ABSTRACT

Title of Dissertation: MODELING MULTIPLE PROBLEM-SOLVING STRATEGIES AND STRATEGY SHIFT IN COGNITIVE DIAGNOSIS FOR GROWTH

Manqian Liao, Doctor of Philosophy, 2020

Dissertation directed by: Associate Professor, Hong Jiao
Measurement, Statistics and Evaluation
Department of Human Development and Quantitative Methodology

Problem-solving strategies, defined as actions people select intentionally to achieve desired objectives, are distinguished from skills that are implemented unintentionally. In education, strategy-oriented instructions that guide students to form problem-solving strategies are found to be more effective for low-achievement students than the skill-oriented instructions designed for enhancing the skill implementation ability. However, conventional cognitive diagnosis models (CDMs) seldom distinguish the concept of skills from strategies. While the existing longitudinal CDMs can model students' dynamic skill mastery status change over time, they did not intend to model the shift in students' problem-solving strategies. Thus, it is hard to use conventional CDMs to identify students who need strategy-oriented instructions or evaluate the effectiveness of the education intervention

programs that aim at training students' problem-solving strategies. This study proposes a longitudinal CDM that takes into account both between-person multiple strategies and within-person strategy shift. The model, separating the strategy choice process from the skill implementation process, is intended to provide diagnostic information on strategy choice as well as skill mastery status. A simulation study is conducted to evaluate the parameter recovery of the proposed model and investigate the consequences of ignoring the presence of multiple strategies or strategy shift. Further, an empirical data analysis is conducted to demonstrate the use of the proposed model to measure strategy shift, growth in the skill implementation ability and skill mastery status.

MODELING MULTIPLE PROBLEM-SOLVING STRATEGIES AND STRATEGY SHIFT IN COGNITIVE DIAGNOSIS FOR GROWTH

by

Manqian Liao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Hong Jiao, Chair
Professor Jeffrey R. Harring
Professor Robert W. Lissitz
Professor Yang Liu
Professor Matthias von Davier
Professor Xin He, Dean's Representative

Dedication

To my beloved parents,

for their endless love, unconditional support and chipper temperament.

Acknowledgements

First, I would like to express my sincere gratitude to the chair of my dissertation committee, Dr. Hong Jiao, who has been my advisor since I joined the EDMS master's program in 2015. I still recall the day when we first met, Dr. Jiao asked whether I planned to pursue a doctoral degree and my answer then was "Maybe not". I cannot emphasize enough the importance of Dr. Jiao's mentorship in cultivating my interest in psychometric research and my determination to set out on the Ph.D. journey, which I now think is one of the most correct decisions I have made in my life. I would like to thank Dr. Jiao for her support and patience in guiding me through all the challenges, including finishing this dissertation, on my way to accomplishing the degree and launching my career. I feel fortunate to have such an excellent mentor who always inspires me with her broad vision, wisdom and open-mindedness.

I am also grateful to the other five extraordinary members on my dissertation committee, Dr. Jeffrey Harring, Dr. Xin He, Dr. Robert Lissitz, Dr. Yang Liu and Dr. Matthias von Davier, not only for their every piece of valuable and constructive advice to this dissertation, but also for the opportunities they have afforded me, and for the skills and qualities that I have learned from them in various phases of my graduate study: Thank Dr. Harring for laying a solid foundation of simulation study and programming for me in my early years in the program, which has made my life in the subsequent years much easier, and thank him for providing me with opportunities to present my dissertation work to a variety of audience. Thank Dr. He for offering the rewarding longitudinal data analysis course, which has equipped me with a

powerful toolbox of longitudinal data analysis that later became one of the most vital components of my dissertation. Thank Dr. Lissitz for always giving me one of the promptest and most insightful feedbacks, for passing on sensible advices whenever I encounter a communication dilemma and for constantly expanding my vocabulary list of words and idiomatic expressions (e.g., I learned the word *chipper*[1] from him). Thank Dr. Liu for suggesting thoughtful and effective solutions to the technical challenges I had in my dissertation, and for sharing the extremely useful data visualization principles and techniques that I would always try to apply to my current and future work. Thank Dr. von Davier for helping me identify the connections that I may have overlooked between my dissertation research and the existing studies, and for having published a great number of inspiring works on cognitive diagnosis modeling, which have been a very crucial theoretical foundation to my dissertation. What I have learned from my committee members has profoundly influenced my ways of thinking and working and will continue being influential to me in the future.

Last but not the least, I would like to express my appreciation to all the faculty members and fellow students in the EDMS family: Thank the faculty members for their wholehearted guidance and assistance; and thank my fellow students and friends for sharing the joys and tears with me as being a graduate student.

Note: This research was supported by the Institute of Education Sciences (IES) grant R324A150035. The opinions expressed are those of the authors and do not necessarily reflect the views of IES.

[1] *adjective.* cheerful and lively.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Statement of Problem

Problem-solving strategies, defined as actions people select intentionally to achieve desired objectives (Alexander et al., 1998), have been distinguished from skills that are applied unintentionally (e.g., Paris et al., 1991). In contrast to strategies, skills are "applied unconsciously for many reasons including expertise, repeated practice, compliance with directions, luck, and naive use" (Paris et al., 1991). Afflerbach et al. (2008) used an example to clarify the difference between strategies and skills in the reading tasks: when a reader intends to understand the meaning of the text by self-questioning "Does that make sense?" , this is a strategy; when the reader comprehends the text "automatically" without deliberate control on awareness, this is a skill. Therefore, one way of determining whether an action is a strategy or a skill is to ask the question: "Is the action under deliberate control or automatic?" A strategy that requires deliberate control could turn into a skill that is implemented automatically with repeated practices (Afflerbach et al., 2008). From the problem-solving perspective, forming a strategy and carrying out the skills required by the strategy can be treated as two steps in solving a problem. For example, in the IDEAL problem-solving model proposed by Bransford and Stein (1993) where the problem-solving process is divided into five steps, "explore possible strategies" and "act" are two of the key steps which correspond to forming the strategies and implementing the skills, respectively.

Instructions have been designed to guide students to develop problem-solving strategies (e.g., Bottge et al., 2003; Jitendra et al., 2002; Mercer & Mercer, 2001;

Paris et al., 1984). For instance, the schema-based instruction guides students through the process of problem solving, including identifying the key part of the problems and forming tentative solutions (Jitendra et al., 2002). This study refers to these instructions as strategy-oriented instructions. In contrast, skill-oriented instructions refer to the instructions that train students to turn the problem-solving process demanding deliberate control into a more automatic process (Afflerbach et al., 2008).

Both strategy-oriented and skill-oriented instructions are beneficial to the students' problem solving performance (Paris et al., 1983; Pressley, 2000), but their effectiveness may differ across students with different achievement levels. Skill-oriented instructions may help the general students to solve problems more efficiently, but they may be less effective to some low-achievement students. For example, while conventional reading programs tend to be skill-oriented, strategy-oriented instructions may be more effective for some struggling readers (Afflerbach et al., 2008). Strategy-oriented instructions have been found to be effective especially at the initial learning stage and for students with low achievement or learning disabilities (e.g., Coughlin & Montague, 2011; Swanson, 2001). Therefore, in educational practice and cognitive diagnosis, it is necessary to identify the type of instructions needed by students. In addition, to assess the effectiveness of the instructions, students' changes in strategy and skill use over time need to be measured.

As skills and strategies are unobservable mental processes, instruments that provide observable indicators paired with latent variable models are needed to measure them. Analyzing the response data to items in an instrument with appropriate

latent variable models makes it possible to draw inferences about unobservable skills and strategies. Given that the concepts of strategies and skills are seldom distinguished in existing latent variable models, the overall goal of this study is to develop a latent variable model that better distinguishes the role of the strategies and skills in the problem-solving process.

Given the complex nature of the problem-solving process, it is indispensable to make assumptions about the nature of the skills and strategies in the latent variable models in order to draw valid inferences from the model parameters. In general, latent variable models designed for understanding skills and strategies explicitly or implicitly make assumptions about the following three questions based on their specific purposes and cognitive theory:

1) How is the problem-solving strategy operationally defined (how to distinguish different strategies)?
2) What is the relationship between the strategies and skills?
3) How to define multiple strategies? The rest of this section reviews the assumptions made in the existing latent variable models and introduces the assumptions to be made in this study about each of the three questions. These assumptions lay a foundation to the proposed model.

How is the problem-solving strategy operationally defined? In the latent variable models for problem-solving strategies, the strategy is usually represented as a discrete variable in a mixture-distribution model (e.g., Mislevy & Verhelst, 1990; Rost, 1990; Yamamoto, 1989). Different strategies are distinguished by their unique cognitive processes and/or by the unique "outcome" they result in, such as different

item functioning. An example drawn from Mislevy (1996) is given below to demonstrate that different strategies can be distinguished by different sets of skills that are intentionally chosen to solve a problem. Table 1 displays a mixed-number subtraction item that can be solved with two approaches. As side notes, this study refers "strategy" as a person property and "approach" as an item property in order to distinguish the person property from the item property; items that can be solved with more than one approach are referred to as "multiple-approach items". Tatsuoka (1987, 1990) found that middle-school students employ two different strategies to solve the mixed-number subtraction problems like the one shown in Table 1. Specifically, using Strategy A, students would convert the mixed numbers into improper fractions and then do the subtraction; using Strategy B, students would separate the mixed numbers into two parts, i.e., a whole number and a fraction, and then do the subtraction separately for each part. Compared to Strategy A, Strategy B is less demanding on the computational skills (Tatsuoka, 1987). In Table 2, Strategy A and Strategy B are distinguished by two matrices where each row represents an item and each column represents a skill. Each entry in the matrix indicates whether a strategy involves a skill to solve an item. For example, to solve the item $2-\frac{1}{3}$, Strategy A involves skills 1, 2 and 5 while Strategy B involves skills 1, 3, 4 and 5. Thus, various problem-solving strategies are operationally defined as various cognitive processes and can be represented as different Q-matrices in cognitive diagnosis models (CDMs) (e.g., de la Torre & Douglas, 2008). Mislevy and Huang (2007) refer to the mixture models designed for such mixed problem-solving strategies as measurement models with narrative structures, the narrative theme of

which is that “Different persons may use different strategies but are presumed to use the same strategy for all items. It is not known which strategy a person is using. Features of tasks that render them difficult are posited for each strategy.”

