Local Optima in Diagnostic Classification Models

By

Hongling Lao

Submitted to the graduate degree program in the Department of Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chair: Neal Kingston

Lesa Hoffman

Jonathan Templin

William Skorupski

Bruce Frey

Date Defended:

The dissertation committee for Hongling Lao certifies that this is the approved version of the following dissertation:

Local Optima in Diagnostic Classification Models

Chair: Neal Kingston

Lesa Hoffman

Jonathan Templin

William Skorupski

Bruce Frey

Date Approved:

Abstract

A Monte Carol simulation study was used to investigate the prevalence of local optima in Diagnostic classification models (DCMs) under multiple conditions. Five variables were manipulated, including *model constraints, starting values, item effect size, attribute correlation, and mastery base rate*. *Model constraints* had two categories (i.e., with and without). *Starting values* had five categories (i.e., true, random 1, random 2, extremely low, and extremely high). The other three independent variables were continuous, sampling from uniform distributions. Other related variables were fixed at simplified yet reasonable values to avoid interference. There were 1000 replications, each with three attributes, 18 items and 5000 examinees.

The simulation design had three levels. Data were at the highest, followed by model constraints, while starting values at the lowest level. For each replication, the same data set was used to estimate parameters 10 times. Half of them were estimated with model constraints. Each had one of those five sets of starting values. The other half were estimated without model constraints, using the same starting values as those in the *model with constraints* condition. The convergence rate was similar across conditions: about 98 percent.

Local optima were identified at each level. At the data level, 11.66 percent of converged replications were identified with local optima, which were exclusively located in the *model with constraints* condition. This indicated model without constraints had better estimation performance than the model with constraints. At the model constraints level, 74.75 percent of the converged replications were local optima in the *model with constraints* condition, whereas it was 0.31 percent for the *model without constraints* condition. This indicated local optima were much more prevalent while estimating with model constraints. At the starting values level, those in the *model with constraints* condition converged to different (and thus local) optima about 12 percent of the time when compared to all 10 estimations, and around 2 percent when compared to only five estimations, except for the *extremely low starting values* condition (78.97 and 76.34 percent respectively). The percentage was much lower in the *model without constraints* condition, ranging from 0 to 0.72.

In conclusion, model constraints had a higher probability for convergence to local optima in general. The worst choice was to use extremely low starting values along with model constraints. As the model constraints were set at the lower limits for main effect and interaction, searching from the nearby boundary made the estimation more unreliable and unpredictable. In other conditions, using different sets of staring values made little impact on the local optima occurrence. Surprisingly, the extremely high starting values without model constraints performed the best, with the fewest local optima, even fewer than the estimations using true parameters as their starting values. In addition, with better item quality and higher proportion of attribute mastery, there were fewer local optima while estimated with constraints. Those variables had little impact while estimating without constraints. Suggestions were provided to practitioners based on the findings.

Acknowledgement

Humans are complicated yet intriguing. An individual may live an ordinary replication of the social norm without an identifiable soul, isolated from their past and future. Sometimes, if lucky enough, that same person may eventually become a unique existence of their own. Silenced in the past, am I voicing now? Not to please. Not to fight. Just to express.

I am learning to craft my own soul and life, thanks to YOU.

Academically, I owe immense amount of gratitude to my advisor, Dr. Neal Kingston. He recognized my talents and invited me into this exciting methodology field. His strong and generous support has provided me with a comfortable environment to learn greedily, to explore bravely and to grow rapidly. He has created a golden age of transformation for me. His wisdom and vision guide me through troubled times.

Dr. Lesa Hoffman is a great story teller in the statistician world. She has the special talents to explain abstract and complicated ideas in simple languages even kids can understand. Without her kindness and devoted support, particularly at my final stage of dissertation, it was impossible to graduate soon. Her tenderness as a female is healing and has created very special positive memories. I am indebted to her deeply.

I have to thank Dr. Jonathan Templin for what he had taught me while co-chairing thesis with my advisor. The conceptual and technical gains from that time made it much easier to implement my dissertation. His expertise continues to guide and provide quality control for my dissertation. Dr. William Skorupski has provided great comments and suggestions to clarify and improve the simulation study. Dr. Bruce Frey has been very supportive for the sake of student's well-being. He observes the environment and finds the best way to help. I am very grateful to have them on my committee.

Beyond my committee, there are other people that have impressed me significantly. Dr.

Paul Johnson is a sincere and caring professor, who introduced me into statistics and programming. He inspires me to keep exceeding myself. It has been a great pleasure to work with Dr. Carol Woods. She treats people equally and respectfully, with a strong sense of integrity and honesty. I will not forget the care and support from Dr. Robert Harrington. I am thankful for the kindness Dr. Sheng has offered me when I did an internship at KSDE. An oversea team has contributed to my data collection at a timely manner, to whom I owe my sincere thanks. They are Congbin Liao, Jin Yao, Xijun Zhou, Huanxian Huang, Jun Tang, Jianhua Shi, Zexi Tang. Never have we met, but we are connected.

It was an intriguing process of "dissertating", experiencing lots of ups from new findings or creative problem solving, along with lots of downs of mistakes or bottleneck. Not only have I learnt so much technically, but also have I made insightful discovery on myself. The dissertation as a project has brought so many warm hearts to me, with their generous offers especially when most needed. I was touched and determined to pay their kindness back or forward.

Personally, I have built strong attachment to my beloved friends. Sylvia Tidwell Scheuring has been an amazing and dear friend to me. Her charisma, intelligence and leadership are still beyond my imagination. She is extremely generous with her knowledge and intelligence, upgrading my mental processors every time we talk. The family time together with Jerome, Killa, Rowan, and Ardian has been the sweetest treats during my stay at this lovely town. Dan He is another charming leader in a different way. She is extremely empathetic and kind, with a respectful sense of moral and principle. I began to learn the language of humanity from her. Her insightful feedback on my weakness opens door to my changes. There has been lots of turbulence in facing the brutal reality. Yet, that was the path to grow at root. My thanks to her will never be enough. Feng Jin is another person that has changed my life completely. A book is

a better place to tell the story than in a sentence. We witness each other's transformation and are part of it. The shared journey has taught me a hard lesson on what matters the most in life. Now I choose differently. We make it!

The love and care from my family has granted me with a good education and a dream for future. My dad teaches me equality and independence, particularly when the environment preaches the opposite. He has been encouraging and inspiring me to pursuit dreams and to be myself. He sets the most important tone in my life. My mom loves in her own way. She has been an extremely responsible and devoted mom. She always places the happiness of her children ahead of her own. A better balance is better. The companion, understanding and love from my siblings have meant so much to me. I am so glad to have you!

Live consciously, as Sylvia recommends. Thank you All!

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Imagine you are in a survival competition on an abandoned island. The next task is to search for the tallest tree in a big forest, which rewards the winner with a one-week supply of cooked food. The forest has multiple mountains. Audiences may find it rather easy to spot the tallest tree from an aerial perspective. An aerial view makes it easy both to move around the forest for a full picture and to identify the tallest tree from a global, comprehensive perspective. However, without aerial equipment, finding the tallest tree from a vast forest is a challenging task for competitors who are constrained to the ground—a local, narrow perspective. It is natural for some teams to identify the tallest tree on one mountain as the tallest tree in the whole forest, as they could not see there is another higher mountain out there, and they are running out of time and resources to explore more.

In statistics, this phenomenon of ending up with a locally-best solution is called a *local optimal solution* (i.e., a *local optimum*). In the survival competition, a local optimum is the tallest tree in one area, but not in the whole forest. Likewise, in maximum likelihood estimation, a local optimum is any peak in the sampling space other than the highest peak. The highest peak is called the global optimum, and the goal of maximization likelihood algorithm is to search for the global optimum. Yet, on some occasions, the search efforts may end after spotting a local optimum instead, because the local optimum satisfies the ending criteria in an algorithm. When this occurs, there could be more than one solution to the models, which is known as the plural of a local optimum: *local optima*. Among all plausible solutions, the best solution is the one with the highest likelihood—this is the *global optimum*. Thus, there could be multiple local optima, but there is only one global optimum.

Similar to an aerial perspective from above, if we have all possible solutions with their corresponding likelihoods lined up in front of us, this task would be straightforward. We could easily pick the highest one in the same way we could pick the tallest tree in a forest. The challenge lies in the fact that a solution is unknown unless the estimation algorithm finds it. We will never know the true solution. Instead, the best we can do is to draw a plausible conclusion based on the empirical evidence presented, with some level of confidence.

By definition, a local optimum is not necessarily the most accurate answer to the maximum likelihood question. It is sometimes unknown whether a local optimum is the global optimum. For estimation with closed form solutions, the proof lies in algebraic calculation. For estimation without closed form solutions, there is no absolute proof, and so there must be sufficient evidence to make an inference. For example, evidence for a global optimum can be provided when multiple estimation replications converge at the same solution (i.e., consistency), even based on different starting values. Otherwise, given that the model parameters will be different at a local optimum than those from the global optimum, the local optima therefore provide a less accurate answer, and avoiding local optima is a critical problem in estimating complex models.

This dissertation focuses on local optima within diagnostic classification models (DCMs; to be described in more detail in chapter 2). Although local optima are likely problematic in the estimation of DCMs, their frequency and impact is not well understood. Generally speaking, the current state of art is to assume the results provided from DCM software reflect the global optimum. However, this may not be the case in some situations. For example, a recent simulation study on diagnostic classification models (Lao, 2016) found that the prevalence of local optima ranged from about 6% to 94% under various conditions using Mplus (Muthén &

Muthén, 1998-2017). Item quality was a key factor that impacted both the model convergence and the occurrence of local optima. With low quality items, the estimation was more likely to end up with either non-convergence or with a local optimum.

However, one concern in Lao's study is that local optima were defined as any difference in the log likelihood in the estimation of a model without constraining parametrization, relative to a model with constraints. Although the two models were identical except the constraints, that is not a strict definition for local optima. This is because the sample space in the model with constraints is usually a subset of the whole sample space in the model without constraints. Thus, it is possible that the global optimum in the subset sample space is different than the global optimum in the whole sample space.

In the present study, local optima were defined as any difference across estimation replications based on the same estimation process multiple times, with identical model specification via the same software, but using different starting values. The same likelihood and parameters were expected upon convergence across these identical replications. When different likelihoods were obtained instead, a local optimum had occurred. Differently from Lao (2016), in the present study, the local optimum was defined more strictly. Here, the results from a model *with constraints* were compared only with the results from the same model *with constraints*, using the same software for estimation, to avoid any potential confounding from the model specification. Similarly, the results from a model *without constraints* were compared only with the results from the same model *without constraints*, using the same software for estimation. In this way, the models being compared shared the same sample space.

