PSYCHOMETRIKA—VOL. 83, NO. 4, 963–990
DECEMBER 2018
https://doi.org/10.1007/s11336-018-9638-5

# ASSESSING GROWTH IN A DIAGNOSTIC CLASSIFICATION MODEL FRAMEWORK

MATTHEW J. MADISON
CLEMSON UNIVERSITY
LAINE P. BRADSHAW
UNIVERSITY OF GEORGIA

A common assessment research design is the single-group pre-test/post-test design in which examinees are administered an assessment before instruction and then another assessment after instruction. In this type of study, the primary objective is to measure growth in examinees, individually and collectively. In an item response theory (IRT) framework, longitudinal IRT models can be used to assess growth in examinee ability over time. In a diagnostic classification model (DCM) framework, assessing growth translates to measuring changes in attribute mastery status over time, thereby providing a categorical, criterion-referenced interpretation of growth. This study introduces the Transition Diagnostic Classification Model (TDCM), which combines latent transition analysis with the log-linear cognitive diagnosis model to provide methodology for analyzing growth in a general DCM framework. Simulation study results indicate that the proposed model is flexible, provides accurate and reliable classifications, and is quite robust to violations to measurement invariance over time. The TDCM is used to analyze pre-test/post-test data from a diagnostic mathematics assessment.

Key words: diagnostic classification model, cognitive diagnosis model, latent transition analysis, item parameter drift, measurement invariance, growth, pre-test/post-test design.

## 1. Introduction

Diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010; see also Bradshaw, 2016) are becoming an increasingly viable alternative to traditional measurement models. The distinguishing feature of DCMs are the discrete latent traits, or *attributes*, in place of the continuous latent traits that underlie more commonly applied measurement models, such as item response theory (IRT) models. In educational contexts, DCM attributes can be used to indicate the knowledge or proficiency level of each examinee. Fulfilling the desires of educators for more detailed and actionable feedback from assessments (Huff & Goodman, 2007), results from DCM analyses can guide instruction by pinpointing students' strengths and weaknesses regarding the assessed domain.

A common assessment research design is the pre-test/post-test design in which examinees are administered an assessment before instruction and then another assessment after instruction. In this type of study, the primary objective is to measure growth in examinees, individually and collectively. In an IRT framework, longitudinal IRT models can be used to assess growth in examinee ability over time (Andersen, 1985; Embretson, 1991; Fischer, 1976, 1989). In a DCM framework, assessing growth translates to measuring changes in attribute mastery status over time, thereby providing a categorical, criterion-referenced interpretation of growth. The focus of this study is to develop and examine methodology to accommodate longitudinal assessment data in a general DCM framework.

Correspondence should be made to Matthew J. Madison, Department of Education and Human Development, Clemson University, 226 Holtzendorff Hall, Clemson, SC 29634, USA. Email: mjmadis@clemson.edu

For latent class models, of which DCMs are constrained versions, there exists methodology to assess change in latent class prevalence over time. Namely, latent transition analysis (LTA; Collins & Wugalter, 1992) can be used to model change over time in a latent class model framework. In addition to the parameters estimated in a traditional latent class analysis, which include latent class proportions and item parameters, LTA models how members of each latent class transition from one latent class to another between each measurement occasion. LTA has been used to study dating and sexual behaviors (Lanza & Collins, 2008), substance use (Collins & Lanza, 2010; Lanza, Patrick, & Maggs, 2010), and the effectiveness of medical interventions (Roberts & Ward, 2011), to name a few examples.

In addition to LTA, Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1995) is another method for assessing change over time in discrete latent traits. BKT is a model used in intelligent tutoring systems to update estimates of user knowledge components as they progress through learning modules. 'Items' in an intelligent tutoring system are moves or actions that a user makes. User moves may include selecting an answer option, inputting a free response, or requesting a hint. BKT is useful in intelligent tutoring systems because the updated estimate can be used to guide the system's next move (e.g., next item, hint, feedback) in an effort to individualize the learning experience for each user. While the BKT modeling framework is similar in form to using LTA, the difference lies in the overall goal of the model. In BKT, the goal is to assess student learning from item to item *within* an assessment. This study proposes a model combining LTA and a DCM, where the goal is to assess student learning *between* assessments.

There has been little research applying DCMs in pre-test/post-test designed studies. Jang (2005) used a DCM to diagnose students' English reading skills before and after instruction. Jang obtained pre- and post-test attribute classifications for 27 students by scoring their responses to different sets of pre-calibrated pre- and post-test items. Jurich and Bradshaw (2014) employed a general DCM to examine four psychoeducational learning outcomes in undergraduate students. They first calibrated item parameters and obtained examinee classifications for the pre-test and then obtained post-test classifications by scoring the post-test item responses with item parameters fixed at their respective pre-test estimates. This modeling approach assumes parameter equality across the pre-test and post-test and, similar to Jang, does not account for pre-test and post-test attribute status dependencies. Li, Cohen, Bottge, and Templin (2015), and Kaya and Leite (2017), combined LTA with constrained DCMs to assess change in attribute mastery. Wang, Yang, Culpepper, and Douglas (2018) used a higher-order hidden Markov model combined with a constrained DCM to evaluate the efficacy of different learning interventions. The DCMs applied by Li et al., Kaya and Leite, and Wang et al. assume particular structures to the item responses and impose extreme parameter constraints a priori across all items and attributes. When these strict assumptions are not met, model classification accuracy decreases and item parameters are inaccurate and misleading (Bradshaw & Templin, 2014). Li et al.'s (2015) and Kaya and Leite's (2017) use of LTA in conjunction with DCMs, and Wang et al.'s (2018) use of a higher-order hidden Markov model are significant advancements for DCM methodology because unlike previous studies, these approaches do account for the dependence in attribute mastery at each time point; however, the use of constrained DCMs limits their application and utility to the special—and likely rare—cases where those strict assumptions are met. This study generalizes previous work by employing a more general DCM in conjunction with LTA.

Additionally, this study examines an important issue with the application of any longitudinal psychometric model: *measurement invariance*. Measurement invariance concerns the extent to which the meaning of the measured construct and its relationship with item responses remains the same over time or across groups. With longitudinal models, measurement invariance is typically assumed so that changes in examinee performance over time can be validly attributed to changes in the examinee latent traits. For IRT models, several studies have noted the detrimental impact of violations of the assumption of measurement invariance (Bock, Muraki, & Pfeiffenberger, 1988;

Han & Guo, 2011; Wells, Subkoviak, & Serlin, 2002; Lee & Cho, 2017). The impact of violations to measurement invariance over time, or *item parameter drift* (IPD; Goldstein, 1983), has not been examined for longitudinal DCMs; this study will contribute to this area of needed research in the DCM literature. More specifically, we examine the robustness of longitudinal DCMs to different levels and types of IPD.

This study advances methodology for assessing growth in a DCM framework by proposing a model that combines LTA with a general DCM, examining the impact of IPD on model performance, and illustrating the proposed model with empirical data. In the next section is a description of the foundations of the proposed model, which are the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009) and latent transition analysis. Then, we present the proposed model, the Transition Diagnostic Classification Model (TDCM). Following that, we move to the simulation study design and empirical analysis description.

## 2. Statistical Foundations of the Transition Diagnostic Classification Model

### *2.1. Log-Linear Cognitive Diagnosis Model*

The LCDM is a general DCM that parameterizes the probability of a correct response as a function of examinee attribute mastery, the attributes measured by the item, and the item parameters. It is similar in form to a multi-dimensional IRT model, with the main difference being that the latent traits are not continuous; rather, they are categorical and commonly dichotomous. In the LCDM, the latent traits are collectively referred to as an *attribute profile*. On a test measuring $A$ attributes, each examinee possesses an $A$ length vector, denoted $\boldsymbol{\alpha_c} = [\alpha_{c1}, \alpha_{c2}, \ldots, \alpha_{cA}]$, where $c$ refers to the index of the specific attribute profile, $\alpha_{ca} = 0$ indicates attribute non-mastery of Attribute $a$, and $\alpha_{ca} = 1$ indicates attribute mastery of Attribute $a$. For example, on a test measuring five attributes, an attribute profile of [1, 1, 0, 1, 0] indicates that the examinee has mastered Attributes 1, 2, and 4 and has not mastered Attributes 3 and 5. The LCDM uses item responses to probabilistically classify each examinee into one of the $2^A$ attribute profiles.

The item parameters in the LCDM have similar interpretations to a reference-coded ANOVA model. That is, for each item measured on a test, the LCDM item response function is comprised of an intercept, a main effect for each attribute measured by the item, and interaction term(s) that correspond to each pair, triplet, etc., of attributes measured by the item. To demonstrate the item response function, consider an item measuring two attributes, Attribute 2 and Attribute 3. The LCDM item response function models the conditional probability of a correct response as

$$P\left(X_{ic} = 1|\boldsymbol{\alpha_c}\right) = \frac{\exp\left(\lambda_{i,0} + \lambda_{i,1,(2)}\left(\alpha_{c2}\right) + \lambda_{i,1,(3)}\left(\alpha_{c3}\right) + \lambda_{i,2,(2,3)}\left(\alpha_{c2} \cdot \alpha_{c3}\right)\right)}{1 + \exp\left(\lambda_{i,0} + \lambda_{i,1,(2)}\left(\alpha_{c2}\right) + \lambda_{i,1,(3)}\left(\alpha_{c3}\right) + \lambda_{i,2,(2,3)}\left(\alpha_{c2} \cdot \alpha_{c3}\right)\right)}. \quad (1)$$

In Eq. 1, $X_{ic}$ represents the random variable for the response to item $i$ by an examinee with attribute profile $\boldsymbol{\alpha_c}$. The first parameter in Eq. 1 is the intercept, $\lambda_{i,0}$. The intercept represents the log-odds of a correct response for the reference group, which is composed of examinees who possess neither Attribute 2 nor Attribute 3. The next parameters, $\lambda_{i,1,(2)}$ and $\lambda_{i,1,(3)}$, are the main effects for Attribute 2 and 3, respectively. They represent the increase in log-odds of a correct response for examinees who have mastered either Attribute 2 or 3, respectively. The last parameter, $\lambda_{i,2,(2,3)}$, is the interaction term and represents the additional change in log-odds of a correct response for examinees who have mastered both Attribute 2 and Attribute 3. The magnitude of these parameters reflects the degree to which attribute mastery statuses affect correct response probabilities.