Table 1
An Example Multiple-Approach Item with Step-by-Step Solutions

Step	Strategy A	Strategy B
1	$2-\frac{1}{3}$ (5) Convert whole number to fraction $=\frac{6}{3}-\frac{1}{3}$	$2-\frac{1}{3}$ (3) Separate whole number from fraction; (4) Borrow one from whole number to fraction; (5) Convert whole number to fraction; $=1\frac{3}{3}-\frac{1}{3}$
2	(1) Basic fraction subtraction $=\frac{5}{3}$	(1) Basic fraction subtraction $=1\frac{2}{3}$
3	(2) Simplify/Reduce $=1\frac{2}{3}$	

Note. The numbers in parentheses correspond to the numbering of the skills in Mislevy (1996).

Table 2
An Example of Strategy Operationally Defined as a Unique Set of Skills

Item	Skill 1	Strategy A				Strategy B			
		Skill 2	Skill 5	Skill 6	Skill 7	Skill 2	Skill 3	Skill 4	Skill 5
$2-\frac{1}{3}$	1	1	1	0	0	0	1	1	1
$3-2\frac{1}{5}$	1	0	1	1	0	1	1	1	1
$3\frac{7}{8}-2$	1	1	1	1	1	0	1	0	0

In contrast, different strategies can be distinguished by different item parameters (e.g., items are deemed more difficult for the subpopulation employing one strategy than another); thus, the strategy is operationally defined by its outcome

in terms of item functioning. Such definition is more likely to be adopted when the measurement model is an item response theory (IRT) model where no parameter directly represents the cognitive process. In sum, the way to distinguish different strategies is dependent on the characteristics of the measurement model, which will be further elaborated in the literature review section.

This study chooses to use the cognitive process (i.e., a set of skills) to operationally define the strategy mainly for two reasons. On one hand, such operational definition of strategy is more aligned with its theoretical definition given that Paris et al. (1983) refer strategies as "skills under consideration". On the other hand, the proposed model is in the CDM framework; the characteristics of CDMs make it more convenient to define the strategy with its cognitive process as opposed to its outcome.

What is the relationship between strategy and skills? Most existing latent variable models designed for measuring strategies do not explicitly distinguish between the concepts of strategies and skills. However, the model specifications imply that the skills and strategies are assumed to be dependent on each other. In the existing CDMs for studying problem-solving strategies, the attribute mastery status is assumed to be dependent on the problem-solving strategies. "Attribute" is a commonly used term for the discrete latent variable in CDMs; it will be used interchangeably with the term "skill" in the subsequent text. In some models, the distributions of the attribute mastery profiles are allowed to vary across strategies (e.g., von Davier, 2007). Alternatively, the choice of strategy can be conditional on the attribute mastery status (e.g., de la Torre & Douglas, 2008; Ma & Guo, 2019). For

instance, individuals who have mastered an attribute could be more likely to choose strategies that involve a specific attribute. However, without explicitly specifying the roles that skills and strategies play in the problem-solving process, it would be hard to verify or assess the modeling assumptions about the relationships between the strategy and attributes. Furthermore, the validity of the strategy and attribute classification results could be questionable.

In order to distinguish the roles of skills and strategies, this study adapts the IDEAL problem-solving model proposed by Bransford and Stein (1993) to model the item responding process. As shown in Figure 1, The IDEAL model consists of five steps required to solve a problem: 1) Identify the problem (I); 2) Define the cause (D); 3) Explore possible strategies (E); 4) Act (A); 5) Look and learn (L). This study, utilizing a simplified version of the IDEAL model (Figure 1), divides the item responding process into two independent stages: strategy choice and skill implementation. The strategy choice stage corresponds to the first three steps of the IDEAL model where respondents intentionally form a strategy by analyzing the item and identifying the skills required to solve the item. The skill implementation stage corresponds to the last two steps in the IDEAL model where respondents utilize the chosen skills to solve the item. In this study, a strategy is said to be successfully applied when all the skills required by the strategy are implemented.

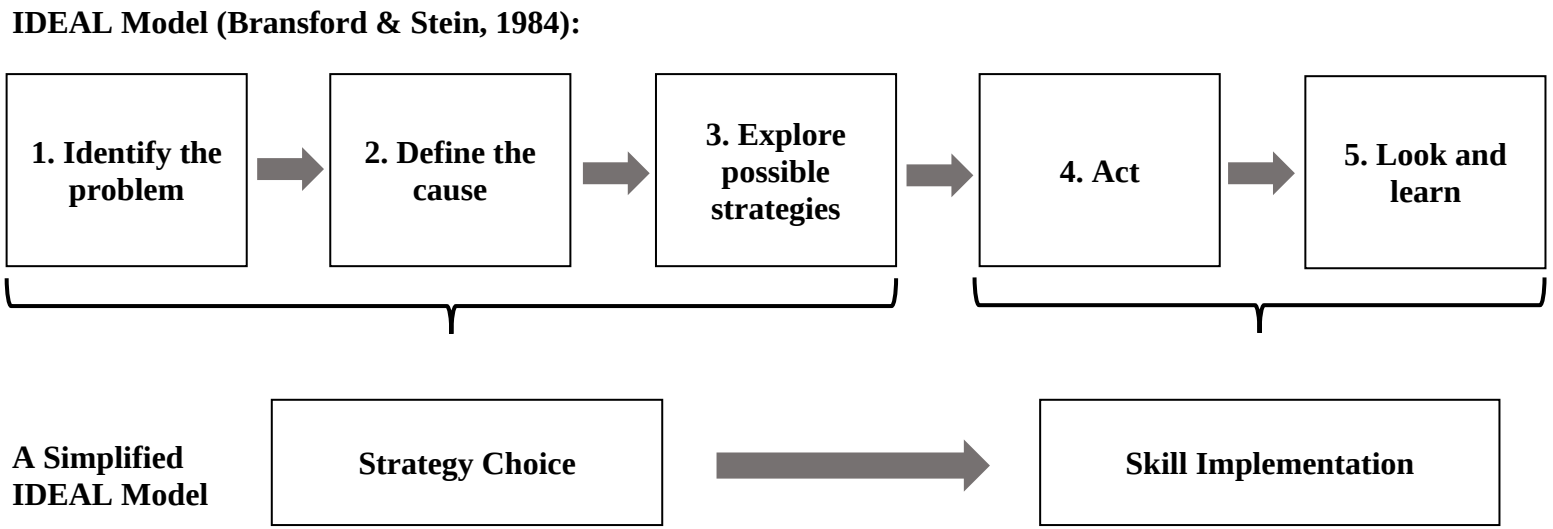


Figure 1. The IDEAL model for problem solving and a simplified IDEAL model.

From a cognitive diagnostic perspective, the separation of the strategy choice process from the skill implementation process is desirable as more targeted instructions could be designed if the diagnostic results can indicate, for example, whether the students have difficulty in choosing a strategy or implementing a skill. As an initial attempt to explicitly separate the roles of the strategy and the skill in a measurement model, this study assumes that the strategy choice and the skill implementation processes are independent for simplicity, given that there is not a consistent theory about the correlational or causal relationships between the strategy choice and skill implementation. While this study does not impose a correlational or causal relationship between the strategy choice and skill implementation, the proposed model will serve as a basis to future studies that have a more compelling theory or hypothesis about the relationship between the strategy choice and skill implementation.

Figure 2 shows the structure underlying an item response. The latent components (i.e., strategy choice, skill implementation and attribute mastery status) that are not directly linked with each other are assumed to be independent. The independence between the strategy choice and skill implementation has two aspects. On one hand, the strategy choice is assumed to be independent from the attribute mastery status that affects the item response probability through the skill implementation stage. Figure 3 elaborates on the independence between the strategy choice (q) and attribute mastery status (α). When solving an item, different combinations of strategy choices and attribute mastery statuses result in 2*2=4 possible mental statuses involving an attribute. In the strategy choice stage, an

individual attempting to use a skill (i.e., identifying the attribute to be required) to solve the item ($q = 1$) may ($\alpha = 1$) or may not ($\alpha = 0$) be able to implement the skill. Similarly, an individual who does not identify an attribute to be required for solving the item may ($\alpha = 1$) or may not ($\alpha = 0$) have mastered the attribute.

On the other hand, it is assumed that the strategy choice and skill implementation stages do not have an interaction effect on the probability of correct item response. While this study makes a strong assumption that the strategy choice and skill implementation processes are completely independent from each other in order to simplify the model structure, such an assumption could be relaxed by allowing the two processes to have some interaction that affects the probability of a correct response. Thus, the proposed model can be easily extended to scenarios where some correlational or causal relationships are hypothesized between strategy choice and skill implementation.