In fact, Mplus has already taken preventive action in some models: it tries to avoid local optima by using 20 different start values in each estimation by default. The optimal solution is selected from those 20 examined solutions. Another commercial software option for latent class analysis, Latent GOLD® (Vermunt & Magidson, 2016), takes a similar automated preventive strategy. In contrast to this practice, there was other software, such as the “CDM” and “GDINA” packages in R (R Core Team, 2017), that has neither included such preventive action, nor warned about the potential existence of local optima.

Furthermore, it is important to know the prevalence of local optima under various conditions in DCMs, which was the main purpose of this study. If the local optima issue turned out to be a rare event in every condition, we may not need to worry about it too much. Otherwise, we should take it more seriously and try to reduce its deleterious impact by making better informed decisions. The findings from this study can offer practical guidance in making better analytic decisions in estimation.

Chapter 2: Literature Review

Diagnostic Classification Models

DCMs are confirmatory latent class models and have received increasing attention in educational and psychological measurement. There are many good introductions to diagnostic classification models (DCMs; e.g., Bradshaw & Templin, 2014; Henson & Douglas, 2005; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Maris & Bechger, 2009; Rupp, Templin, & Henson, 2010; Sheehan, Tatsuoka, & Lewis, 1993; Templin & Bradshaw, 2014; von Davier, 2010, 2014).

DCMs are intended to analyze categorical data. The categorical data are assumed to reflect the underlying categorical unobservable variables (i.e., *attributes)*. For example, in educational measurement, DCMs can be used to model students' binary responses to infer students' mastery status. In counseling psychology, DCMs can be used to model patients' categorical responses to a depression scale, in order to infer whether they are depressed or not.

In a unidimensional test, a DCM measures only one attribute. The attribute may have a few categories. Each category is called a *latent class*. For example, in a test that measures the mastery status of addition, the attribute has two categories: master and non-master. There are two latent classes: master of addition and non-master of addition. For a multidimensional test that measures several attributes simultaneously, latent classes are unique combinations of categories from each attribute. For example, consider a test designed to measure two attributes: the mastery status of addition and the mastery status of subtraction. Both attributes have two categories: master and non-master. In total, there are four latent classes, including 1) those who master neither, 2) those who master addition but not subtraction, 3) those who master subtraction but not addition, and 4) those who master both.

DCMs can provide three types of information: 1) the item parameters (i.e., *the* *measurement model)*; 2) the proportion of examinees in each latent class (i.e., *the structural model*); and 3) the classification of examinees into the most likely latent class. The item parameters provide empirical evidence to evaluate item quality. The proportion of examinees in each latent class provides information on the population. The classification of examines into a latent class provides information on the individual level. Depending on the purpose of a test, researchers can focus their interpretation on different part of the results from DCMs.

The measurement model specifies the relationship between the item and the attributes. The correspondence between an item and the attribute(s) measured by the item is recorded in a q-matrix (de la Torre, 2008; Köhn, Chiu, & Brusco, 2015). A *q-matrix* specifies whether an attribute is measured by an item. It is a binary matrix, in which 0s indicate the item in that row does not measure the attribute in that column, and 1s mean the opposite. Q-matrix specification is an important step in implementing DCMs. If the information in the q-matrix is wrong, the misspecification would jeopardize the estimation (Rupp & Templin, 2008).

Log-Linear Cognitive Diagnosis Model

The following section will introduce a generic framework for the measurement model: the log-linear cognitive diagnosis model (LCDM; Rupp, Templin, & Henson, 2010, pp. 144-168). Some nested models are popular in the literature, such as the deterministic inputs, noisy and gate (DINA) model and the deterministic inputs, noisy or gate model (DINO) model (e.g., de la Torre, 2009, 2011; Junker & Sijtsma, 2001). For one thing, the nested models can be specified by fixing some parameters as zero in the saturated model, except when the saturated model is in fact a nested model itself (e.g., the DINO model). For another, starting with a saturated model and removing parameters based on empirical evidence is a more stringent practice, because whether a parameter is zero is a testable empirical question. Thus, only the saturated LCDM model will be introduced here.

The LCDM is a logistic regression in nature. For a single item, the outcome variable is the probability of observing a correct response, conditional on a latent class. It ranges between 0 and 1. The predictors are the attributes measured by the item, and the attribute interactions. The linear model ranges from negative infinity to positive infinity. A logit link function is used to bridge the different scales between the outcome variable and the linear model. A logit function is the logarithm of the odds. The odds are the ratio between the probability of success over the probability of failure. Mathematically, LCDM can be expressed as Equation 1.

$$\ln\left(\frac{\mathrm{P}(\mathrm{X}_{\mathrm{ic}} = 1|\boldsymbol{\alpha}_{\boldsymbol{c}})}{\mathrm{P}(\mathrm{X}_{\mathrm{ic}} = 0|\boldsymbol{\alpha}_{\boldsymbol{c}})}\right) = \lambda_{\mathrm{i},0} + \boldsymbol{\lambda}_{\mathbf{i}}^{\mathbf{T}}\mathbf{h}(\boldsymbol{\alpha}_{\boldsymbol{c}}, \mathbf{q}_{\mathbf{i}}) \tag{1}$$

$$= \lambda_{\mathrm{i},0} + \sum_{\mathrm{a}=1}^{\mathrm{A}} \lambda_{\mathrm{i},1,(\mathrm{a})}\alpha_{\mathrm{ca}}\mathrm{q}_{\mathrm{ia}} + \sum_{\mathrm{a}=1}\sum_{\mathrm{a}'>1}^{\mathrm{A}} \lambda_{\mathrm{i},2,\left(\mathrm{a},\mathrm{a}'\right)}\alpha_{\mathrm{ca}}\alpha_{\mathrm{ca}'}\mathrm{q}_{\mathrm{ca}}\mathrm{q}_{\mathrm{ca}'} + \ldots$$

The left-hand side of the equation is the logit of the probability of a correct response given a latent class. X_{ic} represents the item response for item *i* for a person in latent class *c*. $\boldsymbol{\alpha}_{\boldsymbol{c}}$ represents the latent class *c*, indicating by its corresponding attribute values. Ln is the natural logarithm function. $\mathrm{P}(\mathrm{X}_{\mathrm{ic}} = 1|\boldsymbol{\alpha}_{\boldsymbol{c}})$ represents the probability of a correct response given a latent class *c* in item *i and* $\mathrm{P}(\mathrm{X}_{\mathrm{ic}} = 0|\boldsymbol{\alpha}_{\boldsymbol{c}})$ is the conditional probability of an incorrect response.

The right-hand side of the equation is the linear model, composed of an intercept, attribute main effects, and all possible attribute interactions. The symbol λ represents the parameters. $\lambda_{\mathrm{i},0}$ represents the intercept parameter, as a baseline value. It is the logit of observing a correct response for an examinee who master none of the attributes measured by the item. The $\boldsymbol{\lambda}_{\mathbf{i}}^{\mathbf{T}}\mathbf{h}(\boldsymbol{\alpha}_{\boldsymbol{c}}, \mathbf{q}_{\mathbf{i}})$ is an abbreviation of the combination of main effects and interactions. Let’s take the compact abbreviation into smaller pieces for better understanding.

$\sum_{a=1}^{A}\lambda_{i,1,(a)}\alpha_{ca}q_{ia}$ represents all main effects. This notation includes all attributes in the test. Only the attributes measured by item *i* will have a non-zero main effect. This is manipulated by the q-matrix parameter (i.e., q_{ia}). $\sum$ is a notation for sum. $\lambda_{i,1,(a)}$ represents the main effect for attribute *a* in item *i*. α_{ca} is the value for attribute *a* in latent class *c*. q_{ia} is the q-matrix value for attribute *a* in item *i*, indicating whether the item measures attribute *a*. If the item measures the attribute (i.e., $q_{ia} = 1$), the main effect parameter for the item is estimated. If the item does not measure the attribute (i.e., $q_{ia} = 0$), the main effect parameter for the item is zero.

$\sum_{a=1}\sum_{a'>1}^{A}\lambda_{i,2,(a,a')}\alpha_{ca}\alpha_{ca'}q_{ca}q_{ca'}$ represents all two-way interactions. This notation includes all possible combination of two attributes. However, only when both attributes are measured by the item will they have a two-way interaction parameter, controlled by the q-matrix parameter (i.e., q_{ca} and $q_{ca'}$). Similarly, $\lambda_{i,2,(a,a')}$ represents the two-way interaction between attribute *a* and attribute a' for item *i*. α_{ca} and $\alpha_{ca'}$ represent the attribute values for attribute *a* and attribute a', respectively. q_{ca} and $q_{ca'}$ represent the q-matrix parameter for the two attributes, specifying whether an attribute is measured by the item. Finally, the ... at the end of Equation 1 represents the higher order interaction terms when item *i* measures more than two attributes.

It is noteworthy that the logit is the model scale, ranging from negative infinity to positive infinity. The model scale can be transformed back into the data scale as shown in Equation 2, using an inverse link function. The probability of a correct response to item *i* for an examinee in latent class *c* equals to the ratio of the exponent of the linear model to the sum of one and the exponent of the linear model.

$$P(X_{ic} = 1|\boldsymbol{\alpha_c}) = \frac{\exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \mathbf{q_i}))}{1 + \exp(\lambda_{i,0} + \boldsymbol{\lambda_i^T}\mathbf{h}(\boldsymbol{\alpha_c}, \mathbf{q_i}))} \tag{2}$$

An item measuring two attributes is shown as an example. The probability of a correct response given a latent class can be specified in Equation 3.

$$P(X_i = 1|\alpha_1, \alpha_2) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_1 + \lambda_{i,1,(2)}\alpha_2 + \lambda_{i,2,(1,2)}\alpha_1\alpha_2\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1,(1)}\alpha_1 + \lambda_{i,1,(2)}\alpha_2 + \lambda_{i,2,(1,2)}\alpha_1\alpha_2\right)} \tag{1}$$

In the linear model, $\lambda_{i,0}$ represents the intercept, $\lambda_{i,1,(1)}$ represents the main effect for attribute α_1, $\lambda_{i,1,(2)}$ represents the main effect for attribute α_2, and $\lambda_{i,2,(1,2)}$ represents the two-way interaction between attribute α_1 and attribute α_2. For the whole test, assuming local independence across items, the probability of observing an examinee's full responses to a test can be modeled as a sum of the products of observing the response to an item given its latent class weighted by the probability in each latent class.

The probability of each latent class is modeled via the structural model in DCMs, which will be briefly introduced below.

Structural Model

Whereas the measurement model explicates the relationship between a latent class and an item response, the structural model depicts the relationship among the latent classes. On the individual level, the structural model allocates each examinee into a most likely latent class. This individual membership is aggregated into the population proportion for each latent class. Thus, on the population level, the structural model estimates the distribution of examinees in each latent class.