The example above presents an item measuring two attributes. The LCDM, however, can accommodate up to $A$ attributes on any given item. Computational demands and model estimation

time are the only limitations to the number of attributes specified on an item or test in the LCDM. In its general form, the LCDM item response function can be expressed as

$$P\left(X_{ic}=1|\boldsymbol{\alpha_c}\right)=\frac{\exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{h}\left(\boldsymbol{\alpha}_c,\boldsymbol{q}_i\right)\right)}{1+\exp\left(\lambda_{i,0}+\boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{h}\left(\boldsymbol{\alpha}_c,\boldsymbol{q}_i\right)\right)}, \tag{2}$$

where $\lambda_{i,0}$ is the intercept parameter described above, $\boldsymbol{\lambda}_i^{\mathrm{T}}$ is a column vector containing all possible $2^A-1$ main effect and interactions terms for Item $i$, and $\boldsymbol{q}_i$ represents the $i$th row of the Q-matrix. The Q-matrix is an item-by-attribute matrix of 0s and 1s indicating which attributes are measured by each item. If item $i$ requires Attribute $a$, then cell $ia$ in the Q-matrix will be 1, and 0 otherwise. The function $\boldsymbol{h}\left(\boldsymbol{\alpha}_c,\boldsymbol{q}_i\right)$ results in a column vector of length $2^A-1$ whose elements are linear combinations of $\boldsymbol{\alpha_c}$ and $\boldsymbol{q_i}$. This combination equals 1 only when the examinee possesses the attributes corresponding to a given parameter $\lambda$ and these attributes are measured by Item $i$. If the examinee has not mastered the attributes corresponding to a particular $\lambda$, or the attributes are not measured by the item, then the linear combination will result in a value of 0. When multiplied by the vector of main effect and interaction parameters, $\boldsymbol{\lambda}_i^T$, this general expansion becomes

$$\boldsymbol{\lambda}_i^{\mathrm{T}}\boldsymbol{h}\left(\boldsymbol{\alpha}_c,\ \boldsymbol{q}_i\right)=\sum_{a=1}^{A}\lambda_{i,1,(a)}\alpha_{ca}q_{ia}+\sum_{a=1}^{A-1}\sum_{a'=a+1}^{A}\lambda_{i,2,(a,a')}\alpha_{ca}\alpha_{ca'}q_{ia}q_{ia'}+\cdots. \tag{3}$$

The first summation includes the $\binom{A}{1}=A$ main effect parameters, the double summation includes the $\binom{A}{2}$ two-way interaction parameters, and the ellipses represent all three-way and higher interaction terms. In the estimation of the LCDM, main effect parameters are constrained to be greater than zero so that examinees who have mastered more of the measured attributes on an item have increasing predicted correct response probabilities, and to prevent latent class switching (Lao & Templin, 2016). Interactions may be negative or positive, but are constrained in magnitude so that the mastery of additional attributes cannot result in decreased predicted correct response probabilities.

### 2.2. *Constrained Versions of the LCDM*

The LCDM is a general DCM in that it subsumes many other common DCMs, including the popular deterministic-inputs, noisy-and-gate (DINA; e.g., Haertel, 1989; Junker & Sijtsma, 2001) model. In the DINA model, the main effects in Eq. 1, $\lambda_{i,1,(2)}$ and $\lambda_{i,1,(3)}$, are constrained a priori to be zero. This constraint forces attributes to behave in a non-compensatory fashion, meaning that non-mastery of one measured attribute cannot be compensated for by mastery of other measured attributes. For each item, the DINA model separates all examinees into exactly two groups: those who have mastered all the measured attributes, and those who have not. For example, on the item illustrated in Eq. 1, the DINA model would have two item response probabilities: one for masters of both Attributes 2 and 3, and another for examinees who have not mastered at least one of the required attributes. Because general DCMs such as the LCDM, the General Diagnostic Model (GDM; von Davier, 2005), or the Generalized DINA model (G-DINA; de la Torre, 2011) subsume many DCMs, including the DINA model, they are often better able to accurately represent the attribute mastery processes underlying the item responses. The benefit of using a general model is that attribute and item effects are freed and item parameter estimates can indicate whether these constraints are warranted. As noted earlier, when these constraints are imposed and do not fit

the data, the model can be severely misspecified, which results in misleading item parameters estimates and decreased classification accuracy (Bradshaw & Templin, 2014).

### 2.3. Latent Transition Analysis

LTA is a special case of the latent or hidden Markov model (HMM; Baum & Petrie, 1966) and is a longitudinal extension to the latent class model. In LTA, class membership at each time point is latent, but measured with a set of observed item responses. The measurement model that parameterizes item response probabilities in LTA at each time point is a latent class model, and class-sequential progressions are provided via latent class transition probabilities from each time point to the next. In a typical latent class analysis, the number of latent classes is determined in an exploratory fashion; several models are fit with different numbers of latent classes, and the model with the best fit (i.e., greatest parsimony) is chosen and subsequently interpreted (e.g., see Collins & Lanza, 2010; Lanza, et al., 2010). Similarly, in an LTA, the number of latent classes at each point is determined through comparisons of LTAs with different numbers of latent classes at each time point.

The LTA model consists of three groups of parameters. The first are the probabilities of belonging to each latent class at the first time point. The second group of parameters consists of the probabilities of transitioning from one latent class to another across each time point. Lastly, the LTA model employs a measurement model to estimate the item response probabilities at each time point. Together, these parameters model the likelihood of starting in a certain latent class, and transitioning to and from other latent classes between each successive time point.

## 3. Transition Diagnostic Classification Model

The proposed model combines LTA with the LCDM as the measurement model. We refer to this model as the Transition Diagnostic Classification Model (TDCM). In the same way that a DCM is a confirmatory latent class model with latent classes specified a priori as the attribute profiles, the TDCM is a confirmatory LTA with the latent classes at each time point specified a priori as the attribute profiles. Consider an examinee $e$ responding to $I$ items over $T$ testing occasions. In the general form of the TDCM, the probability of the item response vector $\boldsymbol{x}_e$ is given by:

$$P\left(\boldsymbol{X}_e = \boldsymbol{x}_e\right) = \sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\cdots\sum_{c_T=1}^{C} \nu_{c_1}\tau_{c_2|c_1}\tau_{c_3|c_2}\cdots\tau_{c_T|c_{T-1}}\prod_{t=1}^{T}\prod_{i=1}^{I}\pi_{ic_t}^{x_{eit}}\left(1-\pi_{ic_t}\right)^{1-x_{eit}}. \quad (4)$$

In Eq. 4, $\nu_{c_1}$ represents the probability of membership in Attribute Profile $c$ at Time Point 1. Each sum ranges over each of the $C$ attribute profiles at each testing occasion. The first product term ranges over the $T$ testing occasions, and the second product terms ranges over the $I$ items. Each $\tau_{c_t|c_{t-1}}$ represents the probability of transitioning between different attribute mastery statuses between Testing Occasion $t-1$ to Testing Occasion $t$; the $\pi_{ic_t}$'s are the item response probabilities, which are estimated with the LCDM described in Sect. 2.1; and $x_{eit}$ is Examinee $e$'s response to Item $i$ at Testing Occasion $t$.

In Li et al.'s study, the DINA model is used in conjunction with LTA; this hybrid model is referred to as the LTA-DINA model (2015). Kaya and Leite (2017) use the deterministic input, noisy-or-gate model (DINO; Templin & Henson 2006) model in conjunction with LTA, which they refer to as the longitudinal DINO model. As noted earlier, the DINA and DINO models are submodels or special cases of the LCDM that can be obtained by constraining parameters in the

TABLE 1.
Example transition probability matrix for four attribute profiles.

| | Attribute profile | Post-test | | | |
|---|---|---|---|---|---|
| | | [0, 0] | [0, 1] | [1, 0] | [1, 1] |
| Pre-test | [0, 0] | .20 | .18 | .35 | .27 |
| | [0, 1] | .10 | .25 | .05 | .60 |
| | [1, 0] | .05 | .15 | .10 | .70 |
| | [1, 1] | .02 | .01 | .01 | .96 |

LCDM. Therefore, the TDCM is a generalization of the LTA-DINA model developed by Li, et al. and the longitudinal DINO model developed by Kaya and Leite. Furthermore, in developing the TDCM, we are developing LTA hybrid versions of all other DCMs subsumed by the LCDM. Among other general DCMs, the LCDM was selected because our study uses binary item response data, so we used the LCDM's corresponding canonical logit link function and because it can be estimated in Mplus (Muthén & Muthén, 2012; e.g., Templin & Hoffman, 2013). Combined with its LTA capabilities, Mplus was capable of estimating the TDCM. In Appendix, we have included abbreviated Mplus syntax for the TDCM.

### *3.1. Attribute Mastery Transitions*

In a pre-test/post-test design (two time points) with $A$ attributes, the transition probabilities would be displayed in $2^A \times 2^A$ matrix with the $jk$th cell representing the probability of transitioning from Attribute Profile $j$ to Attribute Profile $k$. To illustrate, Table 1 shows an example transition matrix for a two-attribute test. Attribute profiles [0, 0] , [0, 1] , [1, 0] , and [1, 1] would serve as the four latent classes at each time point in the TDCM. Here, the probability of transitioning from attribute profile [0, 0] at pre-test into attribute profile [1, 1] at post-test is .27, while the probability of transitioning from [0, 1] to [1, 1] is .60. Although this transition matrix is hypothetical, note that transitions from attribute profiles with more mastered attributes to a profile with fewer mastered attributes also are possible, but generally less likely. For example, the probability of transitioning from the attribute profile [0, 0] to [1, 0] is .35, while the probability of transitioning from [1, 0] to [0, 0] is much smaller at .05. In educational contexts, we would expect this to be the case because students are more likely to obtain or retain knowledge over short periods of targeted instruction than they are to forget or unlearn knowledge (e.g., see Li et al., 2015). In other contexts, such as adolescent depression or substance abuse patterns, one might expect, and studies have observed, more uniform transition probabilities (e.g., see Collins & Lanza, 2010).

The transition probability matrix can be used to evaluate learning across the examinee sample. For example, in Table 1, 80% (18% + 35% + 27%) of examinees starting as non-masters of both attributes transitioned into mastery of one or both attributes. This could be interpreted as an indicator of successful learning. In this two-attribute example, this table is still easily interpretable. As the number of specified attributes becomes large, however, this matrix can quickly become unruly and difficult to interpret. In an educational assessment context, the individual attribute mastery transitions are often of interest because they can be used to evaluate learning for each measured attribute. This information can then be used to guide instructional adjustments. The transition probability matrix, combined with the pre-test attribute profile membership proportions, can be used to compute individual attribute mastery transitions. For example, suppose in Table 1 example that the pre-test attribute profile membership proportions are .40, .25, .20, and .15 for [0, 0] , [0, 1] , [1, 0] , and [1, 1], respectively. Table 2 shows the resulting Attribute 1

TABLE 2.
Example Attribute 1 conditional transition matrix.

| | Mastery state | Post-test | |
|---|---|---|---|
| | | 0 | 1 |
| Pre-test | 0 | .37 | .63 |
| | 1 | .13 | .87 |

transition probability matrix. Focusing on individual attribute mastery transitions provides a more interpretable result. The most likely transition is remaining a master of Attribute 1 at .87. The least likely transition is losing or forgetting Attribute 1 at .13. Note that each row sums to one, as the row probabilities are conditional on the attribute mastery state at pre-test.