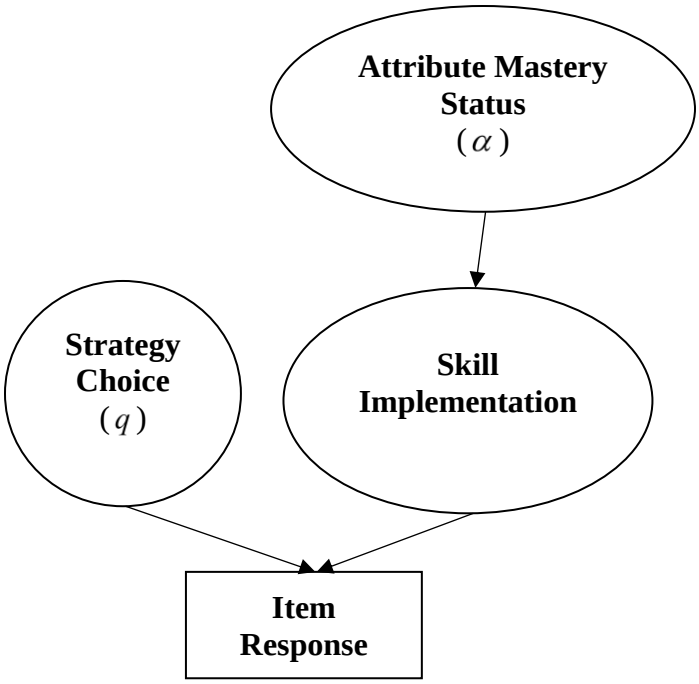


Figure 2. The structure underlying an item response.

Skill implementation (Whether the attribute is mastered?)	Strategy Choice (Whether one attempts to use the attribute to solve the problem?)	
	No ($q = 0$)	Yes ($q = 1$)
No ($\alpha = 0$)	No attempt and non-mastery	Attempted but non-mastery
Yes ($\alpha = 1$)	No attempt but mastery	Attempted and mastery

Figure 3. Cross-classified status of an attribute in the problem-solving process.

How to define multiple strategies? The definitions of multiple strategies can be dependent on the nature of the tasks or items. For example, Mislevy and Verhelst (1990) and Yamamoto (1989) assume that different subpopulations can choose different strategies, but each respondent only uses one strategy throughout a test. In contrast, Rijkes and Kelderman (2007) proposed a strategy-shift model where an examinee may choose to use different strategies for different items within the same test administration. Cho et al. (2010) modeled the shift of strategy over time. In the vocational education setting, Abele and von Davier (2019) have found that the car mechatronics shifted their diagnostic strategies based on the difficulty of the problems. In sum, there is not a universal definition about multiple strategies and different models are designed for different scenarios of multiple strategies.

This study focuses on two types of multiple strategies, the between-person multiple strategies and within-person strategy shift. The between-person multiple strategies refer to the scenario where different respondents on the same test administration choose different strategies to solve an item. It is assumed that each respondent only uses one strategy in one test administration. The within-person strategy shift refers to the scenario where a respondent chooses different strategies to solve the same problems over time; in other words, the respondent shifts his or her strategy over time.

1.2 Purpose

This study proposes a longitudinal CDM that takes into account both between-person multiple strategies and within-person strategy shift. The model, separating the strategy choice process from the skill implementation process, aims at providing

richer diagnostic information compared to the traditional CDMs. In particular, while traditional CDMs diagnose students' skill mastery status, they do not provide information on students' strategy choice especially the shift in their strategy over time. The proposed model, in addition to providing information on whether attributes are mastered as skills, informs that whether attributes are chosen as part of the problem-solving strategy.

A Monte Carlo simulation study is conducted to examine the parameter recovery of the proposed model under several simulated conditions and to investigate the effects of ignoring between-person multiple strategies and within-person strategy shift on the classification accuracy and growth estimates of the longitudinal CDM.

As an empirical data demonstration, the proposed model is applied to the response data from a study with repeated measure pretest-posttest design (Bottge et al., 2015) that assessed the effectiveness of the Enhanced Anchored Instruction (EAI; Bottge, 2001) and compared effectiveness of EAI with that of business as usual (BAU). The empirical data analysis intends to demonstrate the use of the proposed model to provide diagnostic information about strategy choice and skill implementation, respectively.

Specifically, this study aims at addressing the following five research questions, the first three of which are based on the simulation study while the last two of which are based on the empirical data analysis:

1) How do the relative model fit indices perform in identifying the proposed model as the best-fitting model in the presence of both between-person multiple strategies and within-person strategy shift?

2) What is the impact of ignoring between-person multiple strategies and/or within-person strategy shift on the recovery of the model parameters, especially those parameters relevant to diagnostic inferences, of the longitudinal CDMs?
3) How is the parameter recovery of the proposed model affected by the manipulated factors (i.e., the sample size, the initial mixing proportions of strategies, the strategy latent transition probability and the correlation between the initial ability and ability change) in the simulation study?
4) According to the empirical data analysis, how do students' strategy choice, overall skill implementation ability and attribute mastery status change from the pretest to the posttest?
5) According to the empirical data analysis, do Enhanced Anchored Instruction (EAI) and Business as usual (BAU) differ in terms of their effects on students' learning outcomes regarding the strategy choice, overall skill implementation ability and attribute mastery status?

Chapter 2: Literature Review

Given that this study aims at developing a longitudinal CDM for multiple strategies, the literature review is conducted in two topic areas, CDMs and latent variable models for multiple strategies. Within each area, the models are further categorized based on their application settings, that is, whether they are designed for the data from a single time point or for data from multiple time points.

It should be noted that the models in these two areas are not mutually exclusive. Instead, CDMs and the latent variable models for multiple strategies are closely related: 1) Both of them are embraced in the latent variable modeling framework; 2) Both of them are based on mixture-distribution models (McLachlan & Basford, 1988); 3) CDMs can be used as the measurement model in the multiple-strategy models; and 4) In the longitudinal settings, both of them can be extended to model the sequential change of the latent variables by incorporating the latent transition analysis (LTA; Collins & Wugalter, 1992).

2.1 Theoretical Foundation

A brief introduction to the latent variable modeling framework, mixture-distribution model and LTA is provided as the theoretical foundation.

Latent variable modeling framework. By definition, latent variables refer to the variables that are not observable but the values of which can be inferred from the observed variables. Latent variable models are statistical models containing latent variables, and these models are the linkage between the latent and observed variables (Spearman, 1904). Psychometric models are latent variable models and they are used for psychological and educational measurement. In the field of educational

measurement, categorical item responses are usually used as observed variables. This review focuses on the models designed for dichotomous item responses as the proposed model is to be applied to dichotomous response data, but it should be noted most of the models introduced below have been extended to accommodate polytomous responses.

Latent variables can be continuous or categorical. Examples of latent variables are persons' ability and skills. Different psychometric models can be used to measure latent variables of different nature. For instance, when the latent variables are continuous latent traits, IRT models can be used; when the latent variables are attributes or skills whose mastery status is categorical (binary in most cases), CDMs should be considered.

Mixture-distribution model. Mixture-distribution models (McLachlan & Basford, 1988) are used when the sample consists of subjects from different subpopulations or latent classes. The term "latent" is used if the class membership is unobservable. The distribution of the observed data is conditional on the latent class membership. Mathematically, in the mixture model, the marginal probability of the observed data can be written as:

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{c=1}^{C} \pi_c P(\mathbf{Y} = \mathbf{y} \mid c),$$

where c is the discrete latent class variable. π_c is the mixing proportion of class c, which corresponds to the class size. $P(\mathbf{Y} = \mathbf{y} \mid c)$ is the conditional probability of the observed data given class c. When the observed data are categorical, the mixture

model can be referred to as latent class analysis (LCA; e.g., Lazarsfeld & Henry, 1968)

Latent transition analysis. LTA (Collins & Wugalter, 1992) can be treated as a longitudinal extension of the LCA. In the LTA, the data at each time point is modeled with an LCA, but the latent class membership of an individual is allowed to change over time. The latent class progressions are represented with latent transition probabilities. The marginal probability of the observed data in the LTA is specified as

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{c^{(1)}=1}^{C_1} \cdots \sum_{c^{(T)}=1}^{C_T} \pi_{c^{(1)}} \prod_{t'=2}^{T} \tau_{c^{(t')}|c^{(t'-1)}}^{(t'-1)} P(\mathbf{Y} = \mathbf{y} \mid c^*)$$

where $\pi_{c^{(1)}}$ is the mixing proportion of latent class $c^{(1)}$ at the initial time point. $\tau_{c^{(t)}|c^{(t-1)}}^{(t-1)}$ is the latent transition probability from the latent class $c^{(t-1)}$ to $c^{(t)}$ at time point $(t-1)$; that is, for an individual who is classified as $c^{(t-1)}$ at time point $(t-1)$, the probability of this individual transitioning to latent class $c^{(t)}$ at time point t' is $\tau_{c^{(t)}|c^{(t-1)}}^{(t-1)}$. $c^* = (c^{(1)}, ..., c^{(T)})$ represents the latent class progression pattern. $P(\mathbf{Y} = \mathbf{y} \mid c^*)$ is the conditional probability of the observed data given latent class pattern c^*.