There are various approaches to parameterize the structural model. This study adopts the log-linear structural model, consistent with the log-linear cognitive diagnostic measurement model. Statistically, the structural model and the measurement model are quite similar, except the structural model does not contain intercepts. It is under the analysis of variance framework.

Given that the range of a probability is between zero and one, the probability is inappropriate to be the outcome variable of a linear model that ranges from negative infinity to positive infinity. To solve the mismatch in sample space between the linear model and the outcome variable, a log link function is applied to the outcome variable to transform its range to be the same as the linear model.

The log of the probability of an examinee in latent class c, representing via symbol μ_c, is expressed in Equation 4.

$$\mu_c = \sum_{a=1}^{A} \Upsilon_{1,(a)}\alpha_{ca} + \sum_{a=1}^{A-1}\sum_{a'=a+1}^{A} \gamma_{2,(a,a')}\alpha_{ca}\alpha_{ca'} + \cdots + \gamma_{A,(a,a',\dots)}\prod_{a=1}^{A}\alpha_a \tag{4}$$

where, μ_c indicates the log of the probability in latent class c. *A* is the number of attributes. $\Upsilon_{1,(a)}$ is main effect for Attribute a. α_{ca} is a binary indicator representing whether that latent class indicates the mastery status of Attribute a ($\text{i.e.}, \ \alpha_{ca} = 1$), or its non-mastery status ($\text{i.e.}, \alpha_{ca} = 1$). $\gamma_{2,(a,a')}$ is the two-way interaction between Attribute a and Attribute a'. $\gamma_{A,(a,a',\dots)}$ is the *a*-way interaction among all attributes.

Local Optima in DCMs

Local optima are widespread phenomena in statistical estimation (Floudas, Pardalos, Adjinman, Esposito, Gumus, Harding, Klepeis, Mayer, & Schweiger, 2013; Horst, Pardalos, & Van Thoai, 2000). Pursuing global optimization is a common endeavor in various fields, such as in computer science (Chiang & Chu, 1996; Likas, Vlassis, & Verbeek, 2001; Qin, Huang, & Suganthan, 2009; Yao, Liu, & Lin, 1999), in business (Dorsey & Mayer, 1995; Goffle, Ferrier, & Rogers, 1994; Hu, Li, & Liao, 2010), in physics (Liang, Qin, Suganthan, & Baskar, 2006), and in biochemistry (Moles, Mendes, & Banga, 2003), etc. The statistical principle underlying local optima is universal across contexts. Local optima in DCMs have similar features as in other fields.

As described in chapter 1, local optima exist when a better solution exists than the one found in the a local neighborhood of the solution space. Depending on where the search begins in the sample space, the estimation process might either reach the global optimum or be trapped at a local optimum. To alleviate the local optima issue, optimization with multiple starts is a useful technique (e.g., Eddy, 1995; Knox, 1994; Ugray, Lasdon, Plummer, Glover, Kelly, & Marti, 2006). If results from many different starts all produce the same solution, it is persuasive support of this solution as the global optimum.

In addition, convergence is a related issue that needs to be clarified here. An estimation stops by meeting either of two different criteria, whichever comes first. One criterion is the estimation has reached the maximum number of iterations, without finding a plausible solution. This situation leads to a non-convergent estimation. No results are provided from non-convergent estimation. The other stopping criterion is, before reaching the maximum number of iterations, when the difference in the likelihood from the previous iteration is small enough. This situation leads to a convergent estimation, providing a plausible solution in the output. However, the solution may be a local optimum. Model convergence is a prerequisite to decide whether the solution is a global optimum or a local optimum. For more detailed discussion on convergence problems, please refer to other resources (e.g., Muthén & Muthén, 1998-2017, pp. 523-525).

Focus of the Current Study

Avoidance of local optima is necessary for high quality research of DCMs. This main purpose of this study was to provide guidelines with which to make better analytic choices after the data have been gathered (e.g., which model to use for estimation), rather than to provide advice on the research design at the early stage before data collection (e.g., how big of a sample size is needed). Accordingly, in the simulation that followed, the majority of variables that would be decided at the design stage were fixed. They were set at ideal, yet reasonable values to avoid interfering with the estimation. In contrast, key variables reflecting decisions at the analytic stage were manipulated and inspected for their potential influence on the occurrence of local optima. For the purpose of generalization, three variables at the design stage were included as additional manipulated variables. The findings from this study would shed light on the impact of these features in conducting future DCMs analyses.

The following section provides a brief conceptual description and rationale of the variables that were fixed first, followed by the variables that were manipulated instead.

Fixed Variables

The variables to be decided at the design stage included sample size, the number of attributes, and the number of items. To avoid any potential confounds for estimation accuracy, these variables were set at reasonable fixed values. A sample size of 5,000 respondents was chosen to provide sufficient data for accurate parameter estimation. Three attributes were measured at a relatively simple yet reasonable level of test complexity, given that problems in model recovery are exacerbated with more attributes. With three attributes, an interrelated relationship among attributes was a reasonable level of complexity applicable to traditional testing. Although a higher level of complexity in attribute structure is possible to be modeled via high performing computers in dynamic testing environment, such as those in personalized instruction and assessment, we did not address this case here.

The number of items was chosen using the following rationale. First of all, with three attributes measured, there are seven unique ways in how an item can measures the attribute(s). There are three ways in which an item could measure a single attribute (Attribute 1, Attribute 2, or Attribute 3). There are three ways in which an item can measure two attributes simultaneously (i.e., Attribute 1 & 2, Attribute 1 & 3, and Attribute 2 & 3). There is one way an item measures all three attributes (i.e., Attribute 1 & 2 & 3). However, from a pragmatic perspective, rarely is any item designed to measure more than two attributes at once, because it is both a content and a statistical challenge to disentangle effects from different attributes in each item. As a result, in this study, the items were designed to measure one or two attributes only. Each attribute was measured three times, once on its own and twice with another item. However, three items were far from ideal to extract reliable information for an attribute. A multiplication of three was used in order to maintain a balanced q-matrix (meaning all attributes are measured by the same number of items in an equivalent way). By repeating the six unique ways of measurement three times, in total, there were 18 total items in the simulated test, which was a realistic amount for a test diagnosing three skills.

Manipulated Variables

Five variables were manipulated in this study. The first two variables represented the typical decisions to be made during estimation, including 1) whether or not to use model constraints (i.e., *model constraints*), 2) what kinds of starting values to set for model estimation (i.e., *starting values*). The other three variables were included to improve the generalization of the study findings. It is rare each item measures its targeted attributes equally well throughout a test. Thus, item quality was included as the third variable (i.e., *item effect size*). This was for the generalization of items. Furthermore, the different strength of relationship among the attributes was the fourth variable (i.e., *attribute correlation*). This is for the generalization of attributes. Finally, the proportion of attribute mastery varied across populations, which served as the fifth variable (i.e., *mastery base rate*). This was for the generalization of examinees.

From a signal to noise perspective, anything that either strengthened the signal or reduced the noise should make it more likely to find the best solution (and avoid local optima). On the contrary, anything that either weakened the signal or increased the noise should make it less likely to find the best solution. The *model constraints* variable represented an attempt to reduce noise, by narrowing down the searching area based on prior information. The *starting values* variable represented the strategy for setting the starting point in searching, whether to start randomly or from one boundary towards center. The *item effect size* variable represented the strength of the signal. The *attribute correlation* variable represented the relationship between signals, somewhat related or unrelated. The *mastery base rate* variable represented the distribution of the signal strength, evenly or somewhat unevenly.

The following section provides a brief conceptual description for each variable, whereas the Method section provided a more detailed operational description for each.

The use of *model constraints* was the first variable to be manipulated. Model constraints should make a wrong solution less likely by narrowing down the sample space to be searched during maximum likelihood estimation. It was important to note, however, that this strategy only works when the prior information that guided setting the constraints was accurate. Otherwise, given misleading information, it might be impossible to come to the optimal solution at all, if that part of sample space that contains the accurate solution has been excluded from the beginning. In summary, when the constraints are accurate, they guide and speed up the estimation, but when the constraints are inaccurate, they could slow down the estimation, provide inaccurate solutions, or create problems with convergence. As a result, model constraints should be used cautiously, and preferably only when there is strong evidence to do so. In the current study, all main effects in DCMs were set to be positive, following the minimum requirement for an acceptable quality item.

The comparison of five different sets of *starting values* was the second variable to be manipulated. As a baseline for comparison, the true values were used as the starting values for estimation. This condition should result in the best parameter recovery. The other four sets represented two different approaches to select the starting values: one was randomly and the other is from an extreme point. Randomly choosing the starting values from a reasonable sample space is an intuitive and straightforward strategy. By searching from those likely places, it could be efficient sometimes but unorganized to support a thorough search when needed. On the other hand, searching from one extreme end to another is a more thorough yet time consuming strategy, which makes it less likely to miss the signal. In the sample space, it may search from either its extremely low end or its extremely high end. Two sets of random starting values were used to compare with those two sets adopting the extreme strategy.

The third manipulated variable was the *item effect size,* which referred to how well an item measures the attribute(s). It indicated how well an item serves its measurement purpose. The item effect size was largely a consequence of the design and sampling stage. However, item effect size was manipulated here to provide statistical evidence for the impact of including or excluding some items from analysis. For example, if an item measures an attribute rather poorly, which is a less harmful choice, to include it or to exclude it from analysis? To what extent of the item quality does the answer to this question change? In order to provide more details, item effect size was manipulated as a continuous variable, rather than a categorical variable.

The fourth variable was the *attribute correlation*, indicating the relationship between attributes. Attributes measured in the same test are usually related to some extent, yet different in other sense. For example, algebra and geometry in a math test are related to abstract thinking, but differ in the exact type of abstract thinking as well as their prerequisite knowledge. The correlation strengthens the overall signal to be found, at the price of mixing the boundary to separate from each other. Is there a sweet spot that balance the trade off?

The fifth variable was the *mastery base rate*, indexing the population distribution in the sample space. In other words, it depicts the proportion of examinees who had mastered that attribute. It should be easier to identify all signals if the signal strength is distributed evenly. However, in reality, it is unlikely to recruit examinees to form a completely balanced sample for different groups.

In summary, five variables were manipulated to examine the prevalence of local optima in DCMs under different conditions in this study. The main outcome variable was the occurrence of local optima.

Hypotheses

In summary, the hypotheses for this study were as follows:

1) Estimation with constraints would lead to fewer local optima;
2) With higher item effect size, local optima would be less prevalent;
3) There would be an interaction effect between *item effect size* and *model constraints*: with higher item effect size, the use of model constraints would have less impact on the prevalence of local optima.