### *3.2. Pre-test/Post-test Form of TDCM*

For the pre-test/post-test design, the focus of this paper, Eq. 4 reduces to

$$P\left(\boldsymbol{X}_e = \boldsymbol{x}_e\right) = \sum_{c_1=1}^{C}\sum_{c_2=1}^{C} \nu_{c_1}\tau_{c_2|c_1}\prod_{t=1}^{2}\prod_{i=1}^{I}\pi_{ic_t}^{x_{eit}}\left(1-\pi_{ic_t}\right)^{1-x_{eit}}. \tag{5}$$

Here, there are only parameters for two testing occasions. Therefore, $\nu_{c_1}$ represents the pre-test attribute profile proportions and the $\tau$'s represent the attribute mastery transitions from pre- to post-test.

### *3.3. Measurement Invariance in the TDCM*

Generally, measurement invariance exists when the distribution of scores given the latent trait is the same for all groups (Meredith & Millsap, 1992). In practice, measurement invariance for psychometric models is operationalized and tested by examining item parameter invariance. In the case of longitudinal analyses, measurement invariance exists when item parameters for each testing occasion are equal. Mapped on to the TDCM framework, measurement invariance over time exists when the distribution of item responses, conditional on attribute mastery status, is identical over time. More formally, measurement invariance exists in the TDCM if and only if

$$P\left(\boldsymbol{X}_{e1} = \boldsymbol{x}_{e1}|\boldsymbol{\alpha}_c\right) = P\left(\boldsymbol{X}_{e2} = \boldsymbol{x}_{e2}|\boldsymbol{\alpha}_c\right) = \cdots = P\left(\boldsymbol{X}_{eT} = \boldsymbol{x}_{eT}|\boldsymbol{\alpha}_c\right) \tag{6}$$

for all time points $t = 1, \ldots, T$ and all attribute profiles $\boldsymbol{\alpha}_c$. In plain terms, assuming measurement invariance, the TDCM item response function is identical across each individual time point.

In the pre-test/post-test designed study, there are two cases to consider: same items at pre- and post-test and different items at pre- and post-test. In the first case that the same set of items is administered at pre- and post-test, it is customary to constrain the pre-test item parameters to be equal to the post-test item parameters (Collins & Lanza, 2010). All studies using longitudinal DCMs, either in simulation or empirical analyses, have assumed measurement invariance over time (Li et. al, 2015; Kaya & Leite, 2017; Wang et. al, 2018). When pre- and post-test item parameters are constrained to be equal, Eq. 5 reduces to:

$$P\left(\boldsymbol{X}_e = \boldsymbol{x}_e\right) = \sum_{c_1=1}^{C}\sum_{c_2=1}^{C} \nu_{c_1}\tau_{c_2|c_1}\prod_{t=1}^{2}\prod_{i=1}^{I}\pi_{ic}^{x_{eit}}\left(1-\pi_{ic}\right)^{1-x_{eit}}. \tag{7}$$

Notice that the item response probabilities in Eq. 7, the $\pi_{ic}$'s, are no longer subscripted with a $t$ because the equality constraint forces them to be time invariant. In the TDCM, constraining parameters to reflect measurement invariance over time ensures that attribute mastery retains the same meaning over time points and ensures that changes in examinee performance over time can be attributed to changes in examinee mastery. To the degree that measurement invariance is violated, these constraints can introduce measurement error. In particular, if items behave differently and have different item parameters at post-test, constraining the item parameters to be equal introduces error. The effects of IPD have not been examined for longitudinal DCMs. Therefore, in the next section, we conduct a simulation study to systematically examine the robustness of the TDCM to IPD. That is, we assess the performance of the TDCM when measurement invariance is assumed, but IPD is present. To do so, we evaluate the performance of the item parameter-constrained TDCM in Eq. 7 under different types and magnitudes of IPD.

In the case that examinees are given different items at pre- and post-test, the different items' parameters would need to be estimated at each time point without equality constraints, and we would have to rely on the item invariance of LCDM classifications (Bradshaw & Madison, 2016). The LCDM item invariance property states that LCDM classifications are independent of the particular set of items administered. Unlike IRT, when the scale must be set with common items, the scale for DCMs in non-arbitrary, which makes common items unnecessary. This property affords some flexibility in the application of the TDCM as pre- and post-test items can be different, and no common items are necessary. As the case of different sets of items has not been studied for longitudinal DCMs, the simulation study in the next section also examines the performance of the freely estimated TDCM in the case of different items.

## 4. Simulation Study

The simulation study consists of two parts. Part I investigates the effects of IPD in the case that the same items are administered at both time points. In Part I, the parameter-constrained TDCM (see Eq. 7; which assumes parameter invariance) is estimated in the presence of IPD. The objective is to examine the impact this model misspecification, which ignores the IPD. Part II examines the performance of the TDCM in the case that different items are administered at pre- and post-test. In Part II, the correct model, the freely estimated TDCM (see Eq. 5), is specified. The objective in Part II is to demonstrate the LCDM item invariance property in the context of a pre-test/post-test designed study with different items.

### *4.1. Simulation Study Part I Design*

The design of Part I was informed by preliminary TDCM analyses, published LCDM simulation studies, published IRT IPD studies, and existing diagnostic tests. Additionally, we chose conditions that are realistic and obtainable for researchers conducting pre-test/post-test designed studies. We manipulated factors most pertinent to the primary inquiry: the impact of IPD. Manipulated factors included sample size, the type of IPD, magnitude of IPD, and number of IPD items. To keep the simulation manageable, we chose to fix factors that either have been studied in previous research studies or were not expected to moderate the impact of IPD. Fixed factors include the Q-matrix design, attribute correlations, attribute pre-test and post-test mastery base-rates, and attribute mastery transition probabilities and growth. These fixed and manipulated simulation conditions are described in detail below.

*4.1.1. Q-matrix Design* We chose a 4-attribute, 20-item Q-matrix design. The Q-matrix was of moderate complexity, with 60% of the items measuring exactly one attribute, and the other 40% of the items measuring two attributes. The Q-matrix was balanced in that each attribute was measured seven times. This moderate complexity Q-matrix was selected to provide a testing scenario where the TDCM provides utility beyond the LTA-DINA model. For a simple structure Q-matrix, the LTA-DINA model is equivalent to the TDCM, and their performance would be identical. We do not estimate the LTA-DINA or longitudinal DINO models in this study; however, statistical logic suggests, and previous research has shown (Bradshaw & Templin, 2014), that when item responses are not DINA or DINO generated, the flexibility of the LCDM (and hence, TDCM) provides benefits with respect to classification quality and detection of item parameter effects that the DINA and DINO model (and hence, LTA-DINA model and longitudinal DINO) cannot provide.

*4.1.2. Attribute Pre-test Base-Rates and Correlations* An attribute base-rate is the proportion of examinees who are masters of the attribute. Regarding attribute base-rates, we assumed that the majority of examinees were non-masters at pre-test. Therefore, all four attribute base-rates were fixed at .40 at the pre-test occasion. Examinee pre-test attribute profiles were generated with a moderate attribute correlation of .50.

*4.1.3. Marginal Transition Probabilities and Attribute Mastery Growth* To vary growth in attribute mastery from pre-test to post-test, we manipulated transition probabilities for each attribute. Combined with the pre-test attribute base-rates, the marginal attribute transition probabilities were chosen to produce different amounts of growth in attribute mastery proportions. With respect to marginal attribute transitions, we assumed that over the course of a relatively short instructional period, examinees are unlikely to unlearn or forget attributes from pre-test to post-test. Therefore, the transition probabilities for attribute mastery loss were fixed at .15 for each attribute. In our preliminary analyses, we found that in small sample conditions, 5% growth was infrequently detected by a Wald test of growth in attribute mastery proportion over time, while 20% growth was detected virtually all the time. Therefore, marginal transition probabilities were manipulated to produce different levels of growth in attribute mastery corresponding to mastery growth effects of 0, .05, .10, and .15 for Attributes 1–4, respectively.

*4.1.4. Item Parameters and IPD* As IPD is the focus on this simulation study, this is where most of the condition manipulations occurred. Item parameters for the pre-test were fixed with intercepts at $-2$, main effects on simple structure items were set at 3, main effects on complex items were set at 2, and interaction effects were set at 1. IPD items were introduced at the post-test. We first manipulated the magnitude of IPD. In the low-magnitude IPD conditions, the post-test item parameters were $\pm 0.5$ relative to the pre-test parameters. In the high-magnitude IPD conditions, the post-test item parameters were $\pm 1$ relative to the pre-test parameters. We also manipulated the proportion of IPD items. Included were conditions for 0%, 20%, 40%, 60%, 80%, and 100% IPD items. Lastly, we manipulated the type of IPD. In the IRT IPD literature, there are different types of IPD pertaining to which particular parameters are drifting (Holland & Wainer, 1993). To create analogous conditions here, we created three types of DCM IPD: intercept only, main effects and interactions only, and all effects. In the intercept only conditions, IPD was introduced only for LCDM intercepts. In the main effects and interactions only conditions, IPD was introduced only for LCDM main effects and interactions. And in the all parameters conditions, IPD was introduced for all parameter effects (LCDM intercepts, main effects, and interactions).

*4.1.5. Sample Size* Two sample sizes were included (500 and 2000). A sample size of 500 was used to reflect samples attainable in research studies and 2000 was used as a sample attainable in large scale research studies or large scale operational assessments. These sample size conditions are similar to those used in a recent IRT IPD study by Lee and Cho (2017).

*4.1.6. Data Generation and Estimation* There were a total of 72 fully crossed conditions (2 sample sizes, 6 IPD proportions, 2 IPD magnitudes, 3 IPD types). Data were generated in R, version 3.1.1, and all models were estimated in Mplus, Version 7 (Muthén & Muthén, 2012). In Appendix, we include some abbreviated Mplus syntax for the TDCM. More is available upon request from the first author and at the first author's Web site www.matthewmadison.com. Each of the 72 conditions was replicated 100 times.

### *4.2. Simulation Study Part I Results*

For the results of Part I, we summarize item parameter recovery, classification accuracy, and classification reliability for each IPD condition. Then, we evaluate the performance of the likelihood ratio test and information criteria (AIC, BIC, sample-size-adjusted BIC) when comparing the relative fit of parameter-constrained TDCM to the freely estimated TDCM.

*4.2.1. Item Parameter Recovery* Table 3 displays the median absolute deviation (MAD) of the estimated item parameters across sample size and IPD conditions for the parameter-constrained TDCM. We calculated the MAD as

$$\text{MAD} = \text{median}\left(\left|\lambda_i - \hat{\lambda}_i\right|\right), \tag{8}$$

where $\lambda_i$ and $\hat{\lambda}_i$ are the true and estimated parameters, respectively. Item parameters were recovered more accurately in the larger sample size condition across the IPD types and IPD magnitudes. As expected, MAD values increased as the IPD proportion increased, and as the IPD magnitude increased. Averaging across sample size conditions, the MAD values for the 100% IPD items condition were increased by factors of 2.0 and 3.8 compared to the 0% IPD items condition for the low- and high-magnitude IPD conditions, respectively. Lastly, there was an interacting effect of IPD type, with the effect of the proportion of IPD items being more pronounced for the intercept-only and all parameters IPD types than for the all parameters IPD type. Although we did observe an effect of IPD, the item parameters were recovered relatively well, with the mean MAD being 0.17 in the low IPD conditions and 0.24 in the high IPD conditions.