2.2 CDM for a Single Time Point

CDMs are psychometric models that aim at providing fine-grained diagnostic information about students' mastery status on a series of attributes. When applied to response data from cognitive diagnostic assessments, CDMs can be used to classify students into latent classes, each of which is defined by a unique attribute mastery profile. Thus, inferences can be made about students' attribute mastery status. CDMs can be treated as special cases of the discrete mixture-distribution model (McLachlan

& Basford, 1988) where the latent classes are defined by attribute profiles and the outcome variables are categorical (e.g., Rupp et al., 2010). In CDMs, the marginal probability of the observed response patterns of respondent *j* is given as

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{\boldsymbol{\alpha_c} \in \mathbf{A}} \upsilon_{\boldsymbol{\alpha_c}} \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid \boldsymbol{\alpha_c}),$$

where $\boldsymbol{\alpha_c} = (\alpha_1, ..., \alpha_K)$ indicates the latent class defined by an attribute mastery status pattern, assuming that *K* attributes are measured by the assessment and each attribute only has two mastery statuses, i.e., mastery ($\alpha_k = 1$) and non-mastery ($\alpha_k = 0$). For example, when two attributes are measured, $\boldsymbol{\alpha_c} = (1, 0)$ represents a latent class of respondents who have mastered the first attribute but have not mastered the second one. $\mathbf{A}$ represents all the permissible attribute mastery status patterns; the maximum number of latent classes is 2^K, if all the attribute profile patterns are permissible. $\upsilon_{\boldsymbol{\alpha_c}}$ is the mixing proportion of class $\boldsymbol{\alpha_c}$ and $P(Y_{ij} = y_{ij} \mid \boldsymbol{\alpha_c})$ represents the conditional response probability given the latent class.

There are numerous ways to specify the distribution of latent classes, $\boldsymbol{\alpha_c}$. In a saturated form, the mixing proportion parameter, $\upsilon_{\boldsymbol{\alpha_c}}$, of every latent class (except the reference latent class) can be freely estimated, with the constraint, $\sum_{\boldsymbol{\alpha_c} \in \mathbf{A}} \upsilon_{\boldsymbol{\alpha_c}} = 1$. If one or more attribute profile patterns are known to be impermissible, the corresponding mixing proportion(s) can be constrained to zero (e.g., Liu & Huggins-Manley, 2016); the mixing proportions of the remaining classes are freely estimated. Alternatively, a general unidimensional ability parameter, θ_j, can be assumed to underlie all the

attributes as in the higher-order structure specified by de la Torre and Douglas (2004):

$$P(\alpha_{jk}=1|\theta_j)=\frac{\exp(\xi_k\theta_j+\beta_k)}{1+\exp(\xi_k\theta_j+\beta_k)}, \quad (1)$$

where ξ_k and β_k represent the factor loading and intercept corresponding to attribute *k*, respectively. As the parameterization in equation 1 resembles the two parameter logistic (2PL; Birnbaum, 1968) IRT model, ξ_k and β_k are referred to as the attribute discrimination and easiness parameters in the subsequent sections. It is assumed that the attribute mastery probabilities are locally independent given the ability (de la Torre & Douglas, 2004). When the higher-order structure is used, the marginal probability of the observed response patterns is written as

$$P(\mathbf{Y_j}=\mathbf{y_j})=\int\left[\prod_{k=1}^{K}P(\alpha_{jk}=1|\theta)P(\mathbf{Y_j}=\mathbf{y_j}|\boldsymbol{\alpha_c})\right]f(\theta)d\theta.$$

Under the higher-order structure, the attribute discrimination and easiness parameters, instead of the latent class mixing proportions, are to be estimated. Thus, the higher-order structure results in fewer model parameters than the saturated form, which can improve the estimation efficiency when using the Bayesian Markov chain Monte Carlo (MCMC) estimation method (de la Torre & Douglas, 2004).

A variety of CDMs have been proposed to model the item response probability, $P(Y_{ij}=y_{ij}|\boldsymbol{\alpha_c})$, and CDMs were originally used to model data from a single time point. Examples of the commonly-used CDMs are deterministic input, noisy "and" gate (DINA; Junker & Sijtsma, 2001; Macready & Dayton, 1977) model, deterministic input, noisy "or" gate (DINO; Templin & Henson, 2006), log-linear

cognitive diagnosis model (LCDM; Henson et al., 2009) and general diagnostic model (GDM; von Davier, 2005). Different CDMs are designed for different purposes and rest on various assumptions. The DINA model is one of the simplest and most widely-used CDMs. It is a non-compensatory CDM assuming that, ideally, a student can correctly respond to an item only if he or she masters all the required attributes of the item. Specifically, the ideal item response of respondent *j* to item *i* is given as

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ik}} ,$$

where α_{jk} is an indicator of whether respondent *j* masters attribute *k*; and q_{ik}, as an element in the Q-matrix, indicates whether item *i* requires attribute *k*. The Q-matrix is an item-by-attribute matrix. Each element in the Q-matrix, also referred to as "q-entry", is a binary entry with the following denotation:

$$q_{ik} = \begin{cases} 1, & \text{the } i\text{th item requires the } k\text{th attribute} \\ 0, & \text{the } i\text{th item does not require the } k\text{th attribute} \end{cases}$$

The Q-matrix is an important component in CDMs, as it describes the mapping relationships between items and attributes (Tatsuoka, 1983, 1985) and reflects the cognitive specification of a test (Leighton et al., 2004). Thus, in the DINA model, the probability of a correct response is written as

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_{\mathbf{c(j)}}) = (1 - s_i)^{\eta_{ij}} (g_i)^{1-\eta_{ij}} = (1 - s_i)^{\prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}} (g_i)^{1-\prod_{k=1}^{K} \alpha_{jk}^{q_{ik}}} ,$$

where $\boldsymbol{\alpha}_{\mathbf{c(j)}}$ is the attribute mastery status pattern of respondent *j*; $g_i = P(Y_{ij} = 1 \mid \eta_{ij} = 0)$ and $s_i = P(Y_{ij} = 0 \mid \eta_{ij} = 1)$ represent the guessing and slipping probabilities,

respectively. The constraint, $g_i < 1 - s_i$, is set to ensure that individuals who lack one or more required attributes have a lower probability of success than those who master all the required attributes (Junker & Sijtsma, 2001). A compensatory counterpart of the DINA model is the DINO model. The DINO model assumes that students can correctly respond to an item as long as he or she masters one of the required attribute(s), when slipping does not occur.

Both the DINA and DINO models are highly restricted models relying on strong assumptions. For example, the DINA model does not differentiate the correct response probabilities among respondents who lack one or more attributes; the DINO model does not differentiate the correct response probabilities among respondents who master one or more attributes. These assumptions hardly hold true in reality. Some more generalized CDMs based on weaker assumptions have been developed. The LCDM is a generalized CDM where the correct response probability is written as

$$P(Y_{ij} = 1 | \boldsymbol{\alpha}_{\mathbf{c(j)}}) = \frac{\exp[\lambda_{i,0} + \boldsymbol{\lambda}_{\mathbf{i}}^{\mathrm{T}} h(\boldsymbol{\alpha}_{\mathbf{c(j)}}, \mathbf{q}_{\mathbf{i}})]}{1 + \exp[\lambda_{i,0} + \boldsymbol{\lambda}_{\mathbf{i}}^{\mathrm{T}} h(\boldsymbol{\alpha}_{\mathbf{c(j)}}, \mathbf{q}_{\mathbf{i}})]}, \quad (2)$$

where $\lambda_{i,0}$ is the item intercept parameter; it can be interpreted as a guessing parameter in the sense that it is the logit of the correct response probability when no attribute is mastered. $\boldsymbol{\lambda}_{\mathbf{i}}^{\mathrm{T}} h(\boldsymbol{\alpha}_{\mathbf{c(j)}}, \mathbf{q}_{\mathbf{i}})$ is a linear combination of the main and interaction effects of the required attributes, i.e.,

$$\boldsymbol{\lambda}_{\mathbf{i}}^{\mathrm{T}} h(\boldsymbol{\alpha}_{\mathbf{c(j)}}, \mathbf{q}_{\mathbf{i}}) = \sum_{k=1}^{K} \lambda_{i,1,(k)} \alpha_{jk} q_{ik} + \sum_{k_1 < k_2} \lambda_{i,2,(k_1,k_2)} \alpha_{jk_1} \alpha_{jk_2} q_{ik_1} q_{ik_2} + \ldots,$$

where $\lambda_{i,1,(k)}$ is the main effect of attribute *k*. $\lambda_{i,2,(k_1,k_2)}$ is the two-way interaction effect of attributes k_1 and k_2. The higher-order interactions can be represented by

$\lambda_{i,d,(k_1,...,k_d)}$, $d = 3,...,K$, where d is the order of the interaction. In the saturated form of the LCDM, all the main and interaction effects are included. Alternatively, main and/or interaction terms can be dropped to reduce the number of estimated parameters. The LCDM is more generalized than the DINA or DINO models as it allows each attribute to uniquely contribute to the correct response probability. In fact, the DINA model can be written as a constrained version of the LCDM where only the highest-order interaction term is retained for the complex items (Rupp et al., 2010). Another special case of the LCDM is the linear logistic model (LLM; Hagenaars, 1990, 1993; Maris, 1999) that only retains the main effect terms of the LCDM. Note that the ordering constraints need to be set on the main effect and each interaction effect in order to ensure model identification for the LCDM: all the main effects are constrained to be positive; interactions are constrained to ensure that respondents who master more required attributes would have higher success probabilities (Lao, 2016).

The GDM is a generalized model which encompasses the LCDM as a special case (von Davier, 2014). In addition, the GDM allows the latent attributes to be continuous or discrete. A number of other CDMs are not elaborated here as the purpose of this literature review is to provide an overview and to lay a foundation to the proposed model. A more comprehensive review of CDMs that are designed for data from a single time point can be found in Rupp and Templin (2008b).

2.3 CDM for Multiple Time Points

Recent years have seen the development of longitudinal CDMs that model data from multiple time points (Hansen, 2013; Huang, 2017; Kaya & Leite, 2017;

Lee, 2017; F. Li et al., 2016; Madison & Bradshaw, 2018a, 2018b; Pan, 2018; S. Wang et al., 2018; Zhan, Jiao, Liao, et al., 2019). In general, these longitudinal CDMs can be divided into two categories, one based on the LTA and the other based on the growth modeling. The two categories of models assume different attribute distributions: the LTA-based CDMs assume that the attributes follow a discrete distribution, while the growth-model-based CDMs rest on higher-order structures underlying the attributes. Accordingly, the operational definitions of "growth" vary across the two categories of CDMs.

