Psychological Test and Assessment Modeling, Volume 61, 2019 (2), 187-205

Estimation of the parameters of the reduced RUM model by simulated annealing

Youn Seon Lim[1], *Fritz Drasgow* [2] *& Justin Kern*[3]

Abstract

In this study, a simulation-based method for computing joint maximum likelihood estimates of the reduced reparameterized unified model parameters is proposed. The central theme of the approach is to reduce the complexity of models to focus on their most critical elements. In particular, an approach analogous to joint maximum likelihood estimation is taken, and the latent attribute vectors are regarded as structural parameters, not parameters to be removed by integration with this approach, the joint distribution of the latent attributes does not have to be specified, which reduces the number of parameters in the model.

Keywords: cognitive diagnosis model, reduced reparameterized unified model, Markov Chain Monte Carlo, joint maximum likelihood, simulated annealing

[1] *Correspondence concerning this article should be addressed to:* Address of the first Author: Department of Science Education, Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, 500 Hofstra University, Hempstead, NY 11549, USA ; email: YounSeon.Lim@hofstra.edu

[2] School of Labor & Employment Relations, Department of Psychology, University of Illinois at Urbana-Champaign, 504 E. Armory Avenue, Champaign IL 61820, USA

[3] Department of Educational Psychology, University of Illinois at Urbana-Champaign, 1310 S. Sixth St., Champaign IL 61820, USA

Introduction

Cognitive diagnostic models (CDMs) provide diagnostic information about each examinee's abilities which could lead to finer classification and more efficient remediation. Many models for cognitive diagnosis have been introduced over the past decades. A common feature of the model is a matrix that shows the relationship between latent attributes (i.e., knowledge and skills) and each item's response, usually referred to as the Q-matrix. Depending on the relationship, cognitive diagnostic models are classified into three categories: conjunctive, compensatory, and disjunctive (e.g., Henson, Templin, & Willse, 2009).

Some examples of the conjunctive models are the deterministic input, noisy and-gate (DINA) model, the noisy inputs, deterministic and-gate (NIDA) model (e.g. de la Torre, 2009; Junker and Sijtsma, 2001), and the reduced reparameterized unified model (reduced RUM, Hartz, 2002, Hartz & Roussos, 2008). A well-known disjunctive model is the deterministic input, noisy or-gate (DINO) model (see Templin & Henson, 2006). A commonly used compensatory model is the compensatory RUM (see Rupp, Templin, & Henson, 2010). Of these models, the reduced RUM has received a considerable amount of scholarly attention because it includes a greater variety of components than other models, thus giving more flexibility (Hartz & Roussos, 2008).

A variety of possible methods for estimating the reduced RUM have been explored. One classical approach has treated the person parameters as nuisance parameters and simply integrated them out of the likelihood equation. This approach, called marginal maximum likelihood, is thus a function of only the structural (i.e., item) parameters. When the marginal likelihood is evaluated, the Expectation-Maximization (EM) algorithm has been used to some effect (e.g., Feng, Habing, & Huebner, 2014). However, it is sensitive to initial values and can have local maxima (e.g., Doucet, Godsill, & Robert, 2002). Furthermore, it can be a computational burden to deal with high-dimensional integration in the EM which limits the practicality of this approach. In a simulation study, Drasgow (1989) found that marginal maximum likelihood estimates are more accurate than joint maximum likelihood estimates regardless of sample size or test length. However, obtaining marginal maximum likelihood estimates is a complex task because, in some cases, the likelihood function for the structural parameters is not available in closed form and, moreover, may be multimodal (e.g., Doucet, Godsill, & Robert, 2002).

Another approach to estimation is Bayesian estimation of the parameters using prior distributions on the person parameter, or on both person and item parameters (see Hartz, 2002; Henson, Templin, & Willse, 2009; Rupp, Templin, & Henson, 2010). This procedure eliminates the problems sometimes encountered in the marginal maximum likelihood estimation (e.g. Hambleton, Swaminathan, & Rogers, 1991). This approach, however, has its own difficulties. For instance, the prior specification and prior sensitivity are important aspects of Bayesian inferences (e.g., Ghosh, Ghosh, Chen, & Agresti, 2000). In practice, it can be difficult to give a meaningful full prior specification for the reduced RUM, especially as the number of attributes increases. Furthermore, in the Bayesian

framework, the homogenous Markov Chain Monte Carlo (MCMC) methods typically used for the estimation of model parameters are inefficient for maximum a posteriori estimation because a large amount of the computational burden is spent exploring regions of low posterior probability (e.g., Andrieu & Doucet, 2000); for the reduced RUM, MCMC may be prohibitively slow to converge for its complexity. Finally, MCMC methods are often more suited for integration, not optimization problems (e.g., Jacquier, Johnnes, & Polson, 2007).

Several alternative strategies for estimating the reduced RUM parameters, which include the use of simulated annealing (Lim & Drasgow, 2016) and data augmentation (Culpepper & Hudson, 2017) have been proposed. The method of Joint Maximum Likelihood Estimation (JMLE) is another procedure that can be used for the reduced RUM. A JMLE method for cognitive diagnosis models has been proposed (Chiu, Kohen, Zheng, & Henson, 2016). In the method, Birnbaum's paradigm (Birnbaum 1968) was implemented: person parameters and item parameters are considered as two separate sets where one is assumed to be known and the other is estimated.

Neyman and Scott (1948) showed that when the number of structural parameters increases with the number of incidental parameters, estimates may not be consistent. Even when the estimates of structural parameters are consistent, the property of efficiency may not hold. Lord's JMLE procedure (1968) is an example of the situation described by Neyman and Scott, and the consistency of the structural (i.e., item) parameter estimates has been questionable. In the special case of the Rasch model, however, Haberman (1997) proved the joint consistency of maximum likelihood estimates of item and person parameters, obtaining strong consistent estimates of the parameters as the number of items and examinees both go to infinity. Douglas (1997) also proved uniform asymptotic consistency in a unidimensional class of kernel-smoothing-based nonparametric IRT item response function estimation procedures under less restrictive assumptions than Haberman's. Empirical results obtained by Lord (1975) and by Swaminathan and Gifford (1983), for example, showed that the JMLE procedure can give accurate results with as few examinees as $I = 200$ provided the number of items $J \geq 60$. Hulin, Lissak, and Drasgow (1982) conducted a Monte Carlo study to investigate the effects of four sample sizes ($I = 200$, 500, 1000, or 2000 examinees) and three test lengths ($J = 15$, 30, or 60 items) on the accuracy of joint parameter estimation. They found that, for a two-parameter model, there must be at least $J = 30$ and $I = 500$, and for a 3-parameter model, there must be at least $J = 60$ and $I = 1000$.