*4.2.2. Classification Accuracy* To calculate classification accuracy, we compared the generated attribute mastery statuses to the estimated attribute mastery statuses. Table 4 shows the classification accuracy across IPD and sample size conditions. Because marginal classification accuracy and agreement for each attribute was nearly identical, we averaged the four attributes into a single number for each condition. Results indicate that with respect to classification, the TDCM is robust in the presence of IPD. Even in the case of 100% IPD items, the classification accuracy rate decreased by a maximum of .01 compared to the 0% IPD conditions. There was no noticeable effect of IPD type, and a small effect of IPD magnitude with the reduction in accuracy being slightly larger for the high IPD magnitude conditions. Although item parameter estimates were recovered less accurately when IPD was present, the TDCM was still able to recover examinee classifications accurately.

TABLE 3.
Simulation study part I: item parameter recovery (MAD).

| IPD magnitude | Sample size | Proportion IPD items (%) | IPD type | | |
|---|---|---|---|---|---|
| | | | Intercepts | Main + interactions | All parameters |
| Low | 500 | 0 | 0.156 | 0.156 | 0.156 |
| | | 20 | 0.175 | 0.167 | 0.181 |
| | | 40 | 0.187 | 0.172 | 0.201 |
| | | 60 | 0.206 | 0.175 | 0.229 |
| | | 80 | 0.222 | 0.188 | 0.250 |
| | | 100 | 0.234 | 0.196 | 0.273 |
| | 2000 | 0 | 0.077 | 0.077 | 0.077 |
| | | 20 | 0.081 | 0.079 | 0.090 |
| | | 40 | 0.099 | 0.087 | 0.124 |
| | | 60 | 0.119 | 0.099 | 0.170 |
| | | 80 | 0.144 | 0.118 | 0.215 |
| | | 100 | 0.172 | 0.150 | 0.249 |
| High | 500 | 0 | 0.156 | 0.156 | 0.156 |
| | | 20 | 0.186 | 0.180 | 0.202 |
| | | 40 | 0.216 | 0.192 | 0.265 |
| | | 60 | 0.263 | 0.210 | 0.355 |
| | | 80 | 0.323 | 0.255 | 0.428 |
| | | 100 | 0.390 | 0.321 | 0.492 |
| | 2000 | 0 | 0.077 | 0.077 | 0.077 |
| | | 20 | 0.086 | 0.081 | 0.092 |
| | | 40 | 0.108 | 0.092 | 0.131 |
| | | 60 | 0.141 | 0.108 | 0.228 |
| | | 80 | 0.202 | 0.167 | 0.436 |
| | | 100 | 0.300 | 0.356 | 0.495 |

IPD = item parameter drift; MAD = median absolute deviation; all parameters = intercepts, main effects and interactions.

*4.2.3. Classification Reliability* Reliability for classifications refers to the consistency of the model-based classifications. Reliabilities of the classifications were calculated according to the tetrachoric correlation-based metric defined by Templin and Bradshaw (2013). Table 5 shows the classification reliability for the two models across IPD and sample size conditions. Because marginal classification reliability for each attribute was nearly identical, we averaged the four attributes into a single number for each condition. Similar to the classification accuracy, we noticed a minimal effect of IPD on the parameter-constrained TDCM. Comparing 0% IPD to 100% IPD, the largest reduction in classification reliability for the parameter-constrained TDCM was .008. Together with the classification accuracy results, these results suggest that the TDCM can provide accurate and reliable classifications, even in the presence of IPD.

*4.2.4. Model Selection and Comparisons* In traditional latent transition analyses, as the parameter-constrained model is nested within the unconstrained model, a likelihood ratio test (LRT) can be used to test for measurement invariance. For the TDCM, we considered the LRT and additionally investigated related from two information criteria: Akaike information criterion (AIC) and Bayesian information criterion (BIC). In addition to the LRT and information criteria, we included an effect size measure originally introduced by Raju (1988) and mapped onto the DCM framework by George and Robitzsch (2014). The unsigned area (UA) was originally

TABLE 4.
Simulation study part I: classification accuracy.

| IPD magnitude | Sample size | Proportion IPD items (%) | IPD type | | |
|---|---|---|---|---|---|
| | | | Intercepts | Main + interactions | All parameters |
| Low | 500 | 0 | .946 | .946 | .946 |
| | | 20 | .945 | .945 | .945 |
| | | 40 | .945 | .946 | .945 |
| | | 60 | .944 | .945 | .945 |
| | | 80 | .944 | .944 | .944 |
| | | 100 | .944 | .943 | .944 |
| | 2000 | 0 | .952 | .952 | .952 |
| | | 20 | .951 | .951 | .951 |
| | | 40 | .951 | .952 | .951 |
| | | 60 | .951 | .952 | .951 |
| | | 80 | .951 | .950 | .951 |
| | | 100 | .951 | .950 | .950 |
| High | 500 | 0 | .946 | .946 | .946 |
| | | 20 | .944 | .945 | .945 |
| | | 40 | .943 | .946 | .944 |
| | | 60 | .940 | .945 | .943 |
| | | 80 | .938 | .942 | .941 |
| | | 100 | .936 | .938 | .939 |
| | 2000 | 0 | .952 | .952 | .952 |
| | | 20 | .949 | .951 | .950 |
| | | 40 | .950 | .953 | .951 |
| | | 60 | .947 | .952 | .949 |
| | | 80 | .946 | .949 | .948 |
| | | 100 | .944 | .946 | .946 |

IPD = item parameter drift; all parameters = intercepts, main effects and interactions; classification accuracy was averaged over the four attributes and over the two testing occasions.

created to measure the area between two IRT item characteristic curves. George and Robitzsch (2014) used the UA in a DCM framework to analyze different item functioning in multiple group analyses. When applied with the TDCM, the UA can be used to quantify an item's parameter drift over time. For an item $i$, in the pre-test/post-test context applied here, the $\text{UA}_i$ is calculated as:

$$\text{UA}_i = \sum_{c=1}^{C} w\left(\boldsymbol{\alpha}_c\right) \cdot \left| P\left(X_{i1} = 1 | \boldsymbol{\alpha}_c\right) - P\left(X_{i2} = 1 | \boldsymbol{\alpha}_c\right) \right|, \tag{9}$$

where $w\left(\alpha_c\right) = \frac{1}{2}\left(P\left(\boldsymbol{\alpha}_c | t = 1\right) - P\left(\boldsymbol{\alpha}_c | t = 2\right)\right)$. Here, we employ the average of the $\text{UA}_i$ across all test items as a measure of the total IPD present across the entire test. The mean UA has not been employed as an effect size measure in this context; therefore, the simulation results will provide information about its distribution under different IPD and test conditions.

The results of the LRT, information criteria, and UA effect size are displayed in Table 6. The values displayed for the LRT and information criteria are the proportion of replications that the criterion or test preferred the freely estimated TDCM over the parameter-constrained TDCM. Examining the LRT, we observed that in the 0% IPD item conditions, the Type I error rates were close to the .05 nominal rate: .061 and .031 in the 500 and 2000 sample size conditions,

TABLE 5.
Simulation study part I: classification reliability.

| IPD magnitude | Sample size | Proportion IPD items (%) | IPD type | | |
|---|---|---|---|---|---|
| | | | Intercepts | Main + interactions | All parameters |
| Low | 500 | 0 | .980 | .980 | .980 |
| | | 20 | .980 | .979 | .979 |
| | | 40 | .979 | .980 | .980 |
| | | 60 | .979 | .980 | .979 |
| | | 80 | .979 | .978 | .979 |
| | | 100 | .979 | .978 | .978 |
| | 2000 | 0 | .977 | .977 | .977 |
| | | 20 | .976 | .976 | .976 |
| | | 40 | .976 | .977 | .976 |
| | | 60 | .976 | .976 | .976 |
| | | 80 | .975 | .975 | .975 |
| | | 100 | .975 | .974 | .975 |
| High | 500 | 0 | .980 | .980 | .980 |
| | | 20 | .978 | .979 | .978 |
| | | 40 | .978 | .980 | .979 |
| | | 60 | .976 | .980 | .977 |
| | | 80 | .974 | .977 | .976 |
| | | 100 | .972 | .974 | .974 |
| | 2000 | 0 | .977 | .977 | .977 |
| | | 20 | .975 | .975 | .975 |
| | | 40 | .974 | .977 | .975 |
| | | 60 | .972 | .976 | .974 |
| | | 80 | .970 | .973 | .972 |
| | | 100 | .969 | .970 | .971 |

IPD = item parameter drift; all parameters = intercepts, main effects and interactions; classification reliability was averaged over the four attributes and over the two testing occasions.

respectively. The LRT was very sensitive to departures from measurement invariance. With respect to the information criteria, we observed that the AIC was more sensitive to IPD, preferring the freely estimated TDCM more often than BIC. In the small sample size, low-magnitude IPD conditions, BIC consistently preferred the parameter-constrained TDCM, even with 100% IPD items. BIC appears to only prefer the freely estimated TDCM when IPD is more substantial. For both the likelihood ratio test and the information criterion, we observed a strong sample size effect, where the freely estimated TDCM was chosen more frequently in the 2000 sample size condition than in the 500 sample size condition. Unexpectedly, both information criteria and the LRT preferred the freely estimated TDCM more often in the intercepts only and main + interactions only IPD types than in the all parameters IPD type.

