous values were further converted onto the binary scale by comparing the values with zero (i.e., 1 if the value is larger than zero and 0 otherwise). Finally, the item parameters were specified to two level: high-quality group that sets main effects = 2, intercepts = −1.5, and interaction effects = 0.5, and low-quality group that makes main effects = 0.2, intercepts = −0.5, and interaction effects = 0.1.

The traditional EM algorithm was realized via the package *CDM* (George et al., 2016; alternatively, one can choose the package *GDINA* by Ma and de la Torre, 2018), where the EM-DEoptim algorithm was executed in *R* (R Core Team, 2018). The stop criterion in *CDM* was set to 1,000 iterations or the change of likelihood value <0.001, where the EM-DEoptim algorithm was forced to stop if the iteration number reaches to 1,000 or the likelihood value remains identical for 10 iterations. In this study, the DE configurations were set to default (Ardia et al., 2011): DE/current-to-rest/1 with $F_d = 0.8$, $CR = 0.5$, and 500 candidate solutions, where ± 20 is used to constrain the parameter estimates. The machine used was Dell Precision 3520 with 16GB RAM and a 2.90 GHz i7-7820 4-core Intel processor. The study was replicated for 200 times.

The dependent variables in this part of the study are (1) the number of convergence failure of the EM algorithm, (2) relative bias (RBIAS) and root mean squared error (RMSE), and (3) the attribute classification accuracy measured by each attribute and each profile. Overall, there was only two failed convergence failures when the item quality was high, where the low-quality item parameters led to seven failures: two cases in the situation of $A = 4$ and five cases when $A = 5$. On the other hand, EM-DEoptim had no unexpected terminations during the iterations. **Table 3** shows the attribute classification accuracy rates. Both algorithms produced very similar results, where some patterns can be discovered: (1) the more attributes the estimation face, the less accurate the attribute estimates are yielded, (2) the higher the item parameter quality is, the more accurate the attribute estimates are produced, and (3) the profile accuracy is more sensitive to the item parameter quality.

Similar to the attribute estimates, the item parameter recovery presented similar pattern for both algorithms as listed in **Table 4**. The biases and MSEs were higher when (1) the number of attributes was larger and (2) the item parameter quality is higher. In addition, main effect estimates were more accurate and efficient than both interaction and intercept effects. This finding is not uncommon in complex psychometric models (Jiang et al., 2016). When the item parameter quality is low, and/or the number of attribute is large (e.g., 5), the EM-DEoptim performed better than the traditional EM algorithm. An important reason is that the boundary constraints imposed by the EM-DEoptim algorithm can limit the estimates into a certain range. Although not a main focus of the studies, the computing speed showed a substantial difference: the average time (in seconds) for 3-, 4-, and 5 attributes were 4.45, 22.55, and 78.64 for the traditional EM algorithm, while the EM-DEoptim took 61.22, 354.18, and 1228.76.

REAL DATA APPLICATION

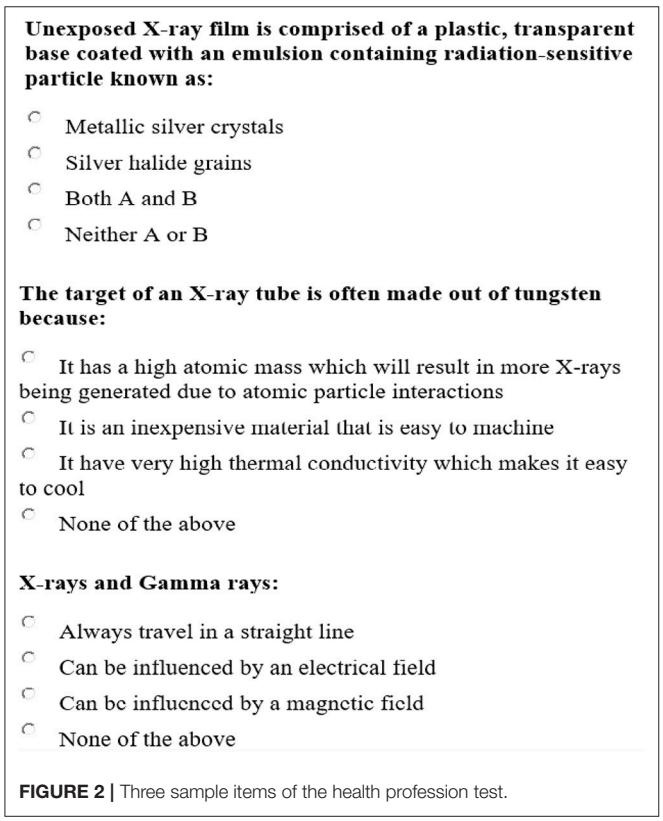
The dataset used in this session is an assessment of a health profession administered to 3491 test takers (Jiang and Raymond, 2018). The number of items is 200 each of which measures

TABLE 3 | attribute accuracy rate of the simulation study.