In this study, we propose a JMLE approach for the reduced RUM, which simultaneously evaluates and optimizes the joint likelihood function for the reduced RUM. This JMLE approach is carried out by means of a combination of the simulated annealing algorithm and stochastic simulation of the hidden Markov chain. The central theme of the approach is to omit variables related to the joint distribution of latent attributes to trim back model complexity. This algorithm is shown to converge for the set of joint maximum likelihood parameter estimates under suitable regularity conditions.

Reduced RUM specification

Let Y_{ij} denote the binary item response of the ith examinee to the jth item, $i = 1, ..., I$, $j = 1, ..., J$, with 1 = correct and 0 = incorrect. Each entry α_{ik} indicates whether the ith examinee has mastered the kth attribute (i.e., the kth knowledge or skill), with $\alpha_{ik} = 1$ indicating examinee i has mastered attribute k and 0 otherwise. The $J \times K$ Q-matrix describes how the items are related to the attributes. Each entry q_{jk} in the matrix indicates whether the kth attribute is relevant for the solution of the jth item: $q_{jk} = 1$ if the attribute is germane, 0 if not (Tatsuoka, 1983).

The reduced RUM is the most flexible of the aforementioned conjunctive models. The item response function in the reduced RUM is

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, \boldsymbol{\pi}^*, \boldsymbol{r}) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{(1-\alpha_k)q_{jk}}, \tag{1}$$

where $0 < \pi_j^* < 1$ denotes the probability of a correct response to the jth item for the ith examinee who possesses and correctly applies all attributes required for that item, and $0 < r^{q_{jk}} < 1$ indicates the penalty to the probability of correct response to item j for not having mastered the kth attributes.

More formally, the reduced RUM is simply a reparameterization of the generalized Noisy Inputs, Deterministic and-gate (NIDA) model (e.g., Junker & Sijtsma, 2001), where $P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, s_j, g_j) = \prod_{k=1}^{K}[(1 - s_{jk})^{\alpha_k} g_{jk}^{1-\alpha_k}]^{q_{jk}}$ and s_{jk} (the slipping parameter) is the probability of the examinee failing to correctly apply attribute k in solving item j, given the examinee has mastered all attribute k, and g_{ik} (the guessing parameter) is the probability of the examinee correctly applying all attribute k in solving item j, given the examinee has not mastered attribute k. In this light, π_j^* can be reinterpreted as an item difficulty parameter and r_{jk} can be viewed as an item discrimination parameter for attribute k. We can see this by setting,

$$\pi_j^* = \prod_{k=1}^{K} \pi_{jk}^{q_{jk}} = \prod_{k=1}^{K} (1 - s_{jk})^{q_{jk}}, \tag{2}$$

where the slipping parameter for item j and attribute k is denoted as s_{jk}.

$$r_{jk} = \frac{g_{jk}}{1 - s_{jk}}, \tag{3}$$

where the guessing parameter for item j and attribute k is denoted as g_{jk}.

Algorithm and properties

Under the assumptions of conditional independence, the joint likelihood L for the item responses is

$$L = L(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{I} \prod_{j=1}^{J} P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j)^{Y_{ij}} [1 - P(Y_{ij} = 1|\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j)]^{1-Y_{ij}}, \quad (4)$$

where $\boldsymbol{\beta}_j = \{\beta_{jk}\}$ denotes the item parameters for item j. The item parameters $\boldsymbol{\beta}_j$ as well as the person parameters $\boldsymbol{\alpha}_i$ are required to be estimated at the same time. The values of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_j$ that maximize the likelihood L are the joint maximum likelihood estimates. One approach is to estimate the values of parameters directly by iteratively setting $\frac{\partial L}{\partial \boldsymbol{\alpha}_i} = 0$, and $\frac{\partial L}{\partial \boldsymbol{\beta}_j} = 0$ (e.g., Lord, 1968). However, some difficulties can be encountered. First, there are some cases in which the maximum likelihood estimates or the likelihood in closed form do not exist (e.g., Hambleton, Swaminathan, & Rogers, 1991). Second, it is a computational burden to iterate between the two sets of partial derivatives; moreover, the numerical optimization on very high dimensional models is time consuming as shown in previous studies. Finally, it is challenging to estimate the standard errors of the maximum likelihood estimates based on the second order derivatives (e.g., Jacquier, Johnnes, & Polson, 2007).

To avoid these potential problems, the current approach modifies the JMLE method in three important ways. One is to implement a regularization term for all model parameters. This is accomplished by establishing uniform (flat) prior distributions, and then obtaining the maximum a posteriori (MAP) values of the parameters. The assumption of flat priors for the parameters means that the prior terms for those parameters can be set to unity, and therefore the MAP updates for $\hat{\boldsymbol{\alpha}}_i$ and $\hat{\boldsymbol{\beta}}_j$ are identical to the maximum likelihood updates for the parameters (e.g., Patz & Junker, 1999) on bounded intervals.

Second, rather than estimate the distribution of $\boldsymbol{\alpha}_i$, each α_{ik} is treated as a parameter to be estimated. This has been problematic in IRT models in which the latent variables are continuous, because something must be done to fix the scale. However, for cognitive diagnosis models in which the latent attributes are binary, the scale is solidly pinned down between the two possible values (0 or 1) in the parameter space. This results in a more streamlined model and yields simpler Markov chains and consistent results, as shown in a later section of this paper.

Third, we propose an algorithm that is a combination of the insights of standard MCMC algorithms and simulated annealing algorithms in the Bayesian framework. The initial value of this algorithm is obtained from the nonparametric estimator of latent attribute variables (Lim & Drasgow, 2017). Given the estimates of $\boldsymbol{\alpha}$, the item parameters are estimated, and then the estimates of item parameters are used to update the estimates of $\boldsymbol{\alpha}$. This procedure is repeated until the convergence criterion is satisfied.