Chapter 3: Method

A Monte Carlo simulation study was used to explore the prevalence of the local optima during estimation of diagnostic classification models under different conditions. The intent of this study was to provide guidance for better analytic decisions after the data are available. Accordingly, two variables in the modeling were selected to be the independent variables, including the *model constraints* and the *starting values*. In addition, in order to generalize the findings on diverse item quality, attribute relationship and population, three variables in the data generation stage were included as the independent variables. Other related variables that would have been decided at the design stage were fixed at relatively ideal yet reasonable values so that they would not interfere with the estimation quality. Before diving into the details, the overall design was introduced next.

Simulation Design

Five independent variables were manipulated in this study. The *model constraints* variable and the *starting values* variable were categorical, whereas the other three were continuous variables. Table 1 provided an overview of the simulation design.

As shown in Table 2, the design had a three-level structure. The highest level reflected the data generation process: 1,000 unique data sets were created using the same sampling process. In other words, the simulation study had 1,000 samples with which to observe sampling error. The 1,000 unique datasets were generated via R (version 3.3.1). The Q-matrix for this simulation were listed in Table 2. Mplus was used to estimate the models.

The lower two levels depict different layers with respect to the model estimation process. The two categorical variables were reflected in this structure. The second highest level is the *model constraints* variable, including two levels (*model with constraints* and *model without* *constraints)*. The third highest level is the *starting values* variable, including five categories (i.e., true, random 1, random 2, extremely low, and extremely high).

Local Optima Identification

In consistent to the hierarchical design of this study, operationally, local optima could be identified at three different levels. The highest level was at the data level. Local optima were identified between the two model constraints conditions. By comparing the best solution from the *model with constraint* condition to the best solution from the *model without constraint* condition. The best solution was indicated by their maximum likelihood. Local optima were identified if the maximum likelihood from two or more conditions differed. Results from this level could inform the choice of adding model constraints or not, based on which condition was more likely to reach a better solution. It was an estimation optimal performance consideration, allowing for multiple trials.

The second level of local optima identification was at the *model constraints* level. Local optima were identified within each model constraints condition. In other word, the five estimations from the *model with constraint* condition were compared to each other. The same was true for the *model without constraint* condition, in order to estimate the occurrence of local optima within each condition. Results from this level could further inform the choice of adding model constraints or not, based on which condition was more likely to produce local optima. It was an estimation reliability consideration, whether results from multiple replications were the same.

Local optima at the model constraints level were identified when there was more than one unique likelihood value across the five replications within the *model with constraint* condition or in the *model without constraint* condition, respectively. In that case, only the largest likelihood may be the global optimum, whereas the smaller one(s) were defined as local optima. Admittedly, although this definition of local optima was not precise, but when there was disagreement in the likelihood, it provided evidence for the presence of a local optimum. However, when there was no disagreement in the likelihood, those identical solutions may be either local optima or global optimum. In other word, this definition underestimated the occurrence of local optima. However, since it was impossible to prove a given solution was really a global optimum, this approximation to identify local optima was used in this study.

The third level was at the *starting value* level. Within each model constraints condition, the likelihoods from the five starting value conditions were compared to the maximum likelihood among the five. In reality, the true values were never known. Yet, they set a standard for best performance. Results from this level could inform the choice of starting values, based on which approach was least likely to produce local optima. It was an estimation optimal performance consideration.

Manipulated Variable Operationalization

With this big picture of the simulation design in mind, the below section described each of the five independent variables in more details. To be temporally consistent with reality, the manipulated variables related with the data generation stage were described first, including the *item effect size*, *mastery base rate*, and *attribute correlation*. Descriptions on the other two manipulated variables followed, including the *model constraints* and *starting values,* which were involved in the model estimation stage.

Item Effect Size

The first manipulated variable *item effect size,* which means how effectively an item measures its attribute(s). It was a quality indicator for an item. The parameterization in the measurement model used the LCDM framework as introduced in the previous chapter. The probability of the expected outcome (local optima here) was predicted by related variables via a logit link function to match the data scale with the model scale. Although the data scale (i.e., probability) was more intuitive and useful in interpreting results, the model scale (i.e., logit) was used to describe the parameters, because the parameters were specified and estimated in the model scale.

Table 3 summarizes the measurement model sampling specification. It begins with the logit of a correct response given a specific level of attribute mastery status. The logit was transformed back into probability to ensure its plausibility. The next step was to calculate the range for the corresponding parameters in the LCDM measurement model, based on the logits given the attribute mastery status. There were three types of parameters: intercept, main effect, and two-way interaction. The intercept parameter represented the logit of a correct response for persons who had not mastered any of the attributes measured by an item.

The main effect was the change in logit of a correct response for those who mastered the attribute, compared with those who did not master it. The lower limit of the main effect parameter equaled the lower limit of those who mastered the attribute minus the upper limit of those who did not master any attribute. The upper limit of the main effect parameter equaled the upper limit of those who master this attribute minus the lower limit of those who mastered none. This was true for both the main effect parameter in the one-attribute items and the two-attribute items. All main effects were set to be positive, to ensure monotonicity which meant with more measured attributes mastered, there was a higher probability of a correct response.

For the two-attribute items, the two attributes measured were conceptualized into a primary attribute and a secondary attribute. The primary attribute was the primary measurement target of the item, whereas the secondary attribute was the other auxiliary dimension reflected in the item. To operationalize the primary–secondary concept, the secondary attribute main effect was proportional to the primary attribute main effect. The proportion was randomly sampled from 20 to 80 percent for each item in order to eliminate sampling dependence between the primary and secondary main effects.

The two-way interaction parameter, unique to two-attribute items, was the change in logit of a correct response for complete masters above the additive two main effects. The lower limit of the two-way interaction parameter equaled the lower limit of those who mastered both attributes minus the sum of the upper limits of those complete non-masters, those who mastered the primary attribute, and those who mastered the secondary attribute. The upper limit of the interaction parameter equaled the upper limit of those who master both attributes minus the sum of the lower limits of those who mastered none, those who mastered the primary attribute, and those who mastered the secondary attribute. To ensure monotonicity, for each two-attribute item, the two-way interaction was constrained to be bigger than the negative of the minimum of the two main effects.

In this study, the *item effect size* was defined as the average main effect for the one-attribute items, and the average primary main effect for the two-attribute items, both in the logit scale. In order to examine the impact of the effect size of items, the whole test needed to share similar level of item quality—otherwise one could not distinguish the mixed effects of both high and low quality items. Accordingly, for each dataset, one value of *item effect size* was drawn from its parameter range. That sample value became the average main effect for the one attribute item, and the average primary main effect for the two-attribute items, with a range of ± 0.50 logits. For example, for dataset 1, a random sample drawn from the range of 1 to 4 ends resulted in value of 3. Then, all the main effects in the one-attribute items and all primary main effects in the two-attribute items would be drawn from a range between 2.5 and 3.5. The secondary main effect for a two-attribute item would be a random draw of the percentage between 20 and 80 of its primary main effect, resulting in a range between 0.5 and 2.8. For other parameters (i.e., intercepts and interaction), they were randomly sampled from their corresponding sample space independently. The same process was repeated for all datasets.

Attribute Correlation & Mastery Base Rate

The second manipulated variable was *attribute correlation* and the third manipulated variable was *mastery base rate.* The value of both variables was directly determined by the values of the structural model parameters. Table 4 contained the details of the sampling distribution for the structural model parameters. The main effects ranged between −2 and −1. The two-way interactions ranged between 1 and 2. Both the intercept and three-way interaction were fixed as zero. Similar to the sampling approach for the *item effect size* variable, for each replication, a mean of the main effect and a mean of the two-way interaction were sampled from their distribution, respectively. The main effect and interaction parameters were then sampled from a uniform distribution that centered at those means, with a range of ± 1.

With regard to the *attribute correlation* variable, since attributes were categorical, a tetrachoric correlation was used to index their relational strength. A tetrachoric correlation was calculated based on Equation 5 and Equation 6.

$$\rho = \frac{\gamma - 1}{\gamma + 1} \tag{5}$$

where

$$\gamma = \left(\frac{a_{00}a_{11}}{a_{10}a_{01}}\right)^{\frac{\pi}{4}} \tag{6}$$

where ρ is the tetrachoric correlation; γ was a place holder; a_{00} is the number of examinees who mastered neither attributes; a_{11} is the number of examinees who mastered both attributes; a_{10} is the number of examinees who mastered the first attribute but not the second attribute; a_{01} is the number of examinees who mastered second attribute but not the first attribute. Both a_{00} and a_{11} indicate agreement between two attributes, whereas both a_{10} and a_{01} represent disagreement between two attributes. These numbers were calculated based on the latent class probabilities.

On the other hand, the *mastery base rate* of an attribute was defined as the proportion of examinees who had mastered the attribute. The *mastery base rate* equaled to the sum of those related latent class probabilities, which indicated a mastery status of the targeted attribute. The latent class probability was calculated via the log linear structural model, given the structural parameters.

Based on the range of the structural model parameters, the range of the *attribute correlation* variable was from 0.34 to 0.79. The range of the *mastery base rate* variable was from 0.18 to 0.89. Since there were three attributes, both the *attribute correlation* variable and the *mastery base rate* variable had three values, respectively. The values were similar as a result of the sampling approach for the main effects and two-way interaction, centering at a mean with a small range for each replication. The average of these three values was used to represent the variable for a replication.

After describing the manipulated variables related to the data generation stage, below was a description for the other two independent variables related to the model estimation stage.

Model Constraints

The forth manipulated variable was *model constraints*. There were two categories (i.e., with and without model constraints). Mplus was used for this study, as it allowed for adding model constraints. The model constraints were placed to ensure monotonicity, which meant an increased probability for a correct response given the mastery of each additional attribute that was measured by an item. Rupp, Templin, and Henson had a detailed discussion on the constraints in their book (Rupp, Templin, & Henson, 2010; pp. 208-211). All main effects were constrained to be positive. The interaction terms were constrained such that there was an increase in the probability to a correct response along with more measured attributes mastered. More specifically, the two-way interaction for each item was constrained to be bigger than the negative of the minimum of their two main effects. It shared the same constraint rule with data sampling.

Starting Values

For the fifth manipulated variable, *starting values,* there were five categories (i.e., true, random 1, random 2, extremely low, extremely high). The detailed specification is shown in Table 5. The starting values in the true condition were the true parameters that generated its corresponding data set. The starting values in the random condition were randomly sampled from a uniform distribution with its range be the same as the parameter sampling range. The starting values in the extremely low condition were fixed, to ensure the starting values were at their permissible sampling space given the constraints placed both on the data and on the *model with constraint*. It would be difficult to predict how an estimation was implemented when the starting values were out of range. The intercepts were fixed at −4.5 at the measurement model. The main effects were fixed at 0.1 and the two-way interactions were fixed at 0 for both the structural model and the measurement model. On the contrary, since the constraints were on the lower limit of main effects and interactions, there was no concern for starting values around their upper limit. The starting values in the extremely high were randomly sampled from a uniform distribution, with the upper limit the same as the upper limit of the sampling distribution. The lower limit was 0.5 less than its upper limit, in order to produce starting values from the high end.