The LTA-based CDMs (Kaya & Leite, 2017; F. Li et al., 2016; Madison & Bradshaw, 2018a, 2018b; S. Wang et al., 2018), using the concept of "transition", assume that the attributes follow a discrete distribution. This category of longitudinal CDMs focuses on the probabilities of respondents transitioning from one attribute mastery status latent class to another over time (usually over adjacent time points). In the LTA-based CDMs, the marginal probability of the observed response pattern is specified as

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{\boldsymbol{\alpha}_c^{(1)} \in \mathbf{A_1}} \cdots \sum_{\boldsymbol{\alpha}_c^{(T)} \in \mathbf{A_T}} \upsilon_{\boldsymbol{\alpha}_c^{(1)}} \prod_{t'=2}^{T} \tau_{\boldsymbol{\alpha}_c^{(t')} | \boldsymbol{\alpha}_c^{(t'-1)}}^{(t'-1)} \prod_{t=1}^{T} \prod_{i=1}^{I} P(Y_{ij}^{(t)} = y_{ij}^{(t)} \mid \boldsymbol{\alpha}_c^{(t)}),$$

where $\upsilon_{\boldsymbol{\alpha}_c^{(1)}}$ is the mixing proportion of latent class $\boldsymbol{\alpha}_c^{(1)}$ at the initial time point. $\tau_{\boldsymbol{\alpha}_c^{(t)} | \boldsymbol{\alpha}_c^{(t-1)}}^{(t-1)}$ is the latent transition probability from the latent class $\boldsymbol{\alpha}_c^{(t-1)}$ to $\boldsymbol{\alpha}_c^{(t)}$ at time point (t-1); that is, for a respondent who is classified as $\boldsymbol{\alpha}_c^{(t-1)}$ at time point (t-1), the probability of this respondent switching to latent class $\boldsymbol{\alpha}_c^{(t)}$ at time point t is $\tau_{\boldsymbol{\alpha}_c^{(t)} | \boldsymbol{\alpha}_c^{(t-1)}}^{(t-1)}$. The latent transition probability can be directly estimated (e.g., Li et al., 2016; Madison &

Bradshaw, 2018) or further decomposed as a combination of covariates (S. Wang et al., 2018). Different measurement models have been used to model $P(Y_{ij}^{(t)} = y_{ij}^{(t)} \mid \boldsymbol{\alpha}_{\mathbf{c}}^{(\mathbf{t})})$ in the LTA-based CDMs, including the DINA (F. Li et al., 2016), DINO (Kaya & Leite, 2017) and LCDM (Madison & Bradshaw, 2018b). The growth in the LTA-based CDMs can be quantified as the change in the proportion of examinees who are classified as attribute mastery over time (Madison & Bradshaw, 2018a, 2018b). Madison and Bradshaw (2018a) further incorporated a multigroup structure into an LTA-based CDM to assess the differential growth among multiple manifest groups (e.g., a control group and a treatment group).

In contrast, most growth-model-based CDMs assume a higher-order structure (de la Torre & Douglas, 2004) with continuous latent trait(s) underlying the attributes (e.g., Lee, 2017; Zhan, Jiao, Liao, et al., 2019), so that the growth can be defined as the change in the continuous latent trait(s) over time. The marginal probability of the observed response pattern is given as

$$P(\mathbf{Y}_{\mathbf{j}}^{(\mathbf{t})} = \mathbf{y}_{\mathbf{j}}^{(\mathbf{t})}) = \int \left[\prod_{k=1}^{K} P(\alpha_{jk}^{(t)} = 1 \mid \boldsymbol{\theta}) \prod_{i=1}^{I} P(Y_{ij}^{(t)} = y_{ij}^{(t)} \mid \boldsymbol{\alpha}_{\mathbf{c}}^{(\mathbf{t})}, \boldsymbol{\theta}) \right] f(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where the latent ability, $\boldsymbol{\theta}$, is multidimensional given that multiple time points are involved.

The denotation, parameterization and structure of $\boldsymbol{\theta}$ vary across different studies. Hansen (2013) applied the hierarchical diagnostic model to model data from multiple time points. Only one attribute is measured at each time point. Correlations of the attribute across time points are allowed by specifying continuous latent variables, $\boldsymbol{\theta}$, underlying the attribute across time points. Zhan, Jiao, Liao, et al.

(2019), following the model of Andersen (1985), denoted $\boldsymbol{\theta}$ as a vector of ability at multiple time points, i.e., $\boldsymbol{\theta}_{\mathbf{j}} = (\theta_j^{(1)}, ..., \theta_j^{(T)})'$, and specified that $\boldsymbol{\theta}_{\mathbf{j}}$ follows a multivariate normal distribution, i.e., $\boldsymbol{\theta}_{\mathbf{j}} \sim MVN(\boldsymbol{\mu}_{(\boldsymbol{\theta})}, \boldsymbol{\Sigma}_{(\boldsymbol{\theta})})$. Lee (2017) used a growth curve model to model the change of latent ability over time. The growth curve model allows individual-specific growth curves. Statistically, the ability parameter at each time point is a function of a time covariate, and that the slope and intercept are random, i.e.,

$$\theta_j^{(t)} = \varsigma_{j0} + (\varpi + \varsigma_{j1}) \times \text{time}_j^{(t)} + \varepsilon_j^{(t)}$$

where $\text{time}_j^{(t)}$ is the time covariate. $\varepsilon_j^{(t)}$ is the error term. In this model, both intercept and slope are specified as random parameters: ς_{j0} is the random intercept; $(\varpi + \varsigma_{j1})$ is the slope consisting of a fixed component, ϖ, and a random component, ς_{j1}. The random effects of the intercept and slope follow a multivariate normal distribution, $\begin{pmatrix} \varsigma_{j0} \\ \varsigma_{j1} \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_{(\boldsymbol{\varsigma})}\right]$, and the error term follows a normal distribution, $\varepsilon_j^{(t)} \sim N(0, \sigma^2_{(\varepsilon)t})$. Further, a multivariate version of the growth curve model can be used to directly model the change of attribute mastery probability over time, without assuming a latent trait underlying the attributes (Pan, 2018).

In the longitudinal CDM studies, there is not a consistent rule of whether and how to handle the dependencies of the repeated items over time. Madison and Bradshaw (2018) did not consider the dependencies of the repeated items over time, while Hansen (2013) and Zhan, Jiao, Liao, et al. (2019) included random effects to account for the dependencies of the repeated items over time. However, all the studies

above assume that the repeated items are measurement invariant, i.e., have the same item parameters over time.

The measurement invariance is an important assumption in longitudinal modeling in the education studies as it is a prerequisite to the meaningful comparison of the latent trait or attribute mastery status across timepoints (e.g., Meredith & Horn, 2001). In the longitudinal CDMs, the measurement invariance assumption holds when the conditional distribution of the observed response patterns given the attribute profile remain identical across time points (Madison & Bradshaw, 2018b). The majority of existing studies on the longitudinal CDM assume measurement invariance without testing the assumption with Madison and Bradshaw (2018) as an exception. Madison and Bradshaw (2018) examined the robustness of an LTA-based CDM to the violation of the measurement invariance assumption due to item parameter drift (IPD). They found that, under the simulated conditions, the item parameter estimates are less accurate when the IPD exists, but the classification accuracy is robust to IPD. Specifically, In the simulation study conducted by Madison and Bradshaw (2018), three manipulated factors considered were related to the item parameter drift (IPD) over time. Specifically, they considered the percentages of items with IPD (0%, 20%, 40%, 60%, 80%, and 100%), the magnitude of IPD (i.e., difference in item parameters) over time (0.5 and 1) and the IPD type (IPD only in the item intercept parameters, IPD only in the main and interaction parameters and IPD in all the item parameters). They found that, in their proposed model (i.e., Transition Diagnostic Classification Model), the classification accuracy rates were higher than 0.9 in all the

simulated conditions. Further, the classification accuracy rate only decreased by 0.01 in the 100% IPD conditions compared to the 0% IPD conditions.

Nevertheless, IPD may not be the only factor leading to the violation of measurement invariance. The shift in problem-solving strategy, manifested as a difference in the Q-matrix (e.g., de la Torre & Douglas, 2008), could also result in the violation of measurement invariance. No study has investigated the impact of the drift of Q-matrix on the performance of longitudinal CDMs. In addition, the drift of Q-matrix could be associated with the Q-matrix misspecification issues in longitudinal CDMs. Studies on single-time-point CDMs have found that misspecified Q-matrices can result in inaccurate item parameter estimates and lower the classification accuracy (Im & Corter, 2011; Rupp & Templin, 2008a). Therefore, the variability of the Q-matrix over time is worth exploring.

Using the same Q-matrix for all the individuals and for all the time points, existing studies on longitudinal CDMs (e.g., Li et al., 2016; Madison & Bradshaw, 2018b; Zhan, Jiao, Liao, et al., 2019) assume that all the respondents employ a uniform type of problem-solving strategy at each time point and each individual would not change the problem-solving strategy over time. Such assumptions may not be realistic given that individuals may choose to use different strategies at different stages of cognitive development (Siegler et al., 1981) and a number of educational programs have been designed for improving students' problem-solving strategies (e.g., Mercer & Mercer, 2001). Thus, it is reasonable to expect that there are a variety of problem-solving strategies among the population at a single time point and an

individual's strategy can change over time, which motivates the proposal of a longitudinal CDM incorporating the multiple strategies and strategy shift.