Simulated annealing is an inhomogeneous variant of MCMC used to perform combinational optimization. This method samples from a sequence of density functions whose

support concentrates itself on the set of maximum likelihood estimates. The power $\gamma(t), t = 1, ..., T$, called the temperature, makes it possible to explore the entire search space systematically by being increased simultaneously as the number of Markov chain iterations increases (e.g., van Laarhoven & Arts, 1989). As in simulated annealing, this proposed algorithm replaces the target joint density $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as

$$\pi_{\gamma(t)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(\boldsymbol{\alpha}, \boldsymbol{\beta})^{\gamma(t)} P(\boldsymbol{\alpha}) P(\boldsymbol{\beta}), \tag{5}$$

where $\lim_{t \to +\infty} \gamma(t) = \infty$. When $\gamma(t) > 1$, $P(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is raised to the $\gamma(t)$ power and the effects of the priors $P(\boldsymbol{\alpha})$ and $P(\boldsymbol{\beta})$ disappear on the range of values (e.g., Jacquier, Johnnes, & Polson, 2007). Nonetheless, they are necessary to ensure their integrability without affecting the maximum joint likelihood estimates.

Formally, the proposed algorithm seeks to maximize the joint likelihood in the Bayesian framework with a constant temperature $\gamma(T)$,

$$\pi_{\gamma(T)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto L^{\gamma(T)} P(\boldsymbol{\alpha}) P(\boldsymbol{\beta}). \tag{6}$$

The likelihood term L reappears in this Bayesian formulation, but is now accompanied by the uninformative prior distributions of the parameters. As it is usually impossible to sample from the density directly, MCMC methods are used to simulate samples from a sequence of joint densities, $\pi_{\gamma(T),n}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, where n indexes the length of the Markov Chain.

It is important to compare and contrast this algorithm with the marginal maximum likelihood (or marginal MAP) estimation methods related to simulated annealing. The basic idea is to generate a sequence of artificial distributions from a density in which the latent variables are replicated temperature $\gamma(t)$ times by data augmentation. Then the sequence concentrates itself on the set of marginal maxima. For generation, non-homogenous MCMC algorithms (Andrieu & Doucet, 2000; Doucet, Godsill, & Robert, 2002), original sequential Monte Carlo methods (Johansen, Doucet, & Davy, 2008), and a standard evolutionary MCMC method (Jacquier, Johannes, & Polson, 2007) have been employed. These researchers advocate that as the chain goes to infinity, the sequence of density concentrates itself upon the marginal maximum of structural parameters. Then the estimates of structural parameters are obtained without resorting to a gradient based method. Temperature $\gamma(t)$ is assumed to be increased as the chain increases, especially in terms of the theoretical foundation.

In contrast, our algorithm estimates the joint maximum likelihood in the Bayesian framework. The joint density is alternately raised to $\gamma(T)$ as in simulated annealing while the priors are not exponentiated unlike simulated annealing. The initial values of this algorithm is obtained from a nonparametric approach. The estimates for the values of parameters are obtained given the estimates of the other parameters. Furthermore, the joint distribution of binary latent variables does not need to be estimated because each component is regarded as an individual parameter. Unlike the algorithm presented here,

the methodologies for marginal maximum likelihood (or marginal maximum a posterior) require or are suitable only for continuous latent variable models.

To summarize, several practical advantages of this algorithm can be gleaned. First, this approach does not require estimating the distribution of $\boldsymbol{\alpha}$ as if it is treated simply as a parameter to be estimated. Second, unlike a Bayesian approach, informative prior distributions for the parameters are not necessary. Third, the MAP estimates (which are equivalent to ML estimates given flat priors) are obtained without exploring regions of low posterior probability or integrating over the incidental parameters. Finally, in combination with the standard MCMC algorithm, simulated annealing maintains the speed and reliability of gradient descent algorithms while simultaneously avoiding local minima (Zomaya & Kazman, 2010).

Convergence properties of the algorithm

Suppose that a function f is defined on a finite set of states S. The purpose of this proposed algorithm is to find a state $\boldsymbol{x} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ such that $f(\boldsymbol{x}) = \max_{\boldsymbol{z} \in S} f(\boldsymbol{z})$. Let $G(\boldsymbol{x}, \boldsymbol{z})$ denote the generation probability, and $\boldsymbol{x}, \boldsymbol{z} \in S$ such that $G(\boldsymbol{x}, \boldsymbol{z}) > 0$. The acceptance probability $A(\boldsymbol{x}, \boldsymbol{z})$ is defined by (13) and (15) below. Let $\boldsymbol{P}^{\gamma(T)}$ denote the transition matrix corresponding to a temperature $\gamma(T)$, $t = 1, 2, ..., T = T_{\text{opt}} < \infty$. Each component $P^{\gamma(T)}(\boldsymbol{x}, \boldsymbol{z})$ of $\boldsymbol{P}^{\gamma(T)}$ can be defined as $G(\boldsymbol{x}, \boldsymbol{z})A(\boldsymbol{x}, \boldsymbol{z})$ for $\boldsymbol{z} \neq \boldsymbol{x}$, otherwise, $1 - P^{\gamma(T)}(\boldsymbol{x}, \boldsymbol{z})$ for $\boldsymbol{z} = \boldsymbol{x}$.

The invariant joint density $\pi_{\gamma(T)}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be rewritten as (8) below like the objective function used in simulated annealing:

$$\pi_{\gamma(T)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(\boldsymbol{\alpha})P(\boldsymbol{\beta}) \exp(\gamma(T) \times \log(L)). \tag{7}$$

Then for the Markov chain $n \to \infty$, the stationary distribution of the time-inhomogeneous Markov chains is obtained as the result of van Laarhoven and Aarts (1989),

Let $S^{\star}$ denote the global maxima set of states. Then

$$\lim_{\gamma(t) \to \infty} \lim_{n \to \infty} P(f(\boldsymbol{x})_{t,n} \in S^{\star}) = 1. \tag{8}$$

Proof See van Laarhoven and Aarts (1989).

Theorem 1 essentially governs the behavior of the SA algorithm. Under certain conditions on the matrices $G(\boldsymbol{x}, \boldsymbol{z})A(\boldsymbol{x}, \boldsymbol{z})$, the SA algorithm converges to a global maximum with probability 1 as temperature goes to ∞ with Markov chain increases.