		EM		DE-EMoptim	
A	Quality	Attribute	Profile	Attribute	Profile
3	High	0.848	0.634	0.847	0.642
4	High	0.816	0.505	0.821	0.489
5	High	0.768	0.336	0.755	0.342
3	Low	0.516	0.104	0.517	0.104
4	Low	0.509	0.039	0.513	0.044
5	Low	0.504	0.017	0.468	0.013

TABLE 4 | Item parameter estimates of the simulation study.

RBIAS		EM			EM-DEoptim		
A	Quality	Main	Intercept	Interaction	Main	Intercept	Interaction
3	High	−0.433	−0.270	−1.896	−0.463	−0.280	−1.196
4	High	−1.744	−0.948	−5.036	−1.740	−0.938	−4.106
5	High	−1.906	−0.957	−4.975	−1.906	−0.957	−2.675
3	Low	−5.389	−8.906	−2.160	−2.389	−4.406	−1.960
4	Low	−9.526	−9.950	−3.696	−8.626	−7.950	−4.026
5	Low	−11.99	−6.419	−10.027	−9.495	−5.419	−7.027
RMSE		**EM**			**EM-Deoptim**		
3	High	6.346	2.124	8.349	6.890	1.924	4.336
4	High	14.576	5.035	15.679	15.079	6.422	13.853
5	High	21.643	6.230	26.327	18.223	7.360	18.707
3	Low	15.423	11.630	18.478	12.863	12.112	18.481
4	Low	22.512	15.165	26.064	19.299	16.865	17.446
5	Low	25.410	15.109	15.319	16.506	14.409	14.319


Unexposed X-ray film is comprised of a plastic, transparent base coated with an emulsion containing radiation-sensitive particle known as:

- Metallic silver crystals
- Silver halide grains
- Both A and B
- Neither A or B

The target of an X-ray tube is often made out of tungsten because:

- It has a high atomic mass which will result in more X-rays being generated due to atomic particle interactions
- It is an inexpensive material that is easy to machine
- It have very high thermal conductivity which makes it easy to cool
- None of the above

X-rays and Gamma rays:

- Always travel in a straight line
- Can be influenced by an electrical field
- Can be influenced by a magnetic field
- None of the above

FIGURE 2 | Three sample items of the health profession test.

one attribute. Therefore, there are five attributes in total: the knowledge of radiation biology (Items #1-45), the knowledge of equipment operation (Items #46-67), the image acquisition and evaluation capacity (Items #68-112), the knowledge of imaging procedures (Items #113-162), and ethics (Items #163-200). Three samples of the items can be found in **Figure 2**.

Two common model fit indices are reported as: (1): mean of absolute deviations in observed and expected correlations (MADcor; DiBello et al., 2007) is 0.041 and standardized mean square root of squared residuals (SRMSR; Maydeu-Olivares, 2013; Maydeu-Olivares and Joe, 2014) is 0.05. Overall, the model has an adequate fit. Note that more model fit indices such as χ^2-like statistics (Orlando and Thissen, 2000) are recommended. This paper focuses on the estimation. More model fit details can be found in Hu et al. (2016) and Sorrel et al. (2017).

Rounding the number of digits to three after the decimal point, one can see that 16 classes are nearly empty and therefore are labeled as "others" in **Figure 3** (see Jiang and Carter, 2018b for more visual aids). Nearly 40% of the test takers master all five attributes. According to Templin and Bradshaw (2014), many empty classes indicate potential hierarchies of attribute structure, however, the parameter estimates can be relatively robust even the non-hierarchical modeling is adopted here. **Figure 4** shows the distributions of the parameter estimates grouped by parameter types and attribute identifications. Attribute #3 had the highest means of both intercepts and main effects: 2.65 and 1.79. The means of intercepts and main effects of Attribute #5 were −1.05 and 0.20.