Overall, the effect size measure UA performed mostly as expected, and appropriately for an effect size measure. As the proportion of IPD items increased, so did the UA. Also, on average, the UA for the high-magnitude IPD conditions increased by a factor of 1.5 relative to the low-magnitude IPD conditions. Also, desirable for an effect size measure, it was unaffected by sample size; UA values were approximately equal in across the two sample size conditions. Consistent with results from the information criteria and the LRT, the UA for the all parameter IPD type was consistently less than the intercepts and mains + interactions IPD types. In terms of the values

TABLE 6.
Simulation study part I: measurement invariance model selection.

| IPD magnitude | Sample size | Proportion IPD items (%) | IPD type | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Intercepts | | | | Main + interactions | | | | All parameters | | | |
| | | | AIC | BIC | LRT | UA | AIC | BIC | LRT | UA | AIC | BIC | LRT | UA |
| Low | 500 | 0 | 0 | 0 | .061 | 0.034 | 0 | 0 | .061 | 0.034 | 0 | 0 | .061 | 0.034 |
| | | 20 | .175 | 0 | .691 | 0.043 | .020 | 0 | .418 | 0.040 | 0 | 0 | .214 | 0.035 |
| | | 40 | .915 | 0 | 1 | 0.054 | .371 | 0 | .918 | 0.047 | .071 | 0 | .566 | 0.038 |
| | | 60 | .978 | 0 | .989 | 0.065 | .750 | 0 | .990 | 0.053 | .273 | 0 | .808 | 0.040 |
| | | 80 | 1 | 0 | 1 | 0.075 | .978 | 0 | .989 | 0.068 | .879 | 0 | .990 | 0.046 |
| | | 100 | .990 | 0 | .990 | 0.087 | 1 | .011 | 1 | 0.086 | .990 | 0 | 1 | 0.052 |
| | 2000 | 0 | 0 | 0 | .031 | 0.017 | 0 | 0 | .031 | 0.017 | 0 | 0 | .031 | 0.017 |
| | | 20 | 1 | 0 | 1 | 0.030 | .979 | 0 | 1 | 0.025 | .380 | 0 | .930 | 0.019 |
| | | 40 | 1 | .095 | 1 | 0.043 | 1 | 0 | 1 | 0.034 | .990 | 0 | 1 | 0.023 |
| | | 60 | 1 | .989 | 1 | 0.054 | 1 | .010 | 1 | 0.042 | 1 | 0 | 1 | 0.025 |
| | | 80 | 1 | 1 | 1 | 0.070 | 1 | 1 | 1 | 0.062 | 1 | 0 | 1 | 0.033 |
| | | 100 | 1 | 1 | 1 | 0.085 | 1 | 1 | 1 | 0.083 | 1 | .890 | 1 | 0.041 |
| High | 500 | 0 | 0 | 0 | .061 | 0.034 | 0 | 0 | .061 | 0.034 | 0 | 0 | .061 | 0.034 |
| | | 20 | .938 | 0 | .948 | 0.059 | .825 | 0 | .990 | 0.048 | .408 | 0 | .888 | 0.040 |
| | | 40 | 1 | .670 | .990 | 0.082 | 1 | 0 | 1 | 0.064 | .990 | 0 | 1 | 0.046 |
| | | 60 | 1 | 1 | 1 | 0.105 | 1 | .162 | 1 | 0.078 | 1 | 0 | 1 | 0.053 |
| | | 80 | 1 | 1 | 1 | 0.134 | 1 | 1 | 1 | 0.106 | 1 | .609 | 1 | 0.069 |
| | | 100 | 1 | 1 | 1 | 0.161 | 1 | 1 | 1 | 0.130 | 1 | 1 | 1 | 0.085 |
| | 2000 | 0 | 0 | 0 | .031 | 0.017 | 0 | 0 | .031 | 0.017 | 0 | 0 | .031 | 0.017 |
| | | 20 | .990 | .960 | .990 | 0.047 | 1 | .152 | 1 | 0.034 | 1 | 0 | 1 | 0.025 |
| | | 40 | 1 | 1 | 1 | 0.072 | 1 | 1 | 1 | 0.052 | 1 | .899 | 1 | 0.032 |
| | | 60 | 1 | 1 | 1 | 0.097 | 1 | 1 | 1 | 0.068 | 1 | 1 | 1 | 0.040 |
| | | 80 | 1 | 1 | 1 | 0.129 | 1 | 1 | 1 | 0.101 | 1 | 1 | 1 | 0.057 |
| | | 100 | 1 | 1 | 1 | 0.161 | 1 | 1 | 1 | 0.129 | 1 | 1 | 1 | 0.075 |

IPD = item parameter drift; AIC = Akaike information criterion; BIC = Bayesian information criterion; LRT = likelihood ratio test with $\alpha = .05$; UA = unsigned area effect size; All = intercepts, main effects and interactions; values for AIC, BIC, and LRT are proportions of replications that criterion or test chose the freely estimated TDCM over the constrained TDCM; values for UA are averaged across replications.

attained, the UA ranged from .035 to .087 in the low IPD magnitude conditions and ranged from .040 to .161 in the high IPD magnitude conditions. While more research and empirical studies are needed to provide recommendations, based on these results, it appears that UA values approaching .09 or above may indicate a large amount of IPD.

### *4.3. Simulation Study Part II Design*

Part II of the simulation study was designed to examine the efficacy of the freely estimated TDCM in the case that pre- and post-test have different items. In this case, it is not expected that item parameter will have the same values, so imposing equality constraints is inappropriate. The item invariance property of the LCDM states that the freely estimated TDCM should be able to provide accurate classification in this case. In terms of design, we used the same, balanced, 4-attribute, Q-matrix as Part I of the simulation study. To simulate different tests, we created three different tests: an easy test, a moderate test, and a hard test. These different tests were distinguished by their correct response probabilities. We defined easy items as items where non-masters have a .30 probability of responding correctly, while masters have a .80 probability of responding

TABLE 7.
Simulation study part II: freely estimated TDCM classification accuracy and reliability for different tests.

| Item difficulty | | Accuracy | | Reliability | | Estimated growth in attribute mastery | | | |
|---|---|---|---|---|---|---|---|---|---|
| Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ |
| Easy | Moderate | .922 | .942 | .953 | .971 | .000 | .052 | .095 | .144 |
| Easy | Hard | .921 | .942 | .952 | .971 | −.002 | .051 | .095 | .143 |
| Moderate | Easy | .921 | .940 | .952 | .971 | .008 | .058 | .100 | .152 |
| Moderate | Hard | .921 | .941 | .951 | .971 | .002 | .054 | .102 | .148 |
| Hard | Easy | .921 | .940 | .950 | .971 | .006 | .059 | .106 | .153 |
| Hard | Moderate | .921 | .941 | .950 | .971 | .007 | .057 | .105 | .153 |

Easy, moderate, and hard tests were distinguished by differing correct response probabilities. Classification accuracy and reliability were averaged over the four attributes. True growth in mastery proportion was 0, .05, .10, and .15 for the four attributes, respectively.

correctly. We defined moderate difficulty items as items where non-masters have a .20 probability of responding correctly, while masters have a .69 probability of responding correctly. We defined hard items as items where non-masters have a .10 probability of responding correctly, and masters have a .52 probability of responding correctly. Though the item discriminations were different for each test (.60 for easy, .59 for moderate, and .42 for hard), these item response probabilities were designed to ensure that item information, as defined by the cognitive diagnostic index (Henson & Douglas, 2005), was held constant across the three tests. The three tests were crossed at pre- and post-test to simulate each possible pattern of different items at pre- and post-test (e.g., easy-moderate, easy-hard, moderate-easy, etc.).

A sample size of 1000 was used as an intermediate sample size to demonstrate that large sample sizes are not necessary to observe the flexibility of the TDCM. Similar to Part I, pre-test attribute mastery base-rates were .40 for each attribute, and transition probabilities were created to produce varying amounts of growth in attribute mastery of 0, .05, .10, and .15 for the four attributes, respectively.

### *4.4. Simulation Study Part II Results*

Table 7 displays the classification accuracy and reliability of the freely estimated TDCM in the six pre- and post-test combinations, as well as the estimated growth in attribute mastery. Across the six different test combinations, the classifications were equally accurate and reliable. The post-test classifications were slightly more accurate and reliable because the pre-test classifications provide extra information upon which to estimate post-test classifications. We also observed that regardless of the difficulty of pre- and post-tests, the estimated growth in attribute mastery was very close to the true growth in attribute mastery. These results suggest that the freely estimated TDCM can be used in cases where different tests are administered at each time point, and common items are not necessary. While these results suggest that it is possible to use different tests with no common items, these results are purely theoretical and demonstrate the LCDM item invariance property in the context of longitudinal designs. We recommend using a couple common items per attribute to provide a mechanism by which to investigate measurement invariance.

## 5. Empirical Analysis with the TDCM

To demonstrate the utility of the TDCM, we analyzed pre-test/post-test data collected in a large scale mathematics education study (Bottge, Ma, Gassaway, Toland, Butler, & Cho, 2014;

Bottge, Toland, Gassaway, Butler, Choo, Griffen, & Ma, 2015). The study included a total of 879 students in middle school. The sample was mostly male (54%), mostly white (78%), and most students were in 7th grade (64%); 15% and 21% were in 6th and 8th grade, respectively. The overall goal of these studies was to examine the difficulties associated with and improve mathematics problem-solving skills for students with disabilities. In particular, these studies examine an innovative instructional program called Enhanced Anchored Instruction (EAI; Bottge, Heinrichs, Chan, Mehta, & Watson, 2003). In EAI, students engage in authentic problem-solving sessions where students watch a 10 to 15-min video presenting a group of adolescents encountering and attempting to solve a problem. Students then search the video for relevant information and use their mathematics knowledge to help the characters in the video develop a solution. Research has shown that embedding mathematics problems in videos can help students with disabilities unlock the meaning of text-based problems (Bottge, Heinrichs, Chan, & Serlin, 2001). It is not known, however, whether engaging in these types of problem-solving activities over a period of instructional time actually improves the problems-solving skills of students with disabilities.

This study provides an analysis of pre-test/post-test data from a researcher-developed mathematics problem-solving test administered to students at the beginning and end of an 18-week instructional period. This problem-solving test had 21 simple structure items measuring four attributes: ratios and proportional relationships, measurement and data, number system (fractions), and geometry (graphing). The items were designed to assess students' problem-solving skills and understanding of four instructional units corresponding to each attribute. Items went through iterations of revisions based on previous research and feedback from mathematics educators and assessment specialists. The four attributes were measured by four, six, five, and six items, respectively. Each item was open-ended and dichotomously scored (correct/incorrect). As each item was simple structured, the TDCM employs the full LCDM, which included intercepts and main effects. Though this test is simple structured, and does not make full use of the LDCM measurement model flexibility, this analysis does demonstrate the utility of the TDCM and also demonstrates the robustness of the model to departures from full measurement invariance. The goal of this analysis is to analyze the growth of the examinees with respect to attribute mastery. We provide item parameter estimates, pre- and post-test attribute reliabilities, pre- and post-test mastery proportions and tests of growth, and individual attribute transition probability matrices.

### *5.1. Assessment of Measurement Invariance and Model Fit*

To assess measurement invariance across the pre- and post-test occasions, we compared the fit of the freely estimated TDCM to the parameter-constrained TDCM. Both models were specified with a two-way structural model at both time points; higher-order structural models did not converge. A likelihood ratio test of these models indicates that the freely estimated TDCM fits the data better than the constrained TDCM ($\chi^2\,(42) = 372.70$, $p < .001$). Additionally, the information criteria preferred the freely estimated TDCM (AIC = 37902.99, BIC = 39475.22) over the parameter-constrained TDCM (AIC = 38196.39, BIC = 39567.9). Together, these results indicate that the data do not exhibit full measurement invariance.