Chapter 4: Results

The 1000 replications were analyzed to answer the research questions. The simulation was run on three machines: 851 replications were estimated using Mplus version 8 and the other 149 were estimated via Mplus version 7. Results from these two different software versions were compared and turned out to be highly similar. Thus, all results were analyzed together. Of the 1000 replications, for 836 all 10 estimations converged.

Results for parameter recovery, prevalence of local optima, and conditions associated with local optima were presented here.

Parameter Recovery

Estimation quality was indexed by the average bias and root-mean-square error (RMSE). Bias was the mean difference between true value and estimated value. RMSE is the square root of the mean of the squared bias. To combine information from dozens of parameters involved in a model, the index from the same type of parameters were aggregated and represented by their average, including intercepts, main effects, and interactions. Table 6 provides details of parameter recovery in the structural model, whereas Table 7 summarizes those in the measurement model.

Based on Table 6, the majority of estimation conditions performed well in recovering parameters in the structural model, except for one condition. Except for the *extremely low starting values* in the *model with constraints* condition, the mean of the bias and of the RMSE for both main effects and interactions were zero or close to zero, with a small standard error, ranging from 0.04 to 0.07. The person classification to latent class recovery was 0.82. On the other hand, for the *extremely low starting values* in the *model with constraints* condition, the bias and RMSE and their standard errors were larger. The person classification recovery statistic was lower as well.

A similar recovery pattern happened in the measurement model, as shown in Table 7, with relatively worse recovery for the interaction parameters in general. Estimation from the *extremely low starting values* in the *model with constraints* condition was relatively worse than the rest. That signaled a caution in interpretation results from this specific estimation condition.

Since replications with estimates based on local optima have worse parameter recovery by definition, parameter recovery was further summarized based on whether that estimation condition had local optima. Local optima were determined at the starting values level, by comparing its likelihood to the maximum likelihood of all 10 estimations for the same data set. Table 8 and Table 9 provided parameter recovery summary for the estimations showing local optima. Table 10 and Table 11 provided parameter recovery summary for the estimations showing no local optima.

In general, results showed that the parameter recovery was much better in the estimations showing no local optima than in the estimations showing local optima, with smaller bias, RMSE and standard errors of both bias and RMSE. The biggest difference was found for the *extremely low starting values* in the *model with constraints* condition, since parameters recovered the worst in this specific condition while having local optima. The parameters in the *extremely low starting values* in the *model with constraints* condition recovered as well as the other estimation conditions while the solutions were not local optima. This finding suggests the problem with using very low starting values is the increased probability of finding a local optimum.

After confirming the quality of the estimation, the next step was to summarize the local optima at different levels, to answer the prevalence question.

Local Optima Prevalence

Local optima were identified at three levels, consistent with the simulation design structure. The highest level was the data level. The medium level was the model constraints level. The lowest was the starting values level. The basic idea was to compare likelihoods across estimations for the same data set. Local optima were identified for the estimation with a smaller likelihood. Table 12 to 15 summarized the prevalence information for these three levels, including a detailed definition of local optima at each level in the table note.

Within the Data Level

At this level, for each data set, the maximum likelihood from the five estimations in the *model with constraints* ("C") condition was compared to the maximum likelihood from those five estimations in the *model without constraints* ("NC") condition. Local optima were identified if the two maximum likelihoods were different. The particular condition that had local optima was further identified, signaled by a smaller maximum likelihood. In order to compare, at least one estimation from both model constraints conditions had to converge, which became the convergence criterion at this level.

As shown in Table 12, 115 out of the 986 (11.66 percent) data sets that always yielded converged solutions were identified to have local optima. All 115 local optima were identified at the *model with constraint* condition.

Within the Model Constraints Level

At this level, for each data set, within each model constraints condition, the likelihoods from the five estimations were compared to each other. Local optima were identified if there were more than two unique likelihoods within each model constraints condition. In order to to be considered, at least two estimations from *each* model constraints condition had to converge. It was a stricter convergence criterion than the data level.

Based on Table 13, both model constraints conditions had the same number of convergent replications. However, the *model with constraints* had a much higher percentage of local optima than the model without constraints, 74.75 versus 0.31.

Within the Starting Values Level

At this level, for each starting values condition, the convergence was counted for each estimation condition, separately for the two model constraints conditions. The local optima were identified at both the data level and the model constraints level. At the data level, if any of the 10 estimations converged, its likelihood was compared to the maximum likelihood among the 10 estimations for the same data set. Local optima were identified for that estimation condition if its likelihood was smaller than maximum likelihood. Similarly, at the data level, if any of the 5 estimations converged, its likelihood was compared to the maximum likelihood among the 5 estimations for the same model constraints condition. Local optima were identified for that estimation condition if its likelihood was smaller than the corresponding maximum likelihood. Results are summarized in Table 14.

Compared to the model constraints level, the rule at data level was stricter to identify local optima, because the maximum likelihood from 10 estimations could be higher than from five of the 10. As a result, more local optima were identified at the data level than the model constraints level. Furthermore, based on the local optima prevalence results at the data level, the maximum likelihood of 10 estimations was from the *model without constraints* condition. As a result, there was a big difference in the number of local optima identified at the data level and at the model constraints level only for the staring values in the *model with constraints* condition, along with a higher prevalence. Particularly, the *extremely low starting values* in the *model with constraints* condition had the highest local optima prevalence: 78.97% at the data level and 76.34% at the model constraints level. As expected, within the *model with constraints* condition, the *true staring values* had the fewest local optima, whereas the *extremely low starting values* had the most. There was little difference across different staring values in the *model without constraints* condition, except the *extremely high staring value* condition had the lowest local optima, even lower than the true starting values condition.

Results from all three levels provided evidence to reject the first hypothesis. Surprisingly, the estimation with model constraints had more prevalent local optima, particularly with extremely low starting values.

Predicting Local Optima

Although it was possible to model the local optima at all three levels, testing the hypotheses required only the model constraints level. Thus, the scope of analysis at this section was narrowed down to the results at that level. Predicting local optima was implemented using an exploratory approach. The first step was to gain insight of the relationship between outcome variables and its predictors. Both correlation and visualization were used to serve this purpose.

Point-Biserial Correlation

Table 13 shows the point-biserial correlation between local optima and their predictors. The negative correlation between the *item effect size* with local optima in the *model with constraints* condition ($r = -0.16$) supported the second hypothesis that with higher item effect size, local optima were less prevalent. Furthermore, the positive correlation ($r = 0.06$) in the *model without constraints* condition was in favor of the third hypothesis that there was an interaction effect between *item effect size* and *model constrains*. For *model without constraints*, item quality had less impact on the estimation, which was true for the other two continuous predictors.

In addition, there was a negative correlation between mastery base rate and attribute correlation for the *model with constraints* condition. When more examinees mastered an attribute, there were fewer local optima ($r = -0.22$). Given a stronger attribute correlation, there were slightly fewer local optima ($r = -0.09$).

Visualization

Various approaches can be used to visualize relationships between categorical and continuous variables. In this study the continuous variables were categorized into ten equal size intervals and drew a line to connect the points to predict a trend. Admittedly, there are no data on the line except for the several points and thus this may be misleading. However, when interpreted along with other evidence, it should provide useful information. Table 11 shows the exact numbers and percentages, and Figure 1 shows the plot.

For each continuous variable (i.e., *mastery base rate, attribute correlation*, and *item effect size*), 10 equal-interval bins were created based on its observed minimum and maximum. By assigning replications into corresponding bins, the continuous variables were transformed into categorical ones. The percentage of local optima at each bin was calculated and compared in order to depict a trend.

In some ways results were consistent with the point-biserial correlations. Since there were only three replications in the *model without constraints* condition that had local optima, it is less meaningful to look at that condition. The percentage of local optima in the *model with constraints* condition was high, ranging from 25.81 to 88.62. There was a general trend of negative correlation, such that with higher values in the continuous variables, there were fewer local optima, but some curvilinearity is evident in the relationships

Logistic Regression

Based on the information from the visualization, both linear and quadratic terms of the three continuous predictors (i.e., *mastery base rate, attribute correlation,* and *item effect size*) were used to build a logistic regression model to predict the occurrence of local optima. In addition, all possible interaction terms were included, primarily for two reasons. Firstly, the *mastery base rate* and *attribute correlation* variables were determined by the structural model parameters together. They were not independent from each other. Secondly, there may be potential interaction effect between the item quality and the attribute features. As such, findings could provide insight on the relationship between items and attributes for test design.

Moreover, since there were only three replications with local optima in the *model without constraints* condition, it was too weak a signal to be modeled accurately. The small sample provided insufficient power to test the third hypothesis with regard to the interaction between model constraints and item effect size. In other word, only replications from the *model with constraints* condition were included in the model. In total, there were three predictors, *mastery base rate, attribute correlation,* and *item effect size.*

In addition, in order to making the intercept parameter meaningful, all three continuous predictors were centered at their mean values. Specifically, the mean of the *mastery base rate* was .51. The mean of the *attribute correlation* was .65. The mean of the *item effect size* was 2.49 in the logit unit. Furthermore, to aid interpretation of the *mastery base rate*, it was rescaled to make one unit corresponding to one percent, instead of the 100 percent. To accomplish this purpose, its centered values were multiplied by 100. The same idea applied to the *attribute* *correlation*, by redefining its one unit corresponding to 0.1 correlation, instead of 1. Its centered values were multiplied by 10. These data preparation steps were conducted before the modeling.

The modeling took a reduction approach, beginning with a full model including all reasonable parameters at the first round of modeling. At the next step, the statistically non-significant parameters were removed from the current model. After removing the extra parameters, the new model was estimated independently. Again, based on the same statistical significance criterion, non-significant parameters were further subtracted from the model. The same process continued iteratively until all parameters remained were statistically significant. Given the big sample size of 1000, a relatively conservative criterion was used to select parameters. Only parameters with p value equal to or lower than 0.01 were retained.

The parameters included at the first round were intercept, three main effects, three quadratic effects, three two-way interactions and one three-way interactions of the *mastery base rate*, *attribute correlation* and *item effect size*. Among the 11 coefficients, six of them were statistically significant at 0.01 level. They were the intercept, the main effect, and the quadratic effect of both the *mastery base rate* and *item effect size*, along with the interaction effect of these two predictors. They were retained for the second-round modeling. All parameters were statistically significant at 0.01 level in this final model.

The estimated intercept was -4.49, with a standard error of 0.77. This means that the logit of observing local optima was -4.49 for a replication with an average of 51% students mastering the attributes being measured, an average of .65 tetrachoric correlation between attributes, and an average of 2.49 logit of item main effects in the LCDM. Using the inverse logit link function (i.e., $\left[\frac{e^x}{(1+e^x)}\right]$), the probability was 0.01.