Many studies have been conducted to determine the optimal temperature $\gamma(T)$ for the efficiency of simulated annealing. Typically an optimal temperature $\gamma(T)$ is experimentally chosen by running simulated annealing until the temperature is frozen at a particular value (e.g., Cohn & Fielding, 1999); by running homogeneous Markov chains at each of a number of fixed temperatures, it is possible to determine which temperature is best, according to an appropriate optimality criterion (e.g., Duong-Ba, Nguyen, & Bose, 2014).

This proposed algorithm uses the optimal $\gamma(T)$ as a fixed temperature $\gamma(T)_{(n+1)} = \gamma(T)_n$. Then, unlike simulated annealing, this proposed algorithm exhibits as a time-homogenous Markov chain at the height of max $f(\boldsymbol{x})$. The temperature which is reached to a frozen status is considered as the optimal $\gamma(T)$ and it is estimated as a parameter in this proposed algorithm. The convergence property of a homogeneous Markov chain has been well investigated in Feller (1950) and others. The homogeneous Markov chain generated by $\boldsymbol{P}^{\gamma(T)}$ has a stationary distribution. Then

$$\lim_{n\to\infty} P\{\max f(\boldsymbol{x})_n \in S^\star\} = 1. \tag{9}$$

Proof See Mitra, Romeo, and Sangiovanni-Vincentelli (1985).

MCMC algorithm

For the proposed approach, the Metropolis-Hastings algorithm with simulated annealing is used for sampling from $\pi_{\gamma(T)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto P(\boldsymbol{\alpha}, \boldsymbol{\beta})^{\gamma(T)} P(\boldsymbol{\alpha}) P(\boldsymbol{\beta})$. Like Birnbaum's two stage paradigm (Birnbaum, 1968), this algorithm starts with the estimated initial values of latent attribute variable $\boldsymbol{\alpha}$ by using a nonparametric technique proposed by Lim and Drasgow (2017). In this approach, the uniform prior distributions are established over the parameters.

Step 1. Estimate the initial value of this algorithm: person parameter $\boldsymbol{\alpha}_{(0)}$

A nonparametric method (Lim & Drasgow, 2017) is used to estimate the initial value of this algorithm. The method estimates the person parameter $\boldsymbol{\alpha}_{(0)}$ based on the Hamming distance between ideal and observed response patterns. This approach consists of two phases, the computation of all possible ideal response vectors and the classification phase.

The ideal responses $\eta_{ij\text{noncompensatory}}$ are defined as $\prod_{k=1}^{K} \alpha_{ik}{}^{q_{jk}}$, $\eta_{ij\text{ disjunctive}}$ are defined as $1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$, and $\eta_{ij\text{compensatory}}$ are defined as rounding of $\sum_{k=1}^{K} (\hat{\alpha}_{ik} \times q_{j_{\text{new}}ck})/K$ for examinee i and assessment item j. All possible ideal response vectors $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, ...\boldsymbol{\eta}_{2^K}$ are constructed from all 2^K possible patterns for $\boldsymbol{\alpha}_i$. In the classification stage, the Hamming distances between Y_i and each of $\boldsymbol{\eta}_m$, for $m = 1, 2, ...2^K$, are computed by simply counting the number of times two vectors disagree as given by

$$D(\boldsymbol{Y}_i, \boldsymbol{\alpha}_m) = \sum_{j=1}^{J} \mid Y_{ij} - \eta_{mj} \mid . \tag{10}$$

The estimator is obtained by minimizing this distance over all possible attribute patterns,

$$\hat{\boldsymbol{\alpha}}_i = \arg \min_{m \in \{1,2,...,2^K\}} D(\boldsymbol{Y}_i, \boldsymbol{\alpha}_m). \tag{11}$$

The theoretical justification is that the true attribute vector minimizes the expected distance between $\boldsymbol{Y}_i$ and $\boldsymbol{\eta}_m$, under some general conditions on the underlying model.

Step 2. Draw $\boldsymbol{\beta}^{\gamma(T)}_{\{n=1,2,...,N\}}|\boldsymbol{\alpha}_{(0)} \sim P(\boldsymbol{\beta}^{\gamma(T)}|\boldsymbol{\alpha}_{(0)}, \boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}^{\gamma(T)})P(\boldsymbol{\beta})$.

Given the estimates of person parameter $\boldsymbol{\alpha}$, the Markov chains of item parameter $\boldsymbol{\beta} = \{\pi^*_j, r_{jk}\}$ are obtained until no values are updated. Here $\boldsymbol{\beta}^{\gamma}(T)$ is considered as the $\gamma(t), t = 1, ..., T$ independent copies of $\boldsymbol{\beta}$. That is,

$$P(\boldsymbol{\beta}^{\gamma(T)}|\boldsymbol{\alpha}, \boldsymbol{Y}) \propto \prod_{t=1}^{T}(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\beta}^{\gamma(t)})P(\boldsymbol{\beta}). \tag{12}$$

Instead of generating $\gamma(T)$ copies, simulated annealing is used for this algorithm. The simulated annealing provides additional flexibility and efficiency in generating $\gamma(T)$ copies of item parameters $\boldsymbol{\beta}$. More specifically, (12) is obtained by

$$(\boldsymbol{\beta}^{\gamma(T)}_{(n+1)}, \boldsymbol{\beta}^{\gamma(T)}_{(n)}) = \min\{1, \exp(\gamma(T) \times (\log P(\boldsymbol{\beta}_{(n+1)}|\boldsymbol{\alpha}, \boldsymbol{Y})) - \log(P(\boldsymbol{\beta}_{(n)}|\boldsymbol{\alpha}, \boldsymbol{Y})))\} \tag{13}$$

Updating π^*_j *for* $j = 1, 2, ..., J$.