In the simulation study, however, we observed that under similar conditions, the parameter-constrained TDCM was able to provide accurate and reliable classification when full measurement invariance was violated. To evaluate the degree to which full measurement invariance was violated, we first examined the pre- and post-test item response probabilities. Figure 1 displays the item response probabilities for the freely estimated TDCM, comparing pre- and post-test item response probabilities for masters and non-masters. From this figure, it is apparent that pre- and post-test item response probabilities were similar, with an average deviation of .06 on the probability scale, and six deviations larger than .15. We then compared the classifications of the freely estimated TDCM and parameter-constrained TDCM and found that the two models agreed 97.9% of the

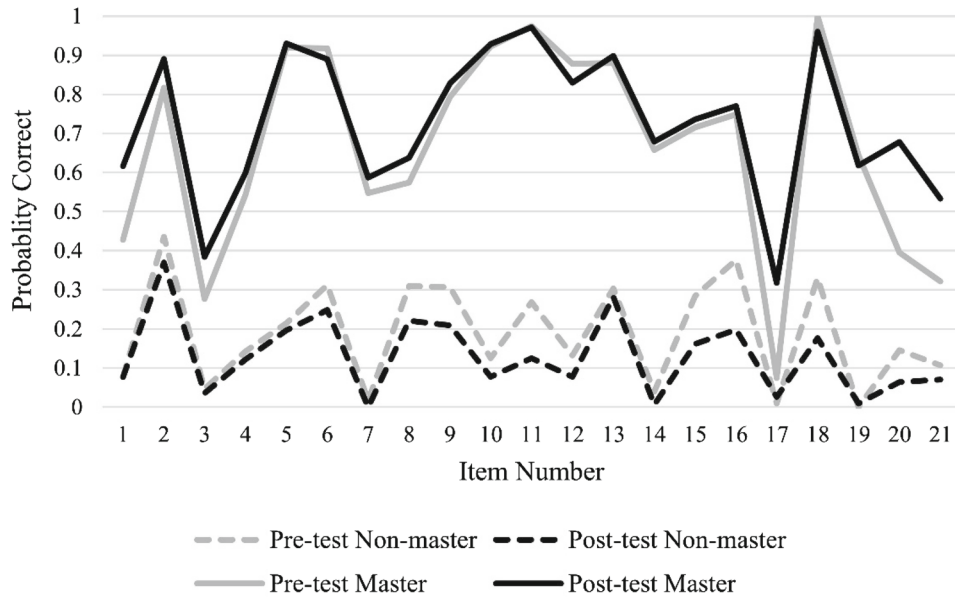


FIGURE 1.
Item response probabilities for the freely estimated TDCM.

time. Additionally, the effect size measure UA was .051, which, according to the simulation study, does not indicate a problematic amount of IPD. Combined with the simulation study results, these findings suggest that IPD is not substantial. The results of the simulation study suggest that under these conditions, the parameter-constrained TDCM provides accurate and reliable classifications. Although partial invariance is sufficient in some cases, in this analysis, we are testing for growth in attribute mastery, which requires an assumption of full measurement invariance. Therefore, we proceed in the following section with interpretations of the results from the parameter-constrained TDCM which enables the comparison of classifications across pre- and post-test and tests of growth in attribute mastery.

To assess the absolute fit of the model, we employed a Monte Carlo bootstrap procedure (Langeheine, Pannekoek, & van de Pol, 1996). We took this approach because traditional goodness-of-fit hypothesis tests are not accurate nor feasible in a DCM context with extreme sparseness in the contingency table of all possible response patterns. The Monte Carlo procedure simulates many data sets according to the known theoretical distributions of the empirical model estimates. In this case, each simulated item parameter was drawn from a normal distribution, centered at the empirical estimate of the parameter with standard deviation equal to the standard error of the empirical estimate. Attribute mastery statuses were drawn from Bernoulli distributions with probabilities equal to each examinee's empirical attribute mastery posterior probabilities. This approach is suggested by Rupp et al. (2010) and was used for DCMs by Templin and Henson (2006). We simulated 500 data sets, estimated them in Mplus, and obtained empirical bootstrap distributions of three model fit statistics: (1) Pearson correlation for item pairs; (2) Cohen's $\kappa$ for item pairs; and (3) raw agreement of observed and predicted item responses. For the Pearson correlation and Cohen's $\kappa$, we computed the median absolute deviation (MAD) by computing the median of the absolute difference between the model-predicted and observed correlation and $\kappa$. In the empirical analysis, we observed a Pearson correlation MAD of .191, Cohen's $\kappa$ MAD of .183, and a raw agreement percentage of .797. Using the bootstrap distributions as estimates of

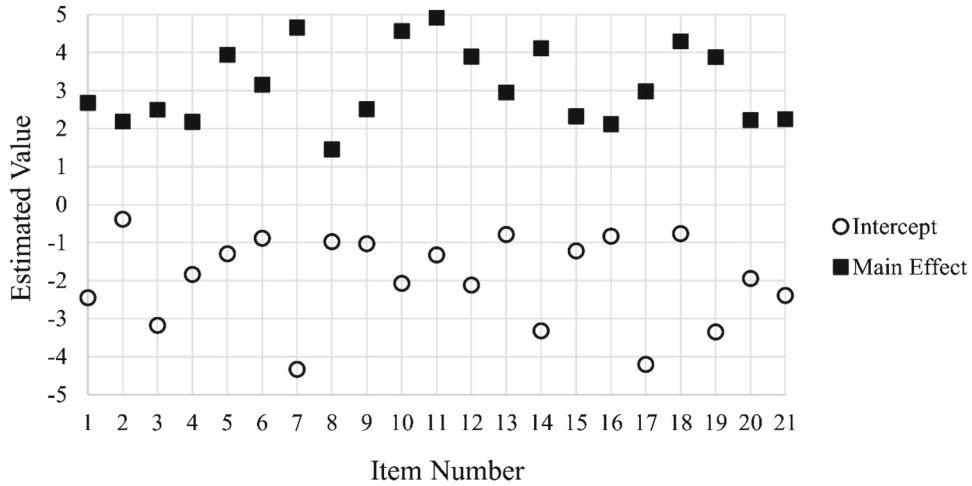


FIGURE 2.
Problem-solving test TDCM item parameters.

the sampling distributions of these statistics, these observed statistics corresponded to $p$ values of .564, .467, and .156, respectively. These $p$ values indicate that the observed fit statistics fall within the range of what we would expect from data generated from a similar model. These model fit results suggest that the model adequately represents the observed data.

### *5.2. Item Parameters*

Figure 2 displays the TDCM item parameters. The median intercept and main effect estimates were − 1.84 and 2.95, respectively. These median parameter estimates correspond to correct response probabilities of .14 and .75 for non-masters and masters, respectively. Item quality can be measured by how well they discriminate between non-masters and masters of the required attribute. Overall, the items discriminated well, with item discriminations ranging from .21 to .81, and a median discrimination of .55.

### *5.3. Classification Reliability*

Classification reliability was calculated according the tetrachoric correlation defined reliability metric defined by Templin and Bradshaw (2013). The pre-test classifications were highly reliable, with reliabilities of .950, .987, .968, .955 for ratios and proportional relationships, measurement and data, number system (fractions), and geometry (graphing), respectively. Classifications were also highly reliable for the post-test, with reliabilities of .969, .994, .971, and .981 for the four attributes, respectively. The simple structure of the items and quality of the items in terms of discrimination contributed to the high reliability estimates. Similar to the simulation study, we observed that the post-test classifications were more reliable than the pre-test classifications. With each respective attribute being measured by between 4 and 6 items, these high reliability estimates highlight the utility and efficiency of using DCMs when classification is the goal of the assessment.

### *5.4. Attribute Mastery Classifications and Growth*

Figure 3 shows the pre-test and post-test mastery proportions for each attribute. Pre-test attribute mastery proportions ranged from .38 to .62, with an average of .47. The most mastered

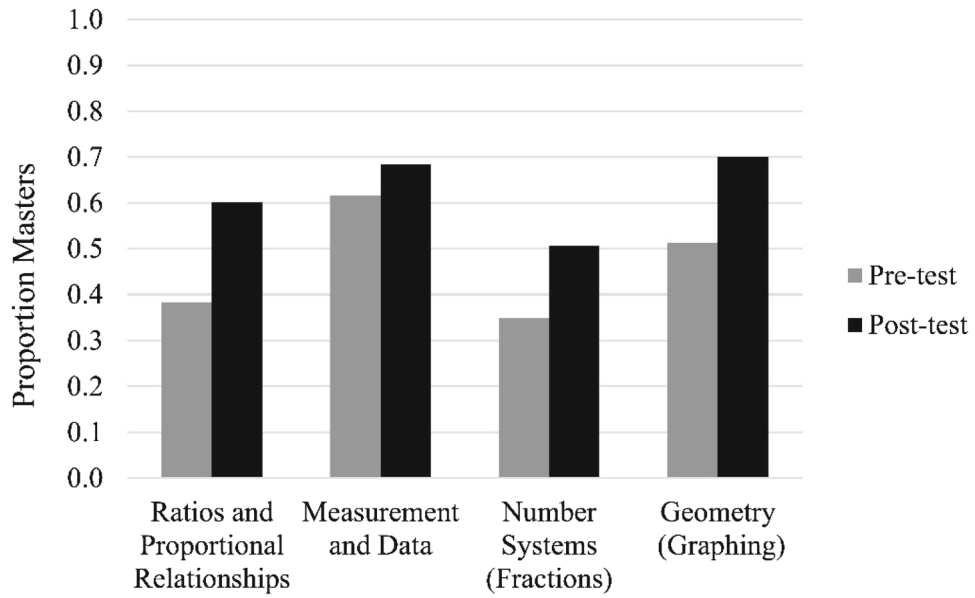


FIGURE 3.
Pre-test (gray) and post-test (black) marginal attribute mastery proportions.

attribute at pre-test was measurement and data, while the least mastered was number systems (fractions). Post-test attribute mastery proportions ranged from .51 to .70, with an average of .62. The most mastered attributed at post-test was geometry (graphing), while the least mastered at post-test was number systems (fractions).

Attribute mastery growth ranged from .068 to .219. Students exhibited the most growth for ratios and proportional relationships (.219) and the least with measurement and data (.068) Measurement data were in one sense the easiest attribute with the highest pre-test mastery proportion at .616, but had the least growth from pre-test to post-test (.068) To assess the statistical significance of mastery growth for each attribute, we employed a Wald test in the TDCM with Hochberg's multiple test correction (1988). Table 8 summarizes the mastery growth and Wald test results for each individual attribute. For each attribute, the examinees exhibited significant growth in mastery from pre-test to post-test for all attributes as indicated by all attribute growth tests resulting in $p$ values less than .001. To facilitate an alternative description of the observed growth in attribute mastery, we included the odds ratio as a growth in attribute mastery effect size.