2.4 Modeling Multiple Strategies at a Single Time Point

Discrete mixture-distribution models combined with the IRT models or CDMs, have been used to model multiple strategies (de la Torre & Douglas, 2008; Mislevy & Verhelst, 1990; Rost, 1990; von Davier, 2010; Yamamoto, 1989). Such mixture models for problem-solving strategies are referred to as measurement models with narrative structures by Mislevy and Huang (2007).

In a single-time-point scenario, the problem-solving process generally consists of two stages. In the first stage, a respondent chooses a strategy; in the second stage, the respondent implements the required skills and makes a response. As the strategy chosen by a person is unobservable, the strategy is treated as a categorical latent variable in the discrete mixture-distribution model and the mixing proportions of the categories are to be estimated. Mathematically, the marginal probability of the observed response pattern of respondent *j* is written as

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{m=1}^{M} \pi_m \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid m) ,$$

where π_m is the mixing proportion of strategy *m*, indicating the proportion of respondents choosing strategy *m*; $P(Y_{ij} = y_{ij} \mid m)$ is the item response probability conditional on strategy *m*, which can be modeled with a chosen measurement model, such as an IRT model or a CDM.

IRT models as the measurement model. When IRT models are used as the measurement model, the item response probability is a function of the continuous

person ability parameter, θ. Different strategies can be characterized by different item parameters, ability distributions and/or different measurement models (e.g., Mislevy & Verhelst, 1990; Rost, 1990; Yamamoto, 1989). One of the simplest mixture IRT models is the mixed Rasch model (Rost, 1990) where the marginal probability of the observed response pattern of respondent *j* is given as (assuming local item independence):

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{m=1}^{M} \pi_m \int \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid m, \theta) f(\theta \mid m) d\theta ,$$

where $f(\theta \mid m)$ is the distribution of the ability parameter conditional on strategy *m*. The response probability of an individual item is written as the Rasch model (Rasch, 1960):

$$P(Y_{ij} = 1 \mid \theta_j, m, \beta_{im}) = \frac{\exp(\theta_j - \beta_{im})}{1 + \exp(\theta_j - \beta_{im})}, \qquad (3)$$

$$P(Y_{ij} = 0 \mid \theta_j, m, \beta_{im}) = 1 - P(Y_{ij} = 1 \mid \theta_j, m, \beta_{im}) .$$

As in the Rasch model, β_{im} is the item difficulty parameter. However, in the mixture Rasch model for multiple strategies, β_{im} is also dependent on the discrete latent class variable, *m*, and, therefore, is strategy-specific.

Mislevy and Verhelst (1990) noted that the strategy-specific item difficulty parameter could be associated with the nature of the strategy and the characteristics of the item. Therefore, they proposed the mixture linear logistic test model (MLLTM). Specifically, the item difficulty parameter in the mixed Rasch model is decomposed as a linear combination of item features associated with the strategy:

$$\beta_{im}(\boldsymbol{\gamma}) = \sum_{l=1}^{L} R_{iml}\gamma_{ml}$$

where each element in matrix $\mathbf{R_{im}}$ indicates whether a strategy-related feature *l* is present in item *i*; the value of the element can be dichotomous or polytomous. γ_{ml} represents the contribution of feature *l* to the item difficulty parameter. Both $\mathbf{R_{im}}$ and γ_{ml} can be predetermined based on theories.

While both the mixed Rasch model and the MLLTM used the Rasch-type models as measurement models for all the strategies, different types of measurement models could be used for different strategies. In the HYBRID IRT model proposed by Yamamoto (1987, 1989), two subpopulations (i.e., normal respondents and random guessers) who implement different strategies are assumed. The normal respondents are modeled with an IRT model while the random guessers are modeled with a latent class analysis (LCA) model**.** As a special case, the HYBRID Rasch model (von Davier & Yamamoto, 2007) consists of the "RASCH" and "LCA" latent classes. In the HYBRID Rasch model, the marginal probability of the observed response pattern is written as

$$P(\mathbf{Y_j} = \mathbf{y_j} \mid \boldsymbol{\pi}) =$$
$$\pi_{RASCH} \int P(\mathbf{Y_j} = \mathbf{y_j} \mid M_{RASCH}, \theta) f(\theta \mid M_{RASCH}) d\theta + \sum_{m \in \mathbf{M_{LCA}}} \pi_m P(\mathbf{Y_j} = \mathbf{y_j} \mid m),$$

The measurement model of the "RASCH" class is the same as that of the mixed Rasch model (equation 3). The measurement model of the "LCA" class is written as:

$$P(Y_{ij} = 1 \mid m) = g_m,\ m \in \mathbf{M_{LCA}}.$$

CDMs as the measurement model. When CDMs are used as the measurement model, the conditional item response probability is a function of the

person attribute profiles, $\boldsymbol{\alpha}$. Different strategies can be characterized by different Q-matrices as well as different distributions of attribute profiles (e.g., de la Torre & Douglas, 2008; von Davier, 2007). As a natural extension of the mixture IRT models to the CDM framework, the mixture GDM (von Davier, 2007) has the marginal probability of the observed response pattern:

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{m=1}^{M} \pi_m \sum_{c=1}^{C} P(\boldsymbol{\alpha}_\mathbf{c} \mid m) \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid \boldsymbol{\alpha}_\mathbf{c}, m),$$

where each strategy *m* corresponds to a unique Q-matrix. In some multiple-strategy CDMs, the strategy choice is specified to be dependent on the attribute mastery status (e.g., de la Torre & Douglas, 2008; Ma & Guo, 2019). When the strategy choice is conditional on the person attribute profiles, $\boldsymbol{\alpha}$, the marginal probability of the response pattern is specified as

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{c=1}^{C} P(\boldsymbol{\alpha}_\mathbf{c}) \sum_{m=1}^{M} \pi_{m|\boldsymbol{\alpha}_\mathbf{c}} \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid \boldsymbol{\alpha}_\mathbf{c}, m),$$

where $\pi_{m|\boldsymbol{\alpha}_\mathbf{c}} = P(m \mid \boldsymbol{\alpha}_\mathbf{c})$ is the probability of choosing strategy *m* given the attribute profiles, $\boldsymbol{\alpha}_\mathbf{c}$. The relationship between strategy choice and the attribute mastery statuses, $\boldsymbol{\alpha}$, can be either deterministic or probabilistic. In the multiple-strategy DINA (MS-DINA; de la Torre & Douglas, 2008) model, it is assumed that respondents would choose a strategy as long as they master all the attributes required by the strategy. Mathematically, the ideal response in the MS-DINA model is specified as

$$\eta_{ij} = \max\{\eta_{ij1}, \eta_{ij2}, ..., \eta_{ijM}\},$$

where $\eta_{ijm} = \prod_{k=1}^{K} \alpha_{jk}^{q_{ikm}}$ is the ideal item response corresponding to strategy *m* and its Q-matrix, $\mathbf{Q_m}$. In this sense, respondents' strategy choice is determined by their attribute mastery statuses. In contrast, Ma and Guo (2019) relates the strategy choice probability to the conditional correct response probability given the attribute profile and strategy:

$$P(m \mid \boldsymbol{\alpha}_\mathbf{j}) = \frac{P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_\mathbf{j}, m)^u}{\sum_{m=1}^{M_i} P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_\mathbf{j}, m)^u},$$

where *u* is referred to as the "strategy selection" parameter. In particular, when *u*=0, all the strategies are equally likely to be chosen; when *u*=1, the probability of choosing a strategy is proportional to the conditional correct response probability given this strategy.

Comparisons between IRT models and CDMs as measurement model.

The mixture-distribution models, incorporating the IRT models or CDMs as measurement models, have been used for multiple-strategy modeling. Regardless of the type of the measurement model, the strategy is indicated by a discrete latent variable in the mixture models. However, the operational definitions of problem-solving strategy can vary across different measurement models. In the CDM framework, strategies are defined by their unique cognitive processes. In particular, each strategy, characterized by a Q-matrix, is defined by a unique combination of attributes involved in the problem-solving process. In contrast, strategies tend to be defined based on their "outcomes" of item functioning in the IRT framework. For

example, items could be deemed more difficult for the subpopulation employing one strategy than another.