A candidate value $\pi^{*\star}_j$ is drawn from the uniform distribution on the interval $(\pi^{*l} - \delta, \pi^{*h} + \delta)$, where π^{*l} and π^{*h} are, respectively, the lower bound and the higher bound of baseline parameters. The π^{*l} and π^{*h} should be no smaller than .5 and no greater than 1, respectively because there is an order constraint such that r_{jk} is smaller than π^*_j (Rupp, Templin, & Henson, 2010). In the studies below, $\delta = .1$ was used in the following simulation studies. Compute the acceptance probability,

$$r_n = \exp(\gamma(T) \times (\log(L(\boldsymbol{Y}_j|\boldsymbol{\alpha}^{(n-1)}, \pi^{*\star}_j, r^{(n-1)}_{jk})) - \log(L(\boldsymbol{Y}_j|\boldsymbol{\alpha}^{(n-1)}, \pi^{*(n-1)}_j, r^{(n-1)}_{jk})))). \tag{14}$$

Let $\pi^{*(n)}_j = \pi^{*\star}_j$ with probability $\min(1, r_n)$, otherwise let $\pi^{*(n)}_j = \pi^{*(n-1)}_j$.

Updating r_{jk} *for* $j = 1, 2, ..., J$, *for* $k = 1, 2, ..., K$

A candidate value for the penalty parameter $r^{\star}_{jk}$ is drawn from the uniform distribution. The lower boundary of the r_{jk} should be 0 and the higher boundary of this parameter is 1. Compute

$$r_n = \exp(\gamma(T) \times (\log(L(\mathbf{Y}_j|\boldsymbol{\alpha}^{(n-1)}, \pi^{*(n-1)}_j, r^{\star}_{jk})) - \log(L(\boldsymbol{Y}_j|\boldsymbol{\alpha}^{(n-1)}, \pi^{*(n-1)}_j, r^{(n-1)}_{jk})))). \tag{15}$$

Let $r^{(n)}_{jk} = r^{\star}_{jk}$ with probability $\min(1, r_n)$, otherwise let $r^{(n)}_{jk} = r^{(n-1)}_{jk}$.

Step 3. Draw $\alpha_{ik,\{1,2,...,N\}}^{\gamma(T)}|\boldsymbol{\beta}_{(0)} \sim P(\alpha_{ik}^{\gamma(T)}|\boldsymbol{\beta}_{(0)}, \boldsymbol{Y}) \propto P(\boldsymbol{Y}|\alpha_{ik}^{\gamma(T)}, \boldsymbol{\beta})P(\alpha_{ik})$.

Now given the estimated item parameters $\boldsymbol{\beta}$ from the previous step, the estimates of person parameter $\boldsymbol{\alpha}$ are updated. The draws of the person parameter $\boldsymbol{\alpha}$ are generated until no values are updated during an iteration like Hartz (2002). The independent draws of each $\gamma(t), t = 1, ..., T$ α_{ik} are,

$$P(\alpha_{ik}^{\gamma(T)}|\boldsymbol{\beta}, \boldsymbol{Y}) \propto \prod_{t=1}^{T} (\boldsymbol{Y}|\alpha_{ik}^{\gamma(t)}, \boldsymbol{\beta})P(\alpha_{ik}). \tag{16}$$

This is obtained by

$$(\alpha_{ik,(n+1)}^{\gamma(T)}, \alpha_{ik,(n)}^{\gamma(T)}) = \min\left\{1, \exp(\gamma(T) \times (\log(P(\alpha_{ik,(n+1)}|\boldsymbol{\beta}, \boldsymbol{Y})) - \log(P(\alpha_{ik,(n)}|\boldsymbol{\beta}, \boldsymbol{Y})))\right\} \tag{17}$$

Updating α_{ik} *for* $i = 1, 2, ..., I$, $k = 1, 2, ..., K$

For $\alpha_{ik}^{\star}$ in $\boldsymbol{\alpha}_i$, a candidate value is drawn from the binomial distribution $(1, .5)$. Compute the acceptable probability,

$$r_{n'} = \exp(\gamma(T) \times (\log(L(\boldsymbol{Y}_i|\alpha_{ik}^{\star}, \boldsymbol{pi}^{*(n'-1)}, \boldsymbol{r}^{(n'-1)})) - \log(L(\boldsymbol{Y}_i|\boldsymbol{\alpha}_i^{(n'-1)}, \boldsymbol{\pi}^{*(n'-1)}, \boldsymbol{r}^{(n'-1)})))). \tag{18}$$

Let $\alpha_{ik}^{(n')} = \alpha_{ik}^{\star}$ with probability $\min(1, r_{n'})$, otherwise let $\alpha_{ik}^{(n')} = \alpha_{ik}^{(n'-1)}$.

Note that flat prior distributions are used throughout. Steps 2 and 3 are repeated until a stopping criterion is met. As in simulated annealing, the stop criterion is either determined by fixing the number of temperature schedule values, or by terminating operation of the algorithm if the Markov chains are identical for a number of chains (e.g., van Laarhoven & Aarts, 1989).

Determining $\gamma(T)$

In the typical implementation of the standard simulated annealing algorithm, the temperature increases with the Markov chains, or is constant for a finite length Markov chain at each increasing temperature schedule. The weakness of this temperature schedule is very slow convergence, especially if the cost function is expensive to compute; for situations where there are few local minima, simpler methods work better. For these reasons, in this study, other than the standard temperature schedule, a constant temperature $\gamma(T) =$ 1, 5, 10, or 20 is proposed as the Temperature schedule. In practice, we suggest that a constant temperature is estimated as a parameter. Like the standard temperature schedule, the temperature is determined by a certain temperature in which the search converges to a frozen state (e.g., van Laarhoven & Aarts, 1989).

Simulation study

A simulation study was carried out to evaluate the performance of the proposed MCMC algorithm under various conditions. Four conditions of item length J (short = 25, long = 50) and size of sample I (small = 250, large = 1000) were considered. A Q-matrix for $J = 25$ was randomly generated from $2^K - 1$ possible q-vectors as presented in Table 1. The Q-matrix for J = 50 was obtained by duplicating the matrix two times.

Table 1:
Correctly Specified Q (K = 7).