With regard to the *mastery base rate* predictor, the estimated coefficient for its main effect was 0.07, with a standard error of 0.02. The estimated coefficient for its quadratic effect was −0.001, with a standard error of 1.32. This means, with one percent higher *mastery base rate*, the change in the probability of observing local optima was 0.07 logit minus the product of 0.001 and the squared change in *mastery base rate*. For the *mastery base rate* smaller than 70 percent, with an increase in its value, there was an increase in the probability of observing local optima. On the contrary, for the *mastery base rate* higher than 70 percent, with an increase in its value, there was a decrease in the probability of observing local optima. However, since there was an interaction effect between the *mastery base rate* and the *item effect size,* this interpretation adjusted with the value of the other variable.

With regard to for the *item effect size* predictor, the estimated coefficient for its main effect was 3.92, with a standard error of 0.14. The estimated coefficient for its quadratic effect was −0.86, with a standard error of 0.17. This means, with one logit change in the *item effect size*, the change in the probability of observing local optima was 3.92 logit minus the product of 0.86 and the squared change in *item effect size*. For the *item effect size* smaller than 4.56 logit, with an increase in its value, there was an increase in the probability of observing local optima. On the contrary, for the *item effect size* higher than 4.56 logit, with an increase in its value, there was a decrease in the probability of observing local optima. However, since there was an interaction effect between the *mastery base rate* and the *item effect size,* this interpretation adjusted with the value of the other variable.

With regard to for the two-way interaction between the mastery base rate and the item effect size, its estimated coefficient was −0.04, with a standard error of 0.007. This means, with one unit increase in either mastery base rate or item effect size, there was 0.04 logit additional decreases in the probability of observing local optima predicted by the other variable.

Chapter 5: Discussion

Summary

This study is intended to guide analytic choice by comparing the prevalence of local optima under various conditions. Surprisingly, the convergence rate was barely affected by adding model constraints, even though it took longer to estimate. Perhaps the most important finding is that local optima were much more frequent when model constraints were imposed, particularly when extremely low starting values were used with model constraints. As the model constraints were set at the lower limits for main effect and interaction, searching from the nearby boundary made the estimation more unreliable and unpredictable.

The negative impact of local optima was further supported by the difference in the parameter recovery between the group with local optima and the group without local optima. The estimations having no local optima recovered the parameters much better than the estimations having local optima. However, when the estimations with extremely low starting values with constraints had no local optima, the estimations showed the same parameter recovery quality as the other estimation conditions.

Interestingly, the average bias in recovering the structural main effects was negative for the extremely low starting values with constraints estimation condition. There were at least two hypotheses that might explain this observation. One hypothesis was that the negative sign might be due to label switching in the parameters. However, no label switching was identified in the replications that had no local optima. No identified label switching was a surprising result, differing from the findings in Lao (2016). In that study, label switching was identified by comparing the estimated structural parameters between the model with constraints and the model without constraints, on the condition that both models converged and had the same likelihood.

The model with constraints and the model without constraints had different sampling space. Differently, in this study, the label switching was evaluated by comparing the estimated structural model parameter from the exactly same model, with constraints to with constraints, and without constraints to without constraints. In addition, this study did not have any three-way interactions which made estimation easier.

The other hypothesis is that local optima are more prevalent at the lower limit of the sample space. Future research on this possibility remains necessary.

Other conditions, using different sets of starting values had little impact on the occurrence of local optima. Surprisingly, the extremely high starting values without model constraints performed the best, with the fewest local optima, even slightly fewer than the estimations using true parameters as their starting values. However, there was no such difference after separating the group with local optima from the group without local optima.

For models estimated with constraints, both item quality and the proportion of attribute mastery affected the prevalence of local optima. The effect was quadratic for both predictors, as well as being interactive between them. Generally speaking, for smaller values, there was an increase in the probability of observing local optima along with the increase in the predictors. However, for higher values, there was a decrease effect. With regard to the interactive term, with one predictor bigger, there was a decrease in the probability of observing local optima predicted by the other variable. This result indicated a connection between the structural model and the measurement model.

It was an arguable choice to use the structural model parameters as predictors because the structural model parameters were manipulated and consequentially determined the values of the *mastery base rate* and *attribute correlation* variable. However, the research used the *mastery* *base rate* and *attribute correlation* as predictors primarily for two reasons. One reason was these two predictors had corresponding substantive meanings, potentially providing useful practical guidance. Without having any substantive meanings, those structural parameters provided limited implications for practitioners. The other reason was that there are other approaches to modeling the structural model, such as the Bayesian approach. The log linear model is only one of them, which makes the results modeling dependent and limited to generalize.

Recommendation for Practitioners

Based on the findings in this study, the most important suggestion is to avoid using extremely low starting values while estimating with model constraints. This approach has a high probability of resulting in local optima. On the contrary, it makes little difference to choose the randomly selected starting values and the extremely high starting values while adding model constraints. However, while estimating without model constraints, estimation with extremely low starting values is not problematic at all. It performs the same as using random starting values.

The second suggestion is to be cautious to add model constraints in general, since 11% of the solutions turned out to be local optima. This suggestion may raise concern on the nature of DCMs. In nature, DCMs are confirmatory latent class models, which are specified by the model constraints. Without adding any model constraints, it may be difficult to stabilize the model in a predictable way. For example, the label switching discussed in Lao's study (2016) is a big concern in the models without any constraints, using Mplus version 7. However, no label switching was identified in this study though, there were some different simulation specifications.

The third suggestion is to be aware of the impact of both item quality and attribute mastery base rate on the occurrence of local optima. When in doubt, replicating estimations multiple times to check for local optima remains a useful strategy. In addition, it may be useful to try different starting values or removing model constraints to look for alternative modeling approaches.

Limitation and Future Direction

The most important limitation of this study is its generalizability. This issue prevailed in multiple layers. Software was the most critical one, because local optima directly results from the searching algorithm in the sample space. Different software programs may adopt completely different search strategies, which could provide different results. This study used Mplus for estimation. Studies with other software might yield different findings. It is valuable to compare findings in other software to exclude conclusions due to software specific factors.

Another layer was the test design. This study fixed some variables in a test design in a simplified manner. In addition, the item quality, attribute mastery base rate and attribute relationship were highly similar within a test. Such settings make it difficult to generalize findings in more complex situations, such as a longer test with diverse item quality and various mastery base rate, using polytomous scoring rules. It is interesting to see how local optima might behave in a more complicated situation.

References

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*(3), 403-425.

Chiang, H. D., & Chu, C. C. (1996). A systematic search method for obtaining multiple local optimal solutions of nonlinear programming problems. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, *43*(2), 99-109.

de La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179-199.

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

de La Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of educational measurement*, *45*(4), 343-362.

Dorsey, R. E., & Mayer, W. J. (1995). Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features. *Journal of Business & Economic Statistics*, *13*(1), 53-66.

Eddy, S. R. (1995, July). Multiple alignment using hidden Markov models. In *Ismb* (Vol. 3, pp. 114-120).

Floudas, C. A., Pardalos, P. M., Adjiman, C., Esposito, W. R., Gümüs, Z. H., Harding, S. T., ... & Schweiger, C. A. (2013). *Handbook of test problems in local and global optimization* (Vol. 33). Springer Science & Business Media.

Goffe, W. L., Ferrier, G. D., & Rogers, J. (1994). Global optimization of statistical functions with simulated annealing. *Journal of econometrics*, *60*(1-2), 65-99.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*(4), 262-277.

Horst, R., Pardalos, P. M., & Van Thoai, N. (2000). *Introduction to global optimization*. Springer Science & Business Media.

Hu, J. Q., Li, J. Z., & Liao, G. (2010). A multi-QoS based local optimal model of service selection. *Chinese Journal of Computers*, *33*(3), 526-534.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Knox, J. (1994). Tabu search performance on the symmetric traveling salesman problem. *Computers & Operations Research*, *21*(8), 867-876.

Köhn, H. F., Chiu, C. Y., & Brusco, M. J. (2015). Heuristic cognitive diagnosis when the Q-matrix is unknown. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 268-291.

Lao, H. (2016). *Estimation of diagnostic classification models without constraints: Issues with class label switching* (Master thesis). University of Kansas, Lawrence, Kansas, USA.

Liang, J. J., Qin, A. K., Suganthan, P. N., & Baskar, S. (2006). Comprehensive learning particle swarm optimizer for global optimization of multimodal functions. *IEEE transactions on evolutionary computation*, *10*(3), 281-295.

Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

Maris, G., & Bechger, T. (2009). Equivalent diagnostic classification models.

Moles, C. G., Mendes, P., & Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome research*, *13*(11), 2467-2474.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Qin, A. K., Huang, V. L., & Suganthan, P. N. (2009). Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE transactions on Evolutionary Computation*, *13*(2), 398-417.

R Core Team (2017). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78-96.

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

Sheehan, K., Tatsuoka, K., & Lewis, C. (1993). A diagnostic classification model for document processing skills. *ETS Research Report Series*, *1993*(2).

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, *79*(2), 317-339.

Yao, X., Liu, Y., & Lin, G. (1999). Evolutionary programming made faster. *IEEE Transactions on Evolutionary computation*, *3*(2), 82-102.

Ugray, Z., Lasdon, L., Plummer, J., Glover, F., Kelly, J., & Martí, R. (2007). Scatter search and local NLP solvers: A multistart framework for global optimization. *INFORMS Journal on Computing*, *19*(3), 328-340.

Vermunt, J. K., & Magidson, J. (2017). *Latent GOLD®*. Statistical Innovations Inc., Belmont, MA.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychology Science Quarterly*, *52*(1), 8.

von Davier, M. (2014). The Log-Linear Cognitive Diagnostic Model (LCDM) as a Special Case of the General Diagnostic Model (GDM). *ETS Research Report Series*, *2014*(2), 1-13.

Table 1 Simulation Design Overview

Data Generation	Model Constraints	Starting Values
Data 1	C	True
		Random 1
		Random 2
		Extremely low
		Extremely high
	NC	True
		Random 1
		Random 2
		Extremely low
		Extremely high
Data 2	…	…
…	…	…
Data 1000	…	…

Note: The “C” means the *model with constraints,* and “NC” means the *model without constraints.* There are three continuous independent variables included in the simulation not shown here, including the *item effect size*, the *mastery base rate,* and the *attribute correlation*.

Table 2 Q-Matrix Specification

Item	A1	A2	A3
1	0	0	1
2	0	0	1
3	0	0	1
4	0	1	0
5	0	1	0
6	0	1	0
7	1	0	0
8	1	0	0
9	1	0	0
10	0	1	1
11	0	1	1
12	0	1	1
13	1	0	1
14	1	0	1
15	1	0	1
16	1	1	0
17	1	1	0
18	1	1	0

Note: A1, A2, and A3 represent Attribute 1, Attribute 2, and Attribute 3, respectively.