To compare the estimates with other estimation approaches, we also implemented a Bayesian technique-Hamiltonian Monte Carlo-to the analysis by adopting uninformative priors for both item parameters and the class membership probability: the mean and standard deviation for item parameters were 0 and 20, while the Dirichelet prior parameters were all set to 1 (see Jiang and Carter, 2018a for details). The correlations of item parameter estimates were relatively high: 0.77, 0.84, and 0.69 for intercept, main effect, and interaction effects. On the other hand, the attribute agreement was lower than that of the item parameter estimates: the average ratio for all attributes was 0.67, where the value dropped to 0.39 when it comes to the match of the class membership classification. This makes sense as the Dirichelet

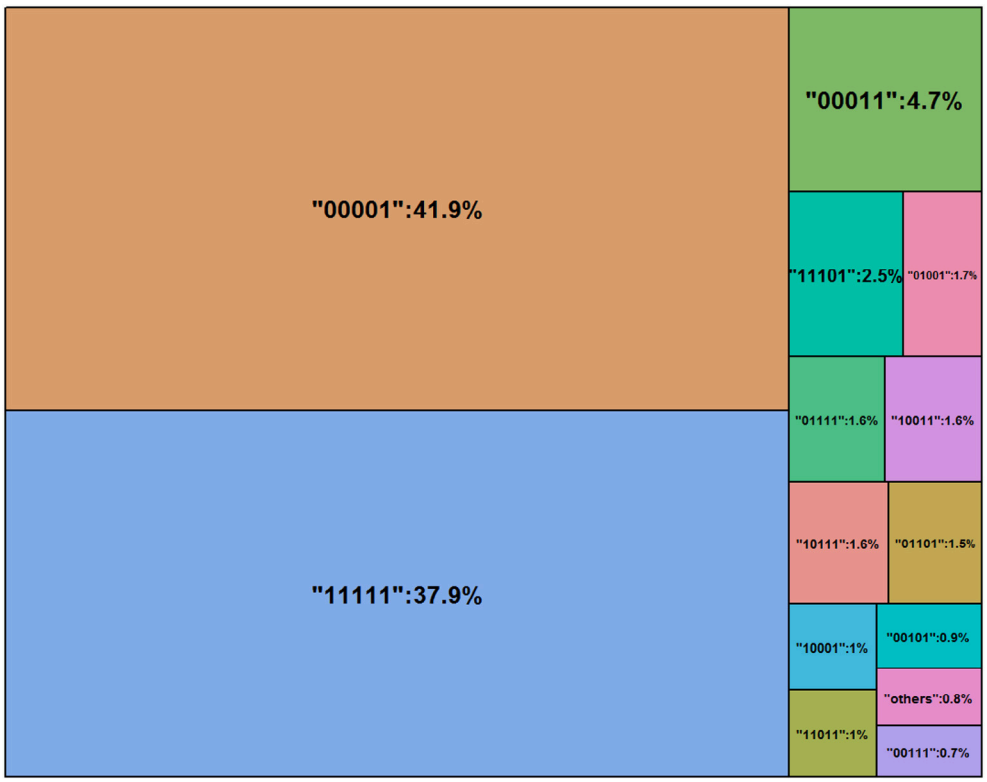


FIGURE 3 | Estimated class probabilities via LCDM.

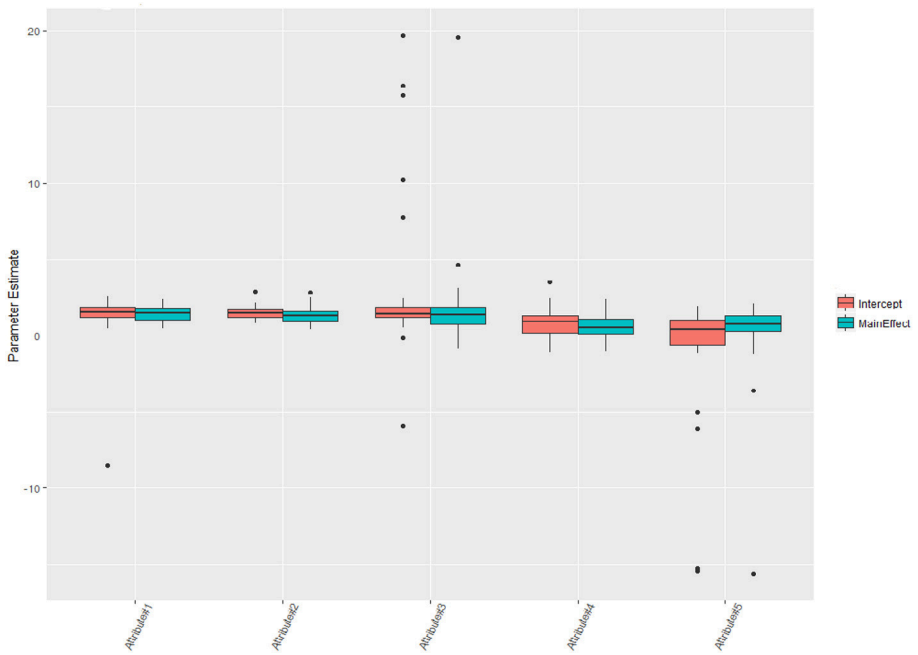


FIGURE 4 | Item parameter estimates of the health profession test.

prior had forced the assignment on each latent class and therefore the result tended to be more different from those that were fully determined by the EM algorithm.