Item	K = 7							Item	K = 7						
1	0	0	1	1	1	1	1	14	1	1	0	0	0	0	0
2	0	0	1	0	1	1	0	15	1	0	1	0	0	0	1
3	1	0	0	1	0	0	0	16	1	1	0	0	0	1	1
4	1	1	0	0	1	1	1	17	0	0	0	1	0	0	1
5	0	1	1	1	1	0	0	18	1	0	1	0	0	0	0
6	1	0	0	0	1	0	1	19	0	0	1	0	1	0	0
7	0	0	0	1	1	1	1	20	0	1	0	0	1	1	0
8	0	0	1	1	0	0	1	21	0	0	0	0	0	0	1
9	1	1	0	1	0	0	1	22	1	0	1	0	1	1	1
10	0	0	1	0	1	1	0	23	0	1	0	0	0	0	1
11	0	0	1	0	1	0	1	24	1	1	1	0	1	1	1
12	1	0	0	0	0	0	1	25	1	1	1	0	1	1	1
13	0	1	1	0	1	1	0								

Each item response data set was generated from the reduced RUM model. The person parameters $\boldsymbol{\alpha}$ were sampled from the bivariate Normal distribution with mean vector $\mu = (0, 0, 0, 0, 0, 0, 0)$ and covariance matrix $\sum$ with 1's on the diagonal and off-diagonal elements of .3. Binary traits were constructed as in previous studies (e.g., Lim & Drasgow, 2017).

$$\alpha_{ik} = \begin{cases} 1, & \text{if } \theta_{ik} \geq \Phi^{-1} \frac{k}{K+1}; \\ 0, & \text{otherwise} \end{cases}$$

The base item parameter π_j^* was sampled from the uniform distribution (.6, 1), and the attribute-level penalty parameter r_{jk} was sampled from a uniform distribution(0, .5). The proposed MCMC algorithm was run with four different values of $\gamma(T)$: 1, 5, 10, and 20. The convergence of item parameters was evaluated by the criterion of normality. The convergence of person parameters was estimated indirectly by evaluating the agreement rates between the true values of α and estimated values $\hat{\alpha}$ (Hartz, 2002). The estimates were compared with the estimates obtained by the fully Bayesian model estimation software called Arpeggio (DiBello & Stout, 2010).

Results

Table 2 reports the RMSE of item parameters for four different conditions. Increasing test length J and sample size I decreased the RMSE for all parameters, and the results are as

expected. However, comparing the results of the fully Bayesian analysis (Arpeggio) with the estimates from the proposed algorithm, some differences can be seen. Specifically, the new method has lower RMSEs, particularly when $\gamma(T) = 20$.

Table 2:
RMSE of Item Parameters

Condition	Parameter	Arpeggio	$\gamma(T) = 20$	$\gamma(T) = 10$	$\gamma(T) = 5$	$\gamma(T) = 1$
$J = 25, I = 250$	Base	.097	.076	.081	.084	.110
	Penalty	.238	.134	.144	.170	.172
$J = 25, I = 1000$	Base	.043	.075	.077	.077	.089
	Penalty	.162	.100	.109	.110	.119
$J = 50, I = 250$	Base	.098	.074	.087	.092	.110
	Penalty	.177	.139	.145	.151	.152
$J = 50, I = 1000$	Base	.045	.039	.042	.052	.056
	Penalty	.117	.098	.104	.110	.109

The theoretical convergence results for simulated annealing indicate that, as $\gamma(T)$ increases, the draw will converge to the joint maximum likelihood estimate. In practice, these results would not be very useful if an inordinately high value of $\gamma(T)$ was required for the algorithm to approach the joint maximum likelihood estimate. This study empirically shows that this is not the case. This algorithm was quite effective for even the moderate values of $\gamma(T) = 10$ or 20.

Figure 1 shows the draws of the item base parameter π_j^* for the four runs of the algorithm with $\gamma(T)$ = 1, 5, 10, and 20. The horizontal red lines show the true parameter value. Each draw of item parameters is conditional on the person parameter $\boldsymbol{\alpha}$ and the penalty parameter r_{jk} As $\gamma(T)$ increases, the variance of draws decreases and the sequence of draws gets closer to the true values.

Figure 1:
Time Plot of Item 2 Base Parameter in $J = 50$, $I = 1000$

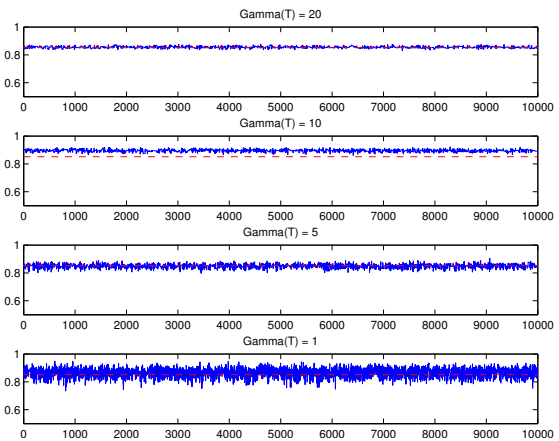
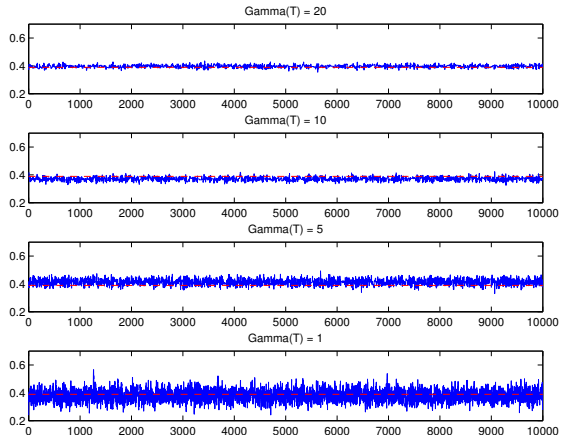


Figure 2 shows the draws of item penalty parameter r_{jk}. As in the previous plots, the horizontal lines denote the true item parameter value. One interesting finding is that the sequence of draws is closer to the true value of the parameter in the condition of $\gamma(T) = 10$ than $\gamma(T) = 20$, albeit with greater variance. This finding confirms the results of Mitra, Romeo, and Sangiovanni-Vincentelli (1986) concerning the optimal $\gamma(T)$. However, this results show the potential problem of this approach: the chain can be biased (trapped in a local maximum) although it has less variance.