### *5.5. Attribute Mastery Transitions*

While the previous section gives information on overall attribute mastery growth, it does not give details on the individual attribute mastery transitions. One benefit of using the TDCM is that the transition probabilities are estimated and accounted for in the estimation of posterior probabilities of mastery. In this case, there were $4^2 \times 4^2 = 256$ transition probabilities. Using rules of conditional probability, mastery transition probability matrices were calculated for each individual attribute. These matrices are shown in Table 9. The second cell (top right) in each $2\times2$ matrix represents the conditional probability of transitioning from non-mastery at pre-test to mastery at post-test. For the first three attributes (ratios and proportional relationships, measurement and data, number systems (fractions)), this cell ranges from .41 to .48, indicating that pre-test non-masters of these attributes were more likely to remain non-masters at post-test than

TABLE 8.
Attribute mastery growth from pre-test to post-test.

| Attribute | Pre-test mastery | Post-test mastery | Growth | $p$ value | Odds ratio |
|---|---|---|---|---|---|
| Ratios and proportional relationships | .384 | .602 | .219 (0.026) | $< .001$ | 2.43 |
| Measurement and data | .616 | .684 | .068 (0.020) | $< .001$ | 1.35 |
| Number systems (fractions) | .348 | .507 | .159 (0.023) | $< .001$ | 1.92 |
| Geometry (graphing) | .512 | .701 | .189 (0.027) | $< .001$ | 2.23 |

Wald test $p$ values are adjusted by Hochberg's (1988) stepwise multiple tests procedure. Odds ratio = odds(mastery at post-test)/odds(mastery at pre-test).

TABLE 9.
Attribute mastery transition probability matrices.

| Attribute | | | Mastery transition matrix | |
|---|---|---|---|---|
| Ratios and proportional relationships | | | Post-test | |
| | | | 0 | 1 |
| | Pre-test | 0 | .52 | .48 |
| | | 1 | .21 | .79 |
| Measurement and data | | | Post-test | |
| | | | 0 | 1 |
| | Pre-test | 0 | .57 | .43 |
| | | 1 | .16 | .84 |
| Number systems (fractions) | | | Post-test | |
| | | | 0 | 1 |
| | Pre-test | 0 | .59 | .41 |
| | | 1 | .31 | .69 |
| Geometry (graphing) | | | Post-test | |
| | | | 0 | 1 |
| | Pre-test | 0 | .42 | .58 |
| | | 1 | .19 | .81 |

to transition into mastery. This is not the case for geometry (graphing), where the probability of transitioning to mastery at post-test, given non-mastery at pre-test, is .58.

As pointed out earlier, measurement and data were the most mastered attribute at pre-test, but here we see it was also one of the hardest attributes to master for examinees who were not masters at pre-test with a non-mastery to mastery transition probability of .43. Number systems (fractions) were consistently difficult to master; it had the fewest masters at pre-test (.35) and was the hardest attribute to master given pre-test non-mastery (.41). Also of note is that pre-test masters of number systems (fractions) were more likely to regress into non-mastery (.31) than the other three attributes (.21, .16, .19).

### *5.6. Teacher and Student Feedback*

To further demonstrate the utility of the TDCM beyond a traditional gain score analysis, Fig. 4 presents the pre- and post-test attribute mastery probabilities for two students with identical gain scores. Student A and Student B both had pre-test scores of 9/21 and post-test scores of 12/21. Using a traditional gain score analysis, these students would have identical gain scores of + 3. Similarly, using a longitudinal Rasch IRT model, these students would have identical

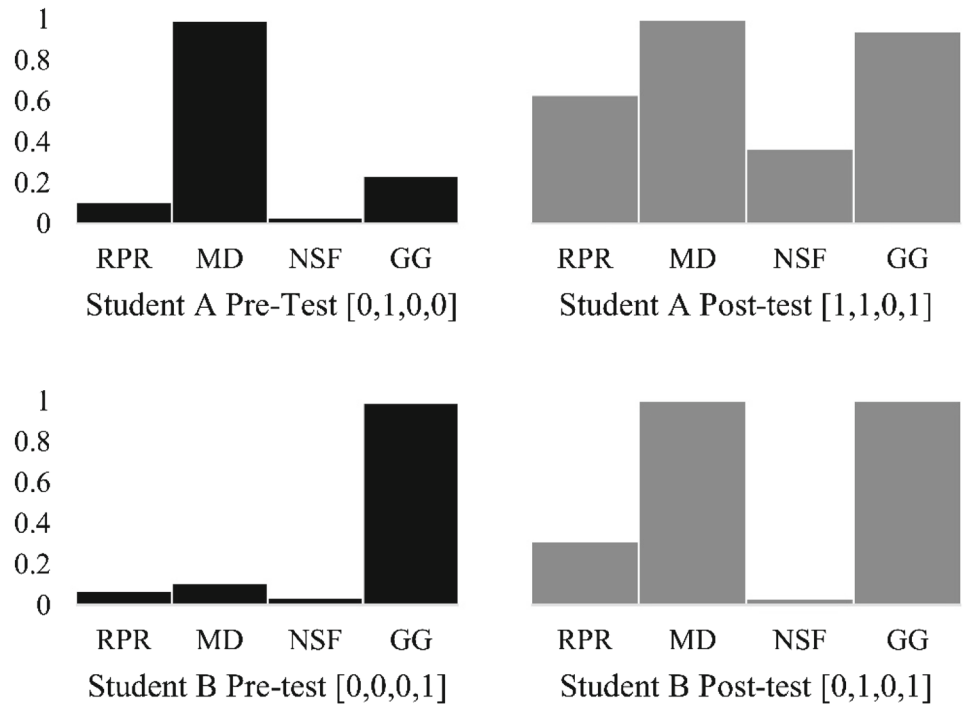


FIGURE 4.
Student changes in attribute mastery probabilities (pre-test is left, post-test is right). Student A (top) and Student B (bottom) both went from a raw pre-test score of 9 to a raw post-test score of 12 (gain = 3). *RPR* ratios and proportional relationships, *MD* measurement and data, *NSF* number systems (fractions), *GG* geometry (graphing).

ability estimates at pre- and post-test, and hence identical ability growth estimates. In the TDCM, however, we see that their transitions in attribute mastery are very different. Student A went from [0, 1, 0, 0] at pre-test to [1, 1, 0, 1] at post-test, while Student B went from [0, 0, 0, 1] to [0, 1, 0, 1]. As a result of this feedback, a teacher could give these students different instruction as it relates to the measured attributes. In addition, a teacher could use the aggregate results from Tables 8 and 9 to evaluate instruction. Results in Tables 8 and 9 indicate that more focus could be directed toward measurement and data and number systems (fractions), which were the hardest attributes to master for students who were non-masters at pre-test.

## 6. Conclusions

This study presents methodology for analyzing longitudinal data in a general DCM framework. In this study, we focused on the pre-test/post-test design where students are administered a set of items before and after instruction. In the pre-test/post-test designed experiment, the goal is to evaluate student learning over time, commonly referred to as growth. In a DCM framework, this equates to analyzing change in attribute mastery. Until now, the only way to model growth in a DCM framework was to either employ the longitudinal DINA or DINO model, or separately estimate the LCDM at each testing occasion. This study generalized and extended previous work by developing the Transition Diagnostic Classification Model (TDCM), which combines latent transition analysis and a general DCM, the LCDM, to model data from longitudinal diagnostic

tests and statistically test for attribute mastery growth. In this way, the TDCM provides a general psychometric method for providing categorical and criterion-referenced interpretations of growth.

We conducted a simulation study to assess the robustness of the TDCM under different IPD conditions. Results from the simulation study indicated the TDCM provided accurate and reliable classifications. Even when IPD was high in magnitude and all items exhibited IPD, the performance of the TDCM did not suffer. With increasing IPD, the TDCM did struggle to accurately recover item parameter estimates. These results suggest that TDCM is quite robust to departures from full measurement invariance in terms of classification accuracy and reliability, and less robust with respect to item parameter estimation. Following the suggestions of IRT researchers to focus on the practical significance of misfit (Maydeu-Olivares, 2015; Sinharay & Haberman, 2014), we observed here that in the presence of IPD, the TDCM still serves its intended purpose of classifying examinees accurately and reliably. Part II of the simulation study showed that when the examinees are administered a different set of items at pre- and post-test, the freely estimated TDCM provides accurate and reliable classifications. While this result does provide flexibility in the construction and administration of tests over several occasions, it is a theoretical result, demonstrating the LCDM item invariance property in a longitudinal context. In practice, we recommend using a few common items per attribute to provide a mechanism to assess measurement invariance.

In practice, these results have significant implications. First, the robustness of the TDCM is critical as full measurement invariance is rarely obtained, but when using the same test over time, assuming measurement invariance over time is necessary to interpret examinee growth and test for examinee growth across testing occasions. Part I of the simulation study indicated that under the conditions simulated, the TDCM can obtain valid classifications when full measurement invariance is not observed. Secondly, Part II of the simulation study suggests that TDCM affords tremendous flexibility in terms of assessment design because the pre- and post-test do not have to have any common items. We note that these conclusions are based on the simulation study conditions, and while they offer insights into the impact of DCM IPD, these conditions do not reflect all possible IPD conditions nor all possible assessment contexts. We also note that the simulation conclusions have adequate model fit, and these results are not expected to hold in the presence of model misfit. And while the TDCM was robust in the presence of IPD, this does not negate the importance of testing for IPD and closely examining the underlying causes for the observed IPD. If the meaning of the attributes has changed, a condition that cannot be simulated, the results of a TDCM growth analysis are meaningless.

The utility and robustness of the TDCM was demonstrated in an analysis of pre-test/post-test data from a mathematics problem-solving test. The TDCM provided highly reliable classifications, even though attributes were only measured by 4–6 items. Statistical tests of attribute mastery growth indicated that students improved from pre-test to post-test with respect to all four attributes. Mastery transition matrices for each attribute gave information detailing how students transitioned between mastery and non-mastery, and gave information regarding which attributes were the most difficult to learn. The TDCM analysis gives considerably more actionable and reliable information than could be obtained from a traditional gain score analysis or longitudinal IRT analysis.

The TDCM modeling framework is not without limitations. The first is that in cases where the pre- and post-test have different items, and item parameter invariance is not expected, the TDCM must rely on the theoretical LCDM classification invariance property (Bradshaw & Madison, 2016). We demonstrated that the freely estimated TDCM can provide accurate and reliable classifications in this case of different items. But if the DCM does not fit the data, the invariance property does not hold, and this could result in a threat to the validity of the classifications and interpretations. Another limitation is model complexity and estimation time. In this study, we simulated a 4-attribute, 20-item test. If there are more attributes, and/or more time points, the number of estimated parameters increases exponentially. Data requirements and estimation time

for additional attributes and times points are unknown, but are likely to be more demanding, and perhaps not feasible. In the future, we hope to examine these extensions of the TDCM and possibly examine model assumptions that would aid in convergence and reducing estimation complexity. One such example could be assuming attribute independence or attribute mastery growth independence (see Li et al., 2015), which would significantly reduce the number of latent classes at each time point. Another option could be to fit a reduced measurement model. In cases with highly complex items measuring many attributes, where the full LCDM may not be an option, previous research suggests that specifying the C-RUM (Hartz, 2002; Henson et al., 2009) as the measurement model in the TDCM is an option that may provide similar classifications (Kunina-Habenicht et al., 2012). This measurement model specification highlights the necessity of the generality of the TDCM. That is, longitudinal variants of many constrained DCMs can be specified within the TDCM framework. This is not the case with previous longitudinal DCMs where researchers were restricted to using constrained diagnostic classification measurement models.