Table 3 LCDM Measurement Model Sampling Specification

Attribute Mastery Status	Logit L	Logit U	Probability L	Probability U	Parameters	Symbol	Logit L	Logit U
One-attribute items								
Complete non-masters	−2	−1	0.12	0.27	Intercept	$\lambda_{i,0}$	−2	−1
Complete masters	0	2	0.50	0.88	Main Effect	$\lambda_{i,1,(a)}$	1	4
Two-attribute items								
Complete non-masters	−2	−1	0.12	0.27	Intercept	$\lambda_{i,0}$	−2	−1
Primary attribute masters	0	2	0.50	0.88	Primary Main Effect	$\lambda_{i,1,(a)}$	1	4
Secondary attribute masters	−0.8	1.2	0.31	0.77	Secondary Main Effect	$\lambda_{i,1,(a')}$	0.2	3.2
Complete masters	2	2.5	0.88	0.92	Two-way Interaction	$\lambda_{i,2,(a,a')}$	−(minimum of two main effects)	−(minimum of two main effects) + 4

Note: "L" and "U" stood for the lower limit and upper limit of a range, respectively, of the logit/probability for a correct response given the attribute mastery status. The primary attribute was the primary measurement target of the item, whereas the secondary attribute was the other auxiliary dimension reflected in the item. The secondary main effect is 20 to 80 percent of the primary main effect.

To ensure monotonicity, for the two-way interaction parameters, the lower limit of sampling distribution was constrained as the negative of the smaller value of the two sampled main effects in that item, and the higher limit was four plus the adjusted lower limit.

Table 4 Structural Model Sampling Specification

Parameters	Symbols	Distribution
Main effect	$(\gamma_{i,1,a})$	$U(-2,-1)$
Two-way interaction	$(\gamma_{i,2,(a,a')})$	$U(1,2)$
Three-way interaction	$(\gamma_{i,3,(1,2,3)})$	0

Note: U (a, b) referred to a uniform distribution, ranging from *a* to *b*. The proportional of attribute mastery ranged between .18 to .89. The tetrachoric correlation between attributes ranged between .44 and .79.

Table 5 Starting Values Specification

Parameters	Symbol	E-Low	True/Random		E-High	
		Fixed	Lower	Upper	Lower	Upper
Structural Model						
Main effect	$\gamma_{i,1,a}$	0.1	−2	−1	−0.5	−1
Two-way interaction	$\gamma_{i,2,(a,a')}$	0.0	1	2	1.5	2
Three-way interaction	$\gamma_{i,3,(1,2,3)}$	0.0	0	0	0.0	0
Measurement Model						
(One-attribute items)						
Intercept	$\lambda_{i,0}$	−4.5	−2	−1	−0.5	−1
Main Effect	$\lambda_{i,1,(a)}$	0.1	1	4	3.5	4
(Two-attribute items)						
Intercept	$\lambda_{i,0}$	−4.5	−2	−1	−0.5	−1
Primary Main Effect	$\lambda_{i,1,(a)}$	0.1	1	4	3.5	4
Secondary Main Effect	$\lambda_{i,1,(a')}$	0.1	1	4	3.5	4
Two-way Interaction	$\lambda_{i,2,(a,a')}$	0.0	−− (minimum of two main effects)	−− (minimum of two main effects) + 4	−− (minimum of two main effects) + 4 − −0.5	−− (minimum of two main effects) + 4

Note: Logit was the unit. The *True/Random* conditions were sampled from a uniform distribution, with the same range as the data generation. In *Extremely Low* condition (“E-Low”), the starting values were fixed at permissible low values. In *Extremely High* condition (“E-High”), the starting values were sampled from a uniform distribution, with the upper limit the same as the sampling upper limit, and the lower limit as 0.5 less than the upper limit.

Table 6 Parameter Recovery for the Structural Model

Model Constraints	**Starting Values**	**Convergence**	**Main Effects**				**Interactions**				**Kappa**	
			Bias		RMSE		Bias		RMSE			
			mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*
C	**True**	978	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	R1	938	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	R2	921	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	E-Low	951	**−0.24**	**0.20**	**0.19**	**0.30**	**0.21**	**0.17**	**0.40**	**0.60**	**0.75**	**0.09**
	E-High	972	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
NC	**True**	980	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	R1	979	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	R2	979	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	E-Low	977	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	E-High	983	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08

Note: "C" stood for *model with constraints* condition. "NC" stood for *model without constraints* condition. "R1" stood for the first *Random* starting value condition. "R2" stood for the second *Random* starting value condition. "E-low" stood for the *Extremely Low* starting value condition. "E-High" stood for the *Extremely High* starting value condition. "RMSE" stood for root mean squared of error. "s.e." stood for standard error.

Table 7 Parameter Recovery for the Measurement Model

Model Constraints	Starting Values	Intercepts				Main Effects				Interactions			
		Bias		RMSE		Bias		RMSE		Bias		RMSE	
		mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*
C	**True**	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.18	0.28	1.67
	R1	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.14	0.19	1.09
	R2	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.12	0.15	0.74
	E-Low	**0.12**	**0.09**	**0.08**	**0.06**	**−0.61**	**0.39**	**1.10**	**0.85**	**1.21**	**0.77**	**2.77**	**2.18**
	E-High	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.17	0.24	1.46
NC	**True**	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.15	0.20	0.95
	R1	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.14	0.19	0.89
	R2	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.14	0.19	0.89
	E-Low	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.13	0.17	0.78
	E-High	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.17	0.23	1.20

Note: "C" stood for *model with constraints* condition. "NC" stood for *model without constraints* condition. "R1" stood for the first *Random* starting value condition. "R2" stood for the second *Random* starting value condition. "E-low" stood for the *Extremely Low* starting value condition. "E-High" stood for the *Extremely High* starting value condition. "RMSE" stood for root mean squared of error. "s.e." stood for standard error.

Table 8 Parameter Recovery for the Structural Model (Local Optima)

Model Constraints	Starting Values	Replications	Main Effects				Interactions				Kappa	
			Bias		RMSE		Bias		RMSE			
			mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*
C	**True**	111	−0.02	0.08	0.01	0.02	0.01	0.05	0.04	0.05	0.81	0.09
	R1	113	−0.02	0.08	0.01	0.02	0.01	0.05	0.04	0.05	0.80	0.09
	R2	120	−0.02	0.08	0.01	0.02	0.01	0.05	0.04	0.05	0.80	0.09
	E-Low	**751**	**−0.30**	**0.18**	**0.23**	**0.31**	**0.26**	**0.15**	**0.48**	**0.63**	**0.73**	**0.07**
	E-High	122	−0.02	0.08	0.01	0.02	0.01	0.05	0.04	0.05	0.80	0.08
NC	**True**	7	0.00	0.03	0.00	0.00	−0.01	0.02	0.01	0.00	0.91	0.01
	R1	7	0.00	0.02	0.00	0.00	0.00	0.03	0.01	0.01	0.91	0.02
	R2	7	0.01	0.02	0.00	0.00	−0.01	0.02	0.01	0.01	0.91	0.01
	E-Low	7	0.01	0.02	0.00	0.00	−0.01	0.03	0.01	0.01	0.91	0.02
	E-High	3	−0.02	0.08	0.01	0.02	0.01	0.05	0.04	0.05	0.81	0.09

Note: "C" stood for *model with constraints* condition. "NC" stood for *model without constraints* condition. "R1" stood for the first *Random* starting value condition. "R2" stood for the second *Random* starting value condition. "E-low" stood for the *Extremely Low* starting value condition. "E-High" stood for the *Extremely High* starting value condition. "RMSE" stood for root mean squared of error. "s.e." stood for standard error.

The converged replications were divided into two groups, "Local Optima" and "NON Local Optima", within the starting values level, based on whether the likelihood from that estimation was smaller than the maximum likelihood of all 10 estimations. This table summarized results from the "Local Optima" group.

Table 9 Parameter Recovery for the Measurement Model (Local Optima)

Model Constraints	**Starting Values**	**Intercepts**				**Main Effects**				**Interactions**			
		Bias		RMSE		Bias		RMSE		Bias		RMSE	
		mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*
C	**True**	0.00	0.02	0.00	0.01	0.01	0.03	0.02	0.01	0.09	0.37	1.01	3.82
	R1	0.00	0.02	0.00	0.01	0.01	0.03	0.02	0.01	0.04	0.23	0.50	2.21
	R2	0.00	0.02	0.00	0.01	0.01	0.03	0.02	0.02	0.02	0.18	0.34	1.55
	E-Low	**0.15**	**0.07**	**0.09**	**0.06**	**−0.76**	**0.27**	**1.36**	**0.73**	**1.51**	**0.54**	**3.40**	**1.86**
	E-High	0.00	0.02	0.00	0.01	0.01	0.03	0.02	0.01	0.07	0.34	0.78	3.42
NC	**True**	0.00	0.01	0.00	0.00	0.00	0.02	0.01	0.01	0.73	0.32	3.91	2.10
	R1	0.00	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.61	0.19	3.50	1.62
	R2	0.00	0.01	0.00	0.00	0.00	0.02	0.02	0.01	0.56	0.19	2.84	1.35
	E-Low	0.00	0.01	0.00	0.00	0.01	0.03	0.02	0.01	0.59	0.17	3.87	2.09
	E-High	0.00	0.02	0.00	0.01	0.01	0.03	0.02	0.01	0.09	0.37	1.01	3.82

Note: "C" stood for *model with constraints* condition. "NC" stood for *model without constraints* condition. "R1" stood for the first *Random* starting value condition. "R2" stood for the second *Random* starting value condition. "E-low" stood for the *Extremely Low* starting value condition. "E-High" stood for the *Extremely High* starting value condition. "RMSE" stood for root mean squared of error. "s.e." stood for standard error.

The converged replications were divided into two groups, "Local Optima" and "NON Local Optima", within the starting values level, based on whether the likelihood from that estimation was smaller than the maximum likelihood of all 10 estimations. This table summarized results from the "Local Optima" group.