Figure 2:
Time Plot of Item 2 and Attribute 6 Penalty Parameter in $J = 50$, $I = 1000$



The proportion of the times for which the true $\boldsymbol{\alpha}_i$ and the estimated $\hat{\boldsymbol{\alpha}}_i$ agreed was summarized for each condition in two different ways: one was the Component-wise Agreement Rate (CAR)= $(\sum_{i=1}\sum_{k=1}|\alpha_{i_k} = \hat{\alpha}_{i_k}|)/(I \times K)$, and the other one was the Vector-wise Agreement Rate (VAR) = $(\sum_{i=1}|\boldsymbol{\alpha}_i = \hat{\boldsymbol{\alpha}}_i|)/I$. The CAR and VAR between the true $\boldsymbol{\alpha}$ and estimated $\hat{\boldsymbol{\alpha}}$ obtained from both methods are reported in Table 3. The component-wise agreement rates of JMLE are relatively consistent and reasonably high (i.e., larger than .897) regardless of J and I unlike the results of the Fully Bayesian model. The estimates from Arpeggio for the condition of $J = 50$ and $I = 1000$ were low.

Analysis of fraction subtraction data with reduced RUM

As an illustration of the model with real data, the reduced RUM model with the proposed algorithm was fitted to the fraction subtraction data that includes the item responses to 20 items with 8 necessary attributes from 536 examinees. The data were originally collected and analyzed by Tatsuoka (1990) and have been analyzed in numerous studies (e.g., de la Torre & Douglas, 2004). Here we use the Q-matrix in Table 4 for the data

Table 3:
Agreement Rates between $\hat{\alpha}$ and α

Condition	Parameter	Arpeggio	$\gamma(T) = 20$	$\gamma(T) = 10$	$\gamma(T) = 5$	$\gamma(T) = 1$
$J = 25$, $I = 250$	CAR	.798	.899	.874	.871	.840
	VAR	.280	.399	.388	.384	.322
$J = 25$, $I = 1000$	CAR	.871	.897	.877	.872	.828
	VAR	.401	.409	.393	.394	.327
$J = 50$, $I = 250$	CAR	.936	.968	.949	.950	.947
	VAR	.648	.656	.624	.616	.632
$J = 50$, $I = 1000$	CAR	.916	.953	.913	.929	.906
	VAR	.603	.598	.574	.568	.535

that appeared in de la Torre and Douglas (2004). The specified attributes are (1) Convert a whole number to a fraction, (2) Separate a whole number from fraction, (3) Simplify before subtracting, (4) Find a common denominator, (5) Borrow from whole number part, (6) Column borrow to subtract the second numerator from the first, (7) Subtract numerators, and (8) Reduce answers to simplest form.

Table 4:
Q for the Fraction Subtraction Data.

Item	K = 8								Item	K = 8							
1	0	0	0	1	0	1	1	0	11	0	1	0	0	1	0	1	0
2	0	0	0	1	0	0	1	0	12	0	0	0	0	0	0	1	1
3	0	0	0	1	0	0	1	0	13	0	1	0	1	1	0	1	0
4	0	1	1	0	1	0	1	0	14	0	1	0	0	0	0	1	0
5	0	1	0	1	0	0	1	1	15	1	0	0	0	0	0	1	0
6	0	0	0	0	0	0	1	0	16	0	1	0	0	0	0	1	0
7	1	1	0	0	0	0	1	0	17	0	1	0	0	1	0	1	0
8	0	0	0	0	0	0	1	0	18	0	1	0	0	1	1	1	0
9	0	1	0	0	0	0	0	0	19	1	1	1	0	1	0	1	0
10	0	1	0	0	1	0	1	1	20	0	1	1	0	1	0	1	0

Unlike the simulation study, the optimal $\gamma(T)$ was empirically determined by searching the uniform simulated annealing schedule until reaching the frozen value. Parameter estimates were obtained by estimating the mode of iterations from the proposed method as in the simulation study. The results were compared with the values from the fully Bayesian MCMC algorithm obtained with the Arpeggio program (DiBello & Stout, 2010).

Results

Table 5 shows the proportion of attribute mastery and non-mastery for each attribute from both methods. The estimates from the JMLE indicate slightly higher proportions of mastery over all attributes (i.e., the overall mean difference is .051).

The base π_j^* and penalty r_{jk} item parameter estimates are reported in Table 6 below.

Table 5:
Attribute Mastery or Non-mastery Rates for Individual Attributes.

Attribute	Number of items	Mastery Proportion		Non-mastery Proportion	
		JMLE	Full	JMLE	Full
1	3	.468	.498	.532	.502
2	13	.806	.593	.194	.407
3	3	.457	.496	.543	.504
4	5	.629	.578	.371	.422
5	8	.578	.483	.422	.517
6	2	.537	.631	.463	.369
7	19	.825	.782	.175	.218
8	3	.713	.543	.287	.457
	Mean	.627	.576	.373	.424

The estimates of base parameters π_j^* obtained from both methods were similar, and their mean difference was .005. However, the values of the penalty parameter were relatively different. The estimates from the fully Bayesian model were larger than those from the proposed algorithm. The reduced RUM has an order constraint in which r_{jk} is smaller than π_j^* (Rupp, Templin, & Henson, 2010). However, the estimates $\hat{\pi}_j^*$ from the fully Bayesian model were sometimes larger than the base parameters $\hat{r}_{jk}$ (i.e., for Item 4, $\pi_4^* = .903$, but $r_{42} = .966$, $r_{47} = .933$).

Discussion

In this study, an MCMC algorithm is proposed for joint maximum likelihood estimation of parameters of the reduced RUM. This MCMC algorithm has the advantage of the standard MCMC algorithm and simulated annealing simultaneously. The significance of this approach is that it enables researchers to trim back model complexity by considering each α as an individual parameter to be estimated; thus it is possible to estimate the item parameters and person parameters simultaneously.

As expected, as $\gamma(T)$ slightly increased and the variance of draws was reduced. The estimates of each model parameter were relatively consistent regardless of the sizes of sample I and item J. This indicates that the approach is appropriate for estimation with small samples and relatively few items. Of course increasing I and J improves results. However, it is unreasonable to determine the performance of the algorithm by comparing the estimates of model parameters given each $\gamma(T) = 1, 5, 10, 20$ with their true values because as $\gamma(T)$ increases up to the optimal value, the draws will be closer to the true values (e.g., Jacquier, Johannes, & Polson, 2007).

Future research might include simulation using more attributes. In addition, the optimal temperature $\gamma(T)$ for each cognitive diagnosis model could be examined empirically. At the present time, however, the new MCMC algorithm appears to be a promising approach for joint maximum likelihood estimation of the parameters of cognitive diagnosis models.