2.5 Modeling Multiple Strategies at Multiple Time Points

In the longitudinal setting, the LTA-mixture Rasch model has been proposed to model strategy shift over time and has been used to evaluate the effectiveness of an education intervention (Cho et al., 2010). In the LTA-mixture Rasch model, the marginal probability of the response pattern at time point *t* is:

$$P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}}) = \sum_{m_1=1}^{M_1} \cdots \sum_{m_T=1}^{M_T} \pi_{m_1} \prod_{t'=2}^{T} \tau_{m_{t'}|m_{t'-1}}^{(t'-1)} P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}} \mid m^*),$$

where $m^* = (m_1, m_2, ..., m_T)$ is the strategy progression over time. $\tau_{m_{t'}|m_{t'-1}}^{(t'-1)}$ is the latent transition probability from strategy $m_{t'-1}$ to $m_{t'}$ at time point $t'-1,\ t' = 2,...T$. In the LTA-mixture Rasch model, each latent class is a strategy pattern, and the mixing proportion of a latent class is $\pi_{m_1} \prod_{t'=2}^{T} \tau_{m_t|m_{t-1}}^{(t'-1)}$. The conditional probability of the response pattern given a latent class is specified as

$$P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}} \mid m^*) =$$
$$\int \prod_{i=1}^{I} P(Y_{ij}^{(t)} = y_{ij}^{(t)} \mid m^*, \boldsymbol{\theta}) f(\boldsymbol{\theta} \mid m^*) d\boldsymbol{\theta},\ \boldsymbol{\theta} = (\theta^{(1)}, ..., \theta^{(T)})',$$

where the ability, $\boldsymbol{\theta}$, is multidimensional and follows the structure specified by Andersen (1985). Specifically, within each latent class m^*, the ability parameters follow a multivariate normal distribution, i.e., $\boldsymbol{\theta}_{j \in m^*} \mid m^* \sim MVN(\boldsymbol{\mu}_{(\theta)\mathbf{m}^*}, \boldsymbol{\Sigma}_{(\theta)\mathbf{m}^*})$. The mean and variance of the ability parameter at the first time point in the first latent class are set to be 0 and 1, respectively, for scale identification. The conditional probability of a correct item response at time point *t* is given as

$$P(Y_{ij}^{(t)} = 1 | \theta_j^{(t)}, m, \beta_{im^*}^{(t)}) = \frac{\exp(\theta_j^{(t)} - \beta_{im^*}^{(t)})}{1 + \exp(\theta_j^{(t)} - \beta_{im^*}^{(t)})}.$$

The item parameter, $\beta_{im^*}^{(t)}$, is allowed to vary across latent classes but is constrained to be invariant over time, i.e., $\beta_{im^*}^{(t)} = \beta_{im^*}$, for scale comparability. Cho et al. (2013) extended the LTA-mixture Rasch model to accommodate a multilevel structure.

2.6 Parameter Estimation of the CDMs for Multiple Time Points or Multiple Strategies

The maximum likelihood estimation (MLE) with the Expectation-Maximization (EM) algorithm and the Bayesian MCMC method are two frequently used parameter estimation methods for the CDMs with multiple strategies or multiple time points (e.g., de la Torre & Douglas, 2008; Huo & de la Torre, 2014; Madison & Bradshaw, 2018b; Zhan, Jiao, Liao, et al., 2019; Zhan, Jiao, Man, et al., 2019). The two estimation methods differ in their assumptions and optimization algorithms. The MLE is a frequentist approach which assumes that each parameter is fixed. The MLE finds the parameter values that maximize the likelihood function and uses them as the parameter estimates. In contrast, the Bayesian approach assumes that each parameter is a random variable which is represented by a probability distribution. In the Bayesian estimation, the prior knowledge about a parameter (i.e., the prior distribution) is updated with the knowledge gained from the observed data (i.e., the likelihood) to yield the updated knowledge about the parameter (i.e., the posterior distribution). Thus, in the Bayesian estimation methods, the estimate and standard error of a parameter is obtained by summarizing the mean and standard deviation, respectively, of the posterior distribution of the parameter. However, despite the

difference in the estimation algorithm, evidence has been found that the MLE and the Bayesian estimation methods yield comparable parameter estimates in the CDMs (e.g., Huo & de la Torre, 2014).

Given that the comparative efficiencies of the two estimation methods vary from case to case, the choice between the MLE and the Bayesian MCMC method depends on a variety of factors. One factor that could affect the choice of the estimation method is the distribution assumed underlying the attribute profile latent classes. In general, the MLE is more likely to be chosen when the attribute profiles are assumed to follow a discrete distribution (e.g., Huo & de la Torre, 2014; Madison & Bradshaw, 2018b), while the Bayesian MCMC is favored when a higher-order structure (equation 1) is assumed underlying the attributes (e.g., de la Torre & Douglas, 2008; Zhan, Jiao, Man, et al., 2019) with Zhan, Jiao, Liao, et al. (2019) being an exception. The MLE tends to be more efficient than the Bayesian MCMC when the attribute distribution is discrete for the MS-DINA model (Huo & de la Torre, 2014). However, the memory required by the MLE increases as the number of the attributes increases and, thus, even the MLE could become burdensome when the number of attributes is extremely large. Compared to the MLE, the Bayesian MCMC is more flexible to be applied to different formulations of the models. For example, it is straightforward to estimate the parameters of the HO-DINA model with the Bayesian MCMC method, which is relatively hard with the MLE (de la Torre, 2009).

As for the software program, flexMIRT (Houts & Cai, 2015) and Mplus (Muthén & Muthén, 2007) have been used to carry out the MLE for the longitudinal CDMs (Madison & Bradshaw, 2018b; Zhan, Jiao, Liao, et al., 2019); JAGS

(Plummer, 2015) has been used to carry out the Bayesian MCMC estimation for the longitudinal CDMs (Zhan, Jiao, Man, et al., 2019). Ox (Doornik, 2009) has been used to implement both the MLE and the Bayesian MCMC for estimating the multiple-strategy CDMs (de la Torre & Douglas, 2008; Huo & de la Torre, 2014). The function and accessibility of the estimation tools could also be taken into account while choosing the appropriate estimation method.

2.7 Summary of the Literature Review

This chapter reviews the literature on CDMs and models for multiple strategies. Three limitations are identified and this study is motivated to filling the gaps in the existing literature. First, the existing longitudinal CDMs, using a single Q-matrix for all the respondents over time, are prone to the Q-matrix misspecification issues. The between-person multiple strategies could result in different Q-matrices for different subpopulations and the within-person strategy shift could result in a drift of Q-matrix over time. Using the same Q-matrix for all the respondents across time points could render the Q-matrix misspecified for at least some of the respondents.

Second, the existing models for within-person strategy shift (e.g., Cho et al., 2010) are originated in the IRT framework and are limited in providing fine-grained diagnostic information. Like the mixture IRT models designed for a single time point, the LTA-mixture Rasch model distinguished different problem-solving strategies by different item functioning as opposed to different cognitive processes. Thus, a shift in strategy is characterized by a shift in the item functioning. For example, a research question that can be answered using the LTA-mixture Rasch model is that “For an individual shifting from Strategy A to Strategy B, which item(s) become easier for

this individual?" However, in the IRT framework, it is hard to answer questions like "For an individual shifting from Strategy A to Strategy B, does this individual tend to utilize additional attributes to solve the problems?" Using CDMs that define strategies as unique cognitive processes could make it possible to answer the latter type of research questions. Currently, no study has been done to model strategy shift over time in the CDM framework.

Third, previous latent variable models seldom explicitly distinguish the concept of strategies from skills. The model specifications (e.g., Rost, 1990; von Davier, 2007) imply dependencies among strategies and skills. For example, von Davier (2007) allowed the distributions of the attribute mastery profiles and/or item parameters to vary across strategy latent classes. Alternatively, Ma and Guo (2019) specified the probability of the strategy choice as a function of the attribute mastery statuses. However, from a diagnostic perspective, it may be worthwhile to separate strategies from skills: diagnostic models that separate the strategy choice process (identifying the attributes required to solve the problem) from the skill implementation process (implementing the attributes to solve the problems) could indicate, for example, whether the students have difficulty in choosing a strategy or in implementing a skill. Thus, skill-oriented or strategy-oriented instructions could be designed accordingly to meet students' needs.

Regarding the limitations of previous studies, this study aims at proposing a longitudinal CDM that makes three contributions. First, the proposed model is designed to reduce the risk of Q-matrix misspecification due to differential problem-solving strategies by considering multiple Q-matrices. While studies with similar

purposes have been done in the IRT framework (e.g., Mislevy & Huang, 2007) or single-time-point CDMs (e.g., de la Torre & Douglas, 2008), this study addresses the purpose in the longitudinal CDMs. Second, the proposed model is intended to measure strategy shift in the CDM framework, which could serve as a measure for the effectiveness of some strategy-oriented intervention programs. Finally, the proposed model aims at providing more informative diagnostic information: In addition to informing students' strengths and weaknesses in terms of their skill implementation, the model provides information on students' strategy choice. This diagnostic information could potentially inform whether strategy-oriented or skill-oriented instructions are needed.

Chapter 3: Methods

3.1 The Proposed Model

This study proposes a longitudinal CDM that takes into account both between-person multiple strategies and within-person strategy transition. The model, separating the strategy choice process from the skill implementation process, aims at providing more fine-grained diagnostic information compared to traditional CDMs. More specifically, in addition to providing diagnostic information on the attribute mastery status, the model is intended to inform the cognitive process underlying the strategy choice and skill implementation, respectively.

Figure 4 demonstrates the relations among strategy choice, skill implementation, Q-matrices, attribute mastery status and item responses. In this study, the strategy choice is represented with a discrete latent variable, *m*, and different strategy choices distinguished by different Q-matrices. The underlying assumption is that people who choose different strategies attempt to use different sets of attributes to solve the same multiple-approach problem. When a Q-matrix represents a problem-solving strategy, each q-entry, q_{ikm}, can be interpreted as "whether a person who chooses strategy *m* would attempt to use attribute *k* to solve item *i*". Whether the attempt would be successful is determined by the attribute mastery status that affects the skill implementation process. In Figure 4, the value of q_{ikm} determines whether or not the path from attribute α_k to item response Y_i is present. Specifically, $q_{ikm} = 0$ denotes that respondents who choose strategy *m* would not attempt to apply attribute *k* to solve item *i*, thus the path from α_k to Y_i is absent

for respondents who choose strategy *m*; when the path is absent, the mastery of α_k does not contribute to the correct item response probability. When $q_{ikm} = 1$, the path from α_k to Y_i is present for respondents who choose strategy *m*; the mastery of α_k is expected to increase the correct item response probability. A continuous latent variable θ is used to represent the skill implementation ability underlying the mastery statuses of all the attributes. Given that the strategy choice and skill implementation stages are assumed to be independent, the attempt to apply an attribute does not imply that attribute is mastered; the mastery of an attribute does not imply the attempt to apply the attribute.