DISCUSSION AND CONCLUSION

The purpose of this paper is to propose a machine-learning based algorithm for the estimation of LCDMs. In particular, the proposed estimator is a combination of the EM framework and the DEoptim algorithm, which has been popular in neural networks and business analytics fields. The performance of the proposed algorithm is evaluated through a simulation study of which the results indicate that it is an appropriate option to handle LCDM estimation task. This paper, however, does not suggest that the proposed algorithm should replace the EM algorithm in practice; at the situations where the EM algorithm fails to produce estimates due to the unsuccessful derivative updates, the EM-DEoptim algorithm can be an alternative.

The proposed EM-DEoptim algorithm and the traditional EM algorithm implemented in M*plus* produced virtually identical parameter estimates, and the former seems less frequently to fail. The average computational time for M*plus* estimation with the multiple-core option is 15 min. The difference is caused by the features of the algorithms: the EM algorithm based upon Quasi-Newton and Fisher scoring updates estimates with directional steps (i.e., the iteration always leads to better solutions), while the DEoptim part is truly stochastic such that the updating procedures may be wasted. Even though the DEoptim mechanism is fundamentally less directional than Quasi-Newton and Fisher scoring, The EM-DEoptim algorithm perform cannot very similar to the EM algorithm. Theoretically, the EM-DEoptim algorithm can be many times faster than what it is now if the entire function is constructed in *C++* or *Fortran*; currently only the DEoptim is implemented in *C++* through the package *RcppDE*, where the entire algorithm is written in base *R* software scripting language. Research has shown that using compiler package with *R* often takes less than half of time executing the same function than that of without packages (e.g., Aruoba and Fernández-Villaverde, 2015). In addition, given the DEoptim algorithm is composed of basic calculation, performing the proposed algorithm in a vectorization approach and therefore with graphics processing units (GPUs) is expected to accelerate the estimations.

AUTHOR CONTRIBUTIONS

ZJ proposes the idea about integrating differential evolution optimization into the EM framework in the LCDM estimation and deploys the functional algorithm in simulation studies; WM produces both literature review and technical detail supports to this manuscript.

FUNDING

ZJ was sponsored by the Research Grants Committee from Research and Economic Development, the University of Alabama (No. RG14790).

REFERENCES

Ardia, D., Boudt, K., Carl, P., Mullen, K. M., and Peterson, B. G. (2011). Differential evolution with DEoptim. *R J.* 3, 27–34.

Aruoba, S. B., and Fernández-Villaverde, J. (2015). A comparison of programming languages in macroeconomics. *J. Econ. Dyn. Control* 58, 265–273. doi: 10.1016/j.jedc.2015.05.009

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459. doi: 10.1007/BF02293801

Bradshaw, L., and Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika* 79, 403–425. doi: 10.1007/s11336-013-9350-4

Celeux, G., Chauveau, D., and Diebolt, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Stat. Comput. Simul.* 55, 287–314. doi: 10.1080/00949659608811772

da Silva, M. A., de Oliveira, E. S., Davier, A. A., and Bazán, J. L. (2017). Estimating the DINA model parameters using the No-U-Turn Sampler. *Biom. J.* 60, 352–368. doi: 10.1002/bimj.201600225

Das, S., Abraham, A., and Konar, A. (2008). Automatic clustering using an improved differential evolution algorithm. *IEEE Trans. Syst. Man Cybernet. A Syst. Hum.* 38, 218–237. doi: 10.1109/TSMCA.2007.909595

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* 56, 463–474. doi: 10.1093/biomet/56.3.463

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

DiBello, L. V., Roussos, L. A., and Stout, W. F. (2007). "Review of cognitively diagnostic assessment and a summary of psychometric models," in *Handbook of Statistics*, Vol. 26, eds C. R. Rao and S. Sinharay (Amsterdam: Elsevier), 979–1030.

George, A. C., Robitzsch, A., Kiefer, T., Gro,ß, J., and Ünlü, A. (2016). The R package CDM for cognitive diagnostic models. *J. Stat. Softw.* 74, 1–24. doi: 10.18637/jss.v074.i02

Hartz, S. M. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory With Practicality*. Ph.D. Doctoral dissertation, ProQuest Information & Learning.