Table 6:
Item Parameter Estimation with RRUM Model via JMLE

J	Base		Penality															
	JM	Full	JM	Full	JM	Full	JM	Full	JM	Full	JM	Full	JM	Full	JM	Full	JM	Full
1	.977	.925							.104	.723			.300	.570	.496	.597		
2	.964	.968							.366	.399					.062	.559		
3	.886	.891							.011	.129					.010	.328		
4	.910	.903			.690	.966	.562	.816			.355	.278			.528	.933		
5	.880	.894			.423	.569			.453	.643					.827	.870	.538	.677
6	.938	.951													.103	.231		
7	.902	.839	.072	.832	.517	.615									.010	.236		
8	.842	.822													.383	.575		
9	.792	.834			.078	.438												
10	.841	.861			.196	.586					.099	.965			.072	.383	.403	.437
11	.925	.933			.273	.873					.124	.143			.102	.283		
12	.971	.964													.151	.181	.314	.739
13	.698	.711			.012	.137			.136	.247	.053	.142			.010	.335		
14	.944	.960			.209	.767									.129	.878		
15	.929	.903	.145	.755											.021	.193		
16	.904	.945			.262	.676									.165	.136		
17	.875	.886			.092	.776					.091	.792			.052	.134		
18	.938	.879			.237	.824					.278	.322	.358	.652	.093	.635		
19	.893	.844	.124	.114	.013	.440	.233	.173			.350	.327			.012	.316		
20	.895	.896			.186	.838	.293	.479			.034	.359			.069	.183		
Mean	.895	.890	0.11	.567	.245	.654	.363	.490	.214	.428	.173	.416	.329	.611	.173	.420	.418	.618

References

Andrieu, C., & Doucet, A. (2000). Simulated annealing for maximum a posteriori parameter estimation of hidden markov models. *IEEE Transactions on Information Theory*, *46*, 994–1004. http://ieeexplore.ieee.org/document/841176/

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397 - 479). Reading, MA: Addison-Wesley.

Cohn, H. & M. Fielding (1999), Simulated annealing searching for an optimal temperature schedule, *SIAM Journal of Optimization*, *9*, 779-802. https://doi.org/10.1137/S1052623497329683

Chiu, C.Y., Koehn, H. F., Zheng, Y., & Henson, R. (2016). Joint maximum likelihood estimation for cognitive diagnostic models. *Psychometrika*, *81*, https://doi.org/10.1007/s11336-016-9534-9

Culpepper, S. A. & Hudson, A. (2017). An improved strategy for Bayesian estimation of the reduced reparameterized unified model. Applied Psychological Measurement. https://doi.org/10.1177/0146621617707511

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

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

Dibello, L., & Stout, W. (2010). *Arpeggio version 3.1* [Computer Program]. Chicago: Applied Informative Assessment Research Enterprises.

Doucet, A., Godsill, S.J., & Robert, C.P. (2002). Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Stat. Comput*, *12*, 77-84. https://doi.org/10.1023/A:1013172322619

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*, 7-28. https://doi.org/10.1007/BF02294778

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*. *13*, 77-90. https://doi.org/10.1177/014662168901300108

Duong, T., Nguyen, T., & Bose, B. (2014). Convergence rate of MCMC and simulated annealing with application to client-server assignment problem. *Stochastic Analysis and Applications*.

Feller, W., (1950). An introduction to probability theory and applications. New York; NY: Wiley.

Feng, Y., Habing, B., & Huebner, A. (2014). Parameter estimation of the reduced RUM using the EM algorithm. *Applied Psychological Measurement*. *38*, 137-150. https://doi.org/10.1177/0146621613502704

Ghosh, M., Ghosh, A., Chen, M., and Agresti, A. (2000). Noninformative priors for one parameter item response models. *Journal of Statistical Planning and Inference*. *88*, 99-115. https://doi.org/10.1016/S0378-3758(99)00201-3

Haberman, S. J. (1977) Maximum likelihood estimates in exponential response models. *Annals of Statistics*, *5*, 815-841. https://projecteuclid.org/euclid.aos/1176343941

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA:Sage.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Hartz, S.M., & Roussos, L.A. (2008). The Fusion model for skills diagnosis: Blending theory

with practice. *ETS Research Report RR-08-71*. Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02157.x

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, *74*, 191-210. https://doi.org/10.1007/s11336-008-9089-5

Hulin, C., Lissak, R.I., and Drasgow, F. (1982) Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*, 249-260. https://doi.org/10.1177/014662168200600301

Jacquier, E., Johannes, M., Polson, N. (2007). MCMC maximum likelihood for latent state models. J. *Econom*. *137*, 615-640. https://doi.org/10.1016/j.jeconom.2005.11.017

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

Lim, Y. S., & Drasgow, F. (2017). Nonparametric calibration of item-by-attribute matrix. *Multivariate Behavioral Research*, *52*, 562-575. https://doi.org/10.1080/00273171.2017.1341829

Lim, Y. S., & Drasgow, F. (2016, July). A joint maximum likelihood approach to cognitive diagnosis, Presented at the meeting of Psychometrics Society Annual Meeting, Asheville, NC.

Lord,F.M.(1975) Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters,(Research Bulletin RB-75-33). Princeton, NJ:Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1975.tb01073.x

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

Mitra, D., Romeo, F., & Sangiovanni-Vincentelli, A. (1986). Convergence and finite-time behavior of simulated annealing. *Advances in Applied Probability*, *18*, 747-771. http://www.jstor.org/stable/1427186

Neyman, J. & Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1-32. http://www.jstor.org/stable/1914288

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178. http://www.jstor.org/stable/1165199

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

Swaminathan, H., & Gifford,J.A. (1983) Estimation of parameters in the three-parameter latent trait model. In D.Weiss (Ed.), *New Horizons in Testing*, 13-20. New York: Academic Press.

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

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

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

Van, Laarhoven, P. J., & Arts, E. H. L. (1989) Simulated annealing: Theory and application, Reidel: Amsterdam.

Zomaya, A.,& Kazman, R. (2010). Simulated Annealing Techniques. In Mikhail J. Atallah and

Marina Blanton (Eds.), Algorithms and Theory of Computation Handbook: General Concepts and Techniques, (pp. 33-1-33-18). Boca Raton: CRC Press.