Table 10 Parameter Recovery for the Structural Model (NON Local Optima)

Model Constraints	**Starting Values**	**Replications**	**Main Effects**				**Interactions**				**Kappa**	
			Bias		RMSE		Bias		RMSE			
			mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*
C	**True**	867	−0.02	0.04	0.00	0.00	0.01	0.04	0.01	0.01	0.92	0.01
	R1	825	0.00	0.06	0.01	0.01	0.00	0.04	0.03	0.04	0.83	0.07
	R2	801	0.00	0.06	0.01	0.01	0.00	0.04	0.03	0.04	0.83	0.07
	E-Low	200	0.00	0.06	0.01	0.01	0.00	0.04	0.03	0.04	0.83	0.07
	E-High	850	−0.03	0.13	0.04	0.22	0.03	0.12	0.09	0.34	0.82	0.10
NC	**True**	973	0.00	0.06	0.01	0.01	0.00	0.04	0.03	0.04	0.83	0.07
	R1	972	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	R2	972	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	E-Low	970	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08
	E-High	980	0.00	0.07	0.01	0.01	0.00	0.04	0.03	0.04	0.82	0.08

Note: “C” stood for *model with constraints* condition. “NC” stood for *model without constraints* condition. “R1” stood for the first *Random* starting value condition. “R2” stood for the second *Random* starting value condition. “E-low” stood for the *Extremely Low* starting value condition. “E-High” stood for the *Extremely High* starting value condition. “RMSE” stood for root mean squared of error. “s.e.” stood for standard error.

The converged replications were divided into two groups, “Local Optima” and “NON Local Optima”, within the starting values level, based on whether the likelihood from that estimation was smaller than the maximum likelihood of all 10 estimations. This table summarized results from the “NON Local Optima” group.

Table 11 Parameter Recovery for the Measurement Model (NON Local Optima)

Model Constraints	**Starting Values**	**Intercepts**				**Main Effects**				**Interactions**			
		Bias		RMSE		Bias		RMSE		Bias		RMSE	
		mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*	mean	*s.e.*
C	**True**	0.00	0.01	0.00	0.00	0.00	0.01	0.02	0.01	0.65	0.22	3.58	2.71
	R1	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.14	0.19	1.11
	R2	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.12	0.15	0.82
	E-Low	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.11	0.12	0.51
	E-High	0.00	0.03	0.01	0.02	-0.05	0.24	0.13	0.51	0.11	0.40	0.43	1.60
NC	**True**	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.03	0.13	0.17	0.85
	R1	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.13	0.17	0.89
	R2	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.13	0.17	0.84
	E-Low	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.13	0.17	0.85
	E-High	0.00	0.02	0.00	0.01	0.00	0.03	0.02	0.01	0.02	0.12	0.14	0.70

Note: "C" stood for *model with constraints* condition. "NC" stood for *model without constraints* condition. "R1" stood for the first *Random* starting value condition. "R2" stood for the second *Random* starting value condition. "E-low" stood for the *Extremely Low* starting value condition. "E-High" stood for the *Extremely High* starting value condition. "RMSE" stood for root mean squared of error. "s.e." stood for standard error.

The converged replications were divided into two groups, "Local Optima" and "NON Local Optima", within the starting values level, based on whether the likelihood from that estimation was smaller than the maximum likelihood of all 10 estimations. This table summarized results from the "NON Local Optima" group.

Table 12 Local Optima Prevalence within the Data Level

Replication	Convergence	Local Optima					
		Total		C		NC	
#	#	#	%	#	%	#	%
1000	986	115	11.66	115	11.66	0	0.00

Note: For each data set, the maximum likelihood from the five estimations in the *model with constraints* ("C") condition was compared to the maximum likelihood from those five estimations in the *model without constraints* ("NC") condition. Local optima were identified for the condition with a smaller maximum likelihood. In order to compare, at least one estimation from both model constraints conditions had to converge, which became the convergence criterion at this level. "#" indicated the number. "%" indicated the percentage.

Table 13 Local Optima Prevalence and Point-Biserial Correlation within the Model Constraints Level

Model Constraints	**Replication** #	**Convergence** #	**Local Optima** #	%	**Point-Biserial Correlation** Mastery Base Rate	Attribute Correlation	Item Effect Size
C	1000	982	734	74.75	−.−0.22	−.−0.09	−.−0.16
NC	1000	982	3	0.31	−.−0.09	−.−0.08	0.06

Note: “C” represented for the *model with constraints* condition. “NC” represented for the *model without constraints* condition. “#” indicated the number. “%” indicated the percentage.

For each data set, within each model constraints condition, the likelihoods from the five estimations were compared to each other. Local optima were identified if there were more than two unique likelihoods within each model constraints condition. In order to compare, at least two estimations from each model constraints condition had to converge, which became the convergence criterion at this level. It was a stricter convergence criterion than the data level. Furthermore, the point-biserial correlation between local optima and three continuous variables for each model constraints condition were listed.

Table 14 Local Optima Prevalence within the Starting Values Level

Model Constraints	Starting Values	Replication	Convergence	Local Optima (10)		Local Optima (5)	
				#	%	#	%
C	**True**	1000	978	111	***11.35***	6	***0.61***
	Random 1	1000	938	113	***12.05***	18	***1.92***
	Random 2	1000	921	120	***13.03***	23	***2.50***
	Extremely Low	1000	951	751	***78.97***	726	***76.34***
	Extremely High	1000	972	122	***12.55***	24	***2.47***
NC	**True**	1000	980	7	***0.71***	6	***0.61***
	Random 1	1000	979	7	***0.72***	6	***0.61***
	Random 2	1000	979	7	***0.72***	6	***0.61***
	Extremely Low	1000	977	7	***0.72***	6	***0.61***
	Extremely High	1000	983	3	***0.31***	0	***0.00***

Note: "C" represented for the *model with constraints* condition. "NC" represented for the *model without constraints* condition. "#" indicated the number. "%" indicated the percentage.

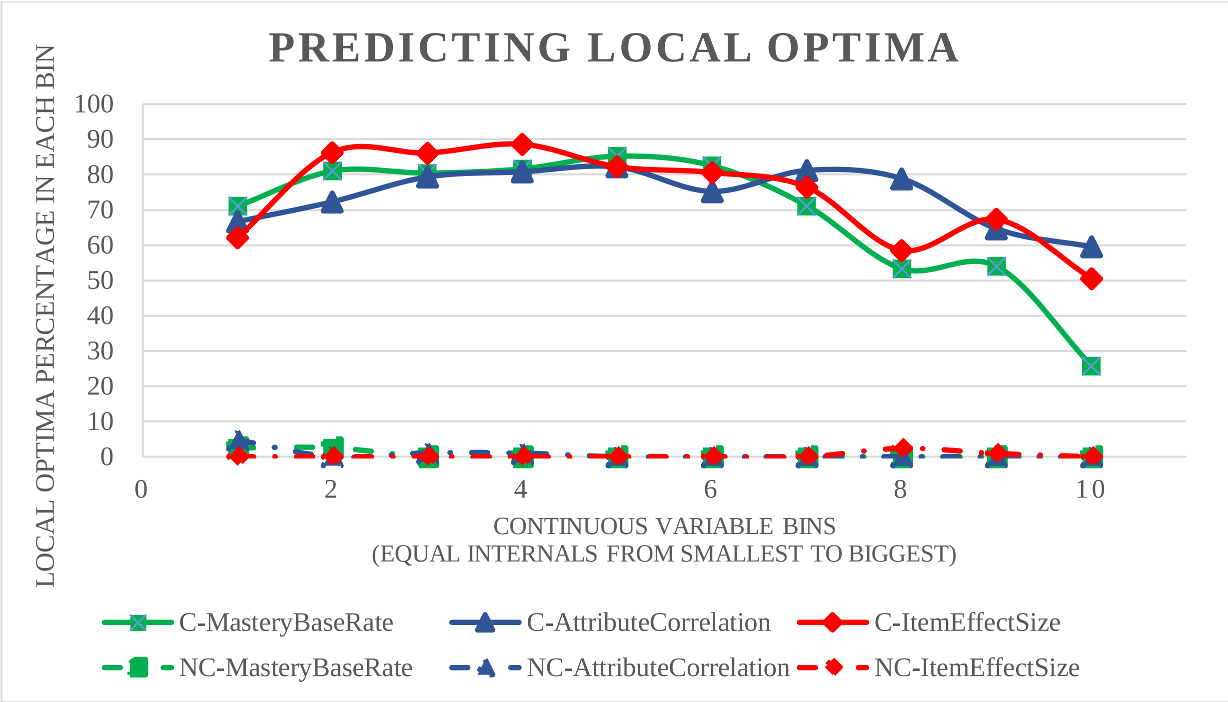
For each starting values condition, the convergence was counted for each estimation, separating the two model constraints conditions. The local optima were identified at both the data level ("Local Optima (10)") and the model constraints level ("Local Optima (5)"). For the data level, if any of the 10 estimations converged, local optima were identified for that specific condition if its likelihood was smaller than the maximum likelihood among the 10 estimations for the same data set. It was a stricter rule to identify local optima, because the maximum likelihood from 10 estimations was the standard to be compared to, whereas at the model constraints level, the standard was the maximum likelihood of five estimations from the same model constraints condition.

Table 15 Predicting Local Optima

			1	2	3	4	5	6	7	8	9	10
C	**Marginal**	Total	38	79	138	147	149	149	111	90	50	31
	Mastery	Local Optima	27	64	111	120	127	123	79	48	27	8
		%	***71.05***	***81.01***	***80.43***	***81.63***	***85.23***	***82.55***	***71.17***	***53.33***	***54.00***	***25.81***
	Tetrachoric	Total	21	65	92	99	101	109	117	142	147	89
	Correlation	Local Optima	14	47	73	80	83	82	95	112	95	53
		%	***66.67***	***72.31***	***79.35***	***80.81***	***82.18***	***75.23***	***81.20***	***78.87***	***64.63***	***59.55***
	Item Effect Size	Total	95	95	101	123	107	98	85	82	107	89
		Local Optima	59	82	87	109	88	79	65	48	72	45
		%	***62.11***	***86.32***	***86.14***	***88.62***	***82.24***	***80.61***	***76.47***	***58.54***	***67.29***	***50.56***
NC	**Marginal**	Total	40	79	137	146	149	149	111	90	50	31
	Mastery	Local Optima	1	2	0	0	0	0	0	0	0	0
		%	***2.50***	***2.53***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***
	Tetrachoric	Total	21	67	92	98	100	109	117	142	147	89
	Correlation	Local Optima	1	0	1	1	0	0	0	0	0	0
		%	***4.76***	***0.00***	***1.09***	***1.02***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***
	Item Effect Size	Total	95	95	101	123	106	98	84	84	108	88
		Local Optima	0	0	0	0	0	0	0	2	1	0
		%	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***0.00***	***2.38***	***0.93***	***0.00***

Note: The three continuous variables were categorized into 10 equal interval bins, respectively, with values in Bin 1 as the smallest and values in Bin 10 as the biggest. The exact values for each Bin differ across variables. The total number of replications may differ across bins.

Figure 1 Predicting Local Optima Plot



Note: Only the points were actual data from the simulation. Lines were artificially added to represent one possible trend between points. The X axis represents the ten equal-interval bins of the three continuous predictors, from smallest values as Bin 1 and the biggest values as Bin 10. “C” stood for the *model with constraints* condition. “NC” stood for *model without constraints* condition.