Henson, R. A., Templin, J. L., and Willse, J. T. (2009). Defining a family of cognitive diagnostic models using log-linear models with latent variables. *Psychometrika* 74:191. doi: 10.1007/s11336-008-9089-5

Hu, J., Miller, M. D., Huggins-Manley, A. C., and Chen, Y. H. (2016). Evaluation of model fit in cognitive diagnostic models. *Int. J. Test.* 16, 119–141. doi: 10.1080/15305058.2015.1133627

Jiang, S., Wang, C., and Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Front. Psychol.* 7:109. doi: 10.3389/fpsyg.2016.00109

Jiang, Z., and Carter, R. (2018a). Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnostic model via Stan. *Behav. Res. Methods*, 1–12. doi: 10.3758/s13428-018-1069-9

Jiang, Z., and Carter, R. (2018b). Visualizing library data interactively: two demonstrations using R language. *Library Hi Tech. News.* 35, 14–17. doi: 10.1108/LHTN-01-2018-0003

Jiang, Z., and Raymond, M. (2018). The use of multivariate generalizability theory to evaluate the quality of subscores. *Appl. Psychol. Meas.* doi: 10.1177/0146621618758698. [Epub ahead of print].

Jin, C., Zhang, Y., Balakrishnan, S., Wainwright, M. J., and 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 (Barcelona), 4116–4124. Available online at: https://papers.nips.cc/book/advances-in-neural-information-processing-systems-29-2016

Junker, B. W., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Lao, H., and Templin, J. (2016). "Estimation of diagnostic classification models without constraints: issues with class label switching, in *Paper Presented at Annual Meeting of the National Council on Measurement in Education* (Washington, DC: District of Columbia).

Li, H., Hunter, V., and Lei, P. (2016). The selection of cognitive diagnostic models for a reading comprehension test. *Lang. Test.* 33, 391–409. doi: 10.1177/0265532215590848

Ma, W., and de la Torre, J. (2018). *GDINA: The Generalized DINA Model Framework [Computer Software Version 2.1]*. Available online at: https://CRAN.R-project.org/package=GDINA

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement* 11, 71–137. doi: 10.1080/15366367.2013.831680

Maydeu-Olivares, A., and Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivar. Behav. Res.* 49, 305–328. doi: 10.1080/00273171.2014.911075

Mohamed, A. W., Sabry, H. Z., and Khorshid, M. (2012). An alternative differential evolution algorithm for global optimization. *J. Adv. Res.* 3, 149–165. doi: 10.1016/j.jare.2011.06.004

Muthén, L. K., and Muthén, B. O. (2013). *Mplus User's Guide (Version 6.1)[Computer Software and Manual]*. Los Angeles, CA: Muthén & Muthén.

Orlando, M., and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Appl. Psychol. Meas.* 24, 50–64. doi: 10.1177/01466216000241003

Paterlini, S., and Krink, T. (2006). Differential evolution and particle swarm optimisation in partitional clustering. *Comput. Stat. Data Anal.* 50, 1220–1247. doi: 10.1016/j.csda.2004.12.004

Rabinowitz, A. R., Fisher, A. J., and Arnett, P. A. (2011). Neurovegetative symptoms in patients with multiple sclerosis: fatigue, not depression. *J. Int. Neuropsychol. Soc.* 17, 46–55. doi: 10.1017/S1355617710001141

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. Available online at: http://www.Rproject.org/

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

Rocca, P., Oliveri, G., and Massa, A. (2011). Differential evolution as applied to electromagnetics. *IEEE Antennas Propagat. Mag.* 53, 38–49. doi: 10.1109/MAP.2011.5773566

Rupp, A., Templin, J., and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford Press.

Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., and Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnostic modeling. *Appl. Psychol. Meas.* 41, 614–631. doi: 10.1177/0146621617707510

Storn, R., and Price, K. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optimiz.* 11, 341–359. doi: 10.1023/A:1008202821328

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconceptions based on item response theory. *J. Educ. Measur.* 20, 345–354. doi: 10.1111/j.1745-3984.1983.tb00212.x

Templin, J., and Bradshaw, L. (2014). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika* 79, 317–339. doi: 10.1007/s11336-013-9362-0

Templin, J., and Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educ. Measur. Issues Prac.* 32, 37–50. doi: 10.1111/emip.12010

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnostic models.*Psychol. Methods* 11:287. doi: 10.1037/1082-989X.11.3.287

Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley.

Wolfe, J. H. (1970). Profile clustering by multivariate mixture analysis. *Multivariate Behav. Res.* 5, 329–350. doi: 10.1207/s15327906mbr0503_6

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.