In sum, the TDCM is a general model that can be used to model and test for attribute mastery growth in pre-test/post-test designed assessment studies. The TDCM provides a modeling framework that is appropriate for modeling data from the commonly used pre-test/post-test assessment design. Now that this methodology is available and estimable in commercially available software (Mplus), and we are hopeful that researchers are encouraged to design diagnostic tests for assessing growth over time in the context of pre-test/post-test designs, and use the fine-grained feedback to improve instruction, thereby improving student learning.

## Acknowledgments

This work was supported by the Institute of Educational Sciences (IES) Grant Number R324A150035. The opinions expressed are those of the authors and do not necessarily reflect the views of IES.

## Appendix

Abbreviated TDCM Mplus Syntax (two-attribute, pre-test/post-test example).
This example combines LCDM syntax from Templin and Hoffman (2013) and Mplus's LTA capabilities to estimate the TDCM.

```
-------------------------------------------------------------------------

TITLE: ! Section that appears in header of output file
     Pre-Post TDCM;
     Item parameter invariance assumed;
     Two attributes and 10 items;

DATA:  ! Location of free format data file
     FILE = prepost1.txt;

VARIABLE:
     NAMES = ID pre1-pre10 post1-post10;      ! List of variables in
data set
     USEVARIABLE = pre1-pre10 post1-post10;   ! Variables to be
     analyzed: 10 pre/post items
     CATEGORICAL = pre1-pre10 post1-post10;   ! Each item is
     dichotomous
     CLASSES = c1(4) c2(4);                   ! Four attribute
profiles at pre and post
     IDvariable = ID;                         ! Person ID variable to
save examinee data

ANALYSIS:
     TYPE = MIXTURE;            ! Estimates latent classes
     PROCESSORS = 8;            ! Number of processors available

MODEL:

%OVERALL%
[c1#1];  ! Latent variable mean for attribute profile [0,0], pre-test

[c1#2];  ! Latent variable mean for attribute profile [0,1], pre-test

[c1#3];  ! Latent variable mean for attribute profile [1,0], pre-test


[c2#1];  ! Latent variable mean for attribute profile [0,0], post-test

[c2#2];  ! Latent variable mean for attribute profile [0,1], post-test

[c2#3];  ! Latent variable mean for attribute profile [1,0], post-test

 c2 on c1;  ! Regress post-test classifications on pre-test
 classification, specifies the LTA model
 Model c1:   !Item parameters for pre-test

 %c1#1%  ! Model for Attribute Profile [0,0]
     [pre1$1] (T1_1);       ! Item 1 Thresh 1
     !----{code for other items omitted}----!
     [pre10$1] (T10_1);     ! Item 10 Thresh 1


 %c1#2%  ! Model for Attribute Profile [0,1]
     [pre1$1] (T1_1);       ! Item 1 Thresh 1
     !----{code for other items omitted}----!
     [pre10$1] (T10_2);     ! Item 10 Thresh 2
```

```
!----{code for other classes omitted}----!

Model c2: ! Item parameters for post-test (same as pre-test)
          ! Measurement invariance specified by keeping threshold
value names the same
          ! Can test invariance specifying different threshold value
names for post-test

%c2#1%  ! Model for Attribute Profile [0,0]
    [post1$1] (T1_1);       ! Item 1 Thresh 1
    !----{code for other items omitted}----!
    [post10$1] (T10_1);     ! Item 10 Thresh 1


%c2#2%  ! Model for Attribute Profile [0,1]
    [post1$1] (T1_1);       ! Item 1 Thresh 1
    !----{code for other items omitted}----!
    [post20$1] (T10_2);     ! Item 10 Thresh 2

!----{code for other classes omitted}----!



MODEL CONSTRAINT:
     ! Used to define LCDM parameters (see Templin & Hoffman, 2013)
     ! Mplus uses P(X=0) rather than P(X=1) so multiply by -1
     ! If invariance not assumed, specify LDCM items for each testing
occasion

! Item 1: Define LCDM parameters for item 1
NEW(L1_0 L1_11);
T1_1=-(L1_0);                          ! Item 1 Thresh 1
T1_2=-(L1_0+L1_11);                    ! Item 1 Thresh 2
! Main effect order constraints
L1_11>0;

!----{code for other items omitted}----!

! Item 10: Define LCDM parameters for item 10
NEW(L10_0 L10_12);
T10_1=-(L10_0);                         ! Item 10 Thresh 1
T10_2=-(L10_0+L10_12);                  ! Item 10 Thresh 2
! Main effect order constraints
L10_12>0;

! Wald test of growth is omitted, but would be in Model Constraint
! Available upon request, see first author website [to be included]

SAVEDATA: ! Format, name of posterior probabilities of class
membership file
    FORMAT = F10.5;
    FILE = prepost1_out.txt;
    SAVE = CPROBABILITIES;
```

```
!------------------------------------------------------------
! OUTPUT
!------------------------------------------------------------

FINAL CLASS COUNTS AND PROPORTIONS FOR EACH LATENT CLASS VARIABLE
BASED ON THE ESTIMATED MODEL

  Latent Class
    Variable    Class

    C1              1       428.70493          0.42870
                    2       243.09047          0.24309
                    3       183.16370          0.18316
                    4       145.04092          0.14504
    C2              1       119.37941          0.11938
                    2       183.41757          0.18342
                    3       170.37970          0.17038
                    4       526.82330          0.52682

LATENT TRANSITION PROBABILITIES BASED ON THE ESTIMATED MODEL

  C1 Classes (Rows) by C2 Classes (Columns)

             1        2        3        4

   1     0.171    0.226    0.369    0.234
   2     0.120    0.250    0.000    0.630
   3     0.093    0.141    0.050    0.716
   4     0.000    0.000    0.020    0.980

!------------------------------------------------------------

New/Additional Parameters
                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

    L1_0              -1.306      0.144     -9.055      0.000
    L1_11              2.341      0.316      7.401      0.000

   !----{other items' estimates omitted}----!

    L10_0             -1.318      0.176     -7.487      0.000
    L10_12             2.286      0.274      8.348      0.000
```

## References


Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*(6), 1554–1563.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275–285.

Bottge, B. A., Heinrichs, M., Chan, S., & Serlin, R. (2001). Anchoring adolescents' understanding of math concepts in rich problem solving environments. *Remedial and Special Education*, *22*, 299–314.
Bottge, B. A., Heinrichs, M., Chan, S. Y., Mehta, Z. D., & Watson, E. (2003). Effects of video-based and applied problems on the procedural Math skills of average- and low achieving adolescents. *Journal of Special Education Technology*, *18*(2), 5–22.
Bottge, B. A., Ma, X., Gassaway, L., Toland, M., Butler, M., & Cho, S. J. (2014). Effects of blended instructional models on math performance. *Exceptional Children*, *80*, 237–255.
Bottge, B. A., Toland, M., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., et al. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Exceptional Children*, *81*(2), 158–175.
Bradshaw, L. (2016). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 297–327). West Sussex: Wiley-Blackwell.
Bradshaw, L., & Madison, M. J. (2016). Invariance properties for general diagnostic classification models. *International Journal of Testing*, *16*(2), 99–118. https://doi.org/10.1080/15305058.2015.1107076.
Bradshaw, L., & Templin, J. (2014). The little model that couldn't: How the DINA model misclassifies students and hides important effects. Paper presented at the annual meeting of the Northeastern Educational Research Association in Trumbull, CT.
Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.
Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*, 131–157.
Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, *4*, 253–278.
de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.
Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.
Fischer, G. H. (1976). Some probabilistic models for measuring change. In D. N. M. de Gruijter & L. J. T. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 97–110). New York: Wiley.
Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*, 599–624.
George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, *56*(4), 405–432.
Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *33*, 315–332.
Han, K. T., & Guo, F. (2011). Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing. GMAC Research Reports, RR-11-02.
Hartz, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.
Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnostics. *Applied Psychological Measurement*, *29*, 262–277.
Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, *74*, 191–210.
Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*(4), 800–802.
Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: theory and applications* (pp. 19–60). London: Cambridge University Press.
Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL (unpublished doctoral dissertation)*. Champaign, IL: University of Illinois.
Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.
Jurich, D. P., & Bradshaw, L. P. (2014). An illustration of diagnostic classification modeling in student learning outcomes assessment. *International Journal of Testing*, *14*, 49–72.
Kaya, Y., & Leite, W. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance. *Educational and Psychological Measurement*, *77*(3), 369–388.
Kunina-Habenicht, O., Rupp, A. A., & Wilhem, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.
Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, *24*, 492–516.
Lanza, S. T., & Collins, L. M. (2008). A new SAS procedure for latent transition analysis: Transitions in dating and sexual behavior. *Developmental Psychology*, *42*(2), 446–456.
Lanza, S. T., Patrick, M. E., & Maggs, J. L. (2010). Latent transition analysis: Benefits of a latent variable approach to modeling transitions in substance use. *Journal of Drug Issues*, *40*, 93–120.

Lao, H., & Templin, J. (2016). Estimation of diagnostic classification models without constraints: Issues with class label switching. Paper presented at the annual meeting of the National Council on measurement in education in Washington, DC.

Lee, W., & Cho, S. J. (2017). The consequences of ignoring item parameter drift in longitudinal item response models. *Applied Measurement in Education*, *30*(2), 129–146.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2015). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*(2), 181–204.

Maydeu-Olivares, A. (2015). Evaluating the fit of IRT models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 111–127). New York, NY: Taylor & Francis (Routledge).

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement invariance. *Psychometrika*, *57*(2), 289–311.

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

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502.

Roberts, T. J., & Ward, S. E. (2011). Using latent transition analysis in nursing research to explore change over time. *Nursing Research*, *60*(1), 73–79.

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

Sinharay, S., & Haberman, S. J. (2014). How often is the misfit of item response theory models practically significant? *Educational Measurement: Issues and Practice*, *33*(1), 23–35.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification*, *30*, 251–275.

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

Templin, J., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice*, *32*(3), 37–50.

von Davier, M. (2005). A general diagnostic model applied to language testing data (Research Report No. RR-05–16). Princeton, NJ: Educational Testing Service.

Wang, S., Yang, Y., Culpepper, S., & Douglas, J. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, *47*, 57–87.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26*, 77–87.

*Manuscript Received: 15 SEP 2016*
*Published Online Date: 27 SEP 2018*