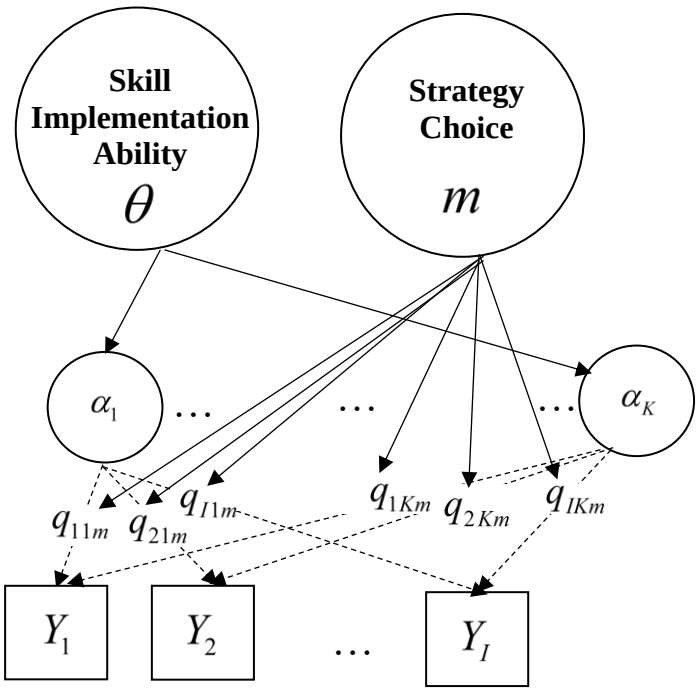


Figure 4. The relations among strategy choice, skill implementation ability, Q-matrices, attribute mastery status and item responses. The strategy choice is represented with a discrete latent variable *m*; the skill implementation ability is represented with a continuous latent variable θ. The dashed lines indicate that the paths may or may not be present, depending on the values of *q*.

As one of the goals of the proposed model is to quantify the growth of the ability of skill implementation, the model is based on growth modeling and a higher-order structure is specified to underlie the attributes as shown in Figure 4. The continuous ability parameter θ in the higher-order structure render it feasible to quantify the growth on a continuous scale. The LLM is used as the measurement model as it is flexible to incorporate the unique contribution of each attribute to the correct response probability. Nevertheless, as the LLM can be easily generalized to the LCDM – one of the most widely-used generalized CDMs – by incorporating the attribute interaction terms, the proposed model is presented in the form of the LCDM for the convenience of future generalization. Following the model specification of the mixture GDM (von Davier, 2007), this study assumes that the population consists of subpopulations who choose different strategies (i.e., between-person multiple strategies) and the Q-matrices are different for different subpopulations. In a mixture CDM (MCDM) with a higher-order attribute structure (equation 1) and with the LCDM as the measurement model (equation 2), the marginal probability of the observed item responses is written as

$$P(\mathbf{Y_j} = \mathbf{y_j}) = \sum_{m=1}^{M} \pi_m \int \left[\prod_{k=1}^{K} P(\alpha_{jk} = 1 \mid \theta, m) \prod_{i=1}^{I} P(Y_{ij} = y_{ij} \mid \boldsymbol{\alpha}_\mathbf{c}, \theta, m) \right] f(\theta \mid m) d\theta,$$

where π_m is the mixing proportion of strategy *m*. The conditional item response probability given strategy *m* is:

$$P(Y_{ij}=1|\boldsymbol{\alpha}_{\mathbf{c(j)}},\theta_j,m)=$$
$$\frac{\exp(\lambda_{im,0}+\sum_{k=1}^{K}\lambda_{im,1,(k)}\alpha_{jk}q_{ikm}+\sum_{k_1<k_2}\lambda_{im,2,(k_1,k_2)}\alpha_{jk_1}\alpha_{jk_2}q_{ik_1m}q_{ik_2m}+...)}{1+\exp(\lambda_{im,0}+\sum_{k=1}^{K}\lambda_{im,1,(k)}\alpha_{jk}q_{ikm}+\sum_{k_1<k_2}\lambda_{im,2,(k_1,k_2)}\alpha_{jk_1}\alpha_{jk_2}q_{ik_1m}q_{ik_2m}+...)}, \quad (4)$$

where q_{ikm} denotes whether individuals who chooses strategy *m* would attempt to use attribute *k* to solve item *i*. $\lambda_{im,0}$ is the item intercept parameter and equals to the logit of the correct response probability when no attributes are attempted and/or mastered. $\lambda_{im,1,(k)}$ and $\lambda_{im,2,(k_1,k_2)}$ are the main and interaction effects of the attributes on the correct response probability of item *i* when the attributes are attempted. When $\lambda_{im,0}$, $\lambda_{im,1,(k)}$ and $\lambda_{im,2,(k_1,k_2)}$ are strategy-specific, another way of interpreting these parameters is that they are the interaction effects of the strategy choice and skill implementation on the correct response probability. Recall that this study assumes that the strategy choice and skill implementation stages are independent and no interaction between strategy choice and implementation is considered for simplicity, $\lambda_{im,0}$, $\lambda_{im,1,(k)}$ and $\lambda_{im,2,(k_1,k_2)}$ are constrained to be equal across strategies, i.e., $\lambda_{im,0}=\lambda_{i,0}$ $\lambda_{im,1,(k)}=\lambda_{i,1,(k)}$, $\lambda_{im,2,(k_1,k_2)}=\lambda_{i,2,(k_1,k_2)}$; thus, $\lambda_{i,1,(k)}$ and $\lambda_{i,2,(k_1,k_2)}$ denote the main and interaction effects of the skill implementation on the correct response probability. When dependencies between the strategy choice and skill implementation are considered, the above equality constraints imposed on $\lambda_{im,0}$, $\lambda_{im,1,(k)}$ and $\lambda_{im,2,(k_1,k_2)}$ can be relaxed. In the LLM, the interaction term, $\lambda_{i,2,(k_1,k_2)}$, is dropped.

Empirical data analyses using the LCDM have yielded two-way attribute interaction estimates of different sizes and directions, but most of these empirical

findings supported the compensatory relationships among the required attributes (e.g., Templin & Hoffman, 2013; Toprak et al., 2019). Templin and Hoffman (2013) indicated that the attribute interactions being zero has suggested a compensatory item response function. Further, Toprak et al. (2019) pointed out that the LCDM with negative attribute interaction terms resembled the DINO model where, when one of the required attributes is mastered, the mastery of additional required attribute will not increase the correct item response probability. Templin and Hoffman (2013) applied the LCDM to the Examination for the Certificate of Proficiency in English (ECPE) dataset and found that, among the 9 items requiring two attributes, 4 items had attribute interaction effects being negative or positive but smaller than 0.1. Templin and Hoffman (2013) also suggested removing the interaction terms when their estimates are close to 0. Toprak et al. (2019) analyzed the Michigan English Test data and found negative attribute interaction estimates for all the four items requiring more than one attributes. In sum, conclusions about whether the attribute interaction terms should be included in the model vary across items and datasets. However, given that it is challenging to accurately recover the interaction term in the LCDM (e.g., Sen & Bradshaw, 2017), the data-generating and data-fitting models in this study are based on the LLM that does not contain the interaction terms in order to reduce estimation difficulty. Constraining the interaction terms at 0 implies the assumption that the non-mastery of a required attribute can be compensated, at least partially, by the mastery of the other required attributes when an item requires more than one attributes (Templin & Hoffman, 2013). Future studies could consider including some two-way attribute interactions and examine how the specification of the interaction

terms affects the estimation accuracy. Nevertheless, three-way interactions or interactions of an even higher order are likely to cause estimation issues, thus are seldom included (e.g., Templin & Hoffman, 2013; Toprak et al., 2019). To ensure model identification, all the attribute main effects in the LLM are constrained to be positive, i.e., $\lambda_{im,1,(k)} > 0$. Such constraint is adapted from the ordering constraints required for the LCDM to ensure model identification and prevent label switching issue (Lao, 2016).

The conditional attribute mastery probability given the general ability and strategy *m* is specified as

$$P(\alpha_{jk} = 1 | \theta_j, m) = \frac{\exp(\xi_{km}\theta_j + \beta_{km})}{1 + \exp(\xi_{km}\theta_j + \beta_{km})}.$$

Given that the strategy choice is assumed to be independent from attribute mastery status, the attribute discrimination and easiness parameters are constrained equal across strategies, i.e., $\xi_{km} = \xi_k$, $\beta_{km} = \beta_k$. Such equality constraints denote the assumption that the relationships between the general skill implementation ability and the attribute mastery statuses are invariant across different strategies. Furthermore, the distribution of the latent ability parameter is independent from the strategy latent class membership, i.e., $f(\theta | m) = f(\theta)$, implying that θ affects the item response probability only through the skill implementation stage and that θ does not affects the strategy choice. Therefore, in this study, θ is interpreted as the ability to implement the strategy. With these equality constraints, the conditional attribute mastery probability is written as

$$P(\alpha_{jk}=1|\theta_j,m)=P(\alpha_{jk}=1|\theta_j)\frac{\exp(\xi_k\theta_j+\beta_k)}{1+\exp(\xi_k\theta_j+\beta_k)}.$$

Incorporating between-person multiple strategies into longitudinal CDMs. When response data from multiple time points are available, growth-model-based longitudinal CDMs (e.g., Lee, 2017; Zhan, Jiao, Liao, et al., 2019) can be used to provide diagnostic information on the attribute mastery status as well as estimate the growth in the latent ability underlying the attributes. This study illustrates the longitudinal models under a repeated-measure pretest-posttest design as it is a simple and widely-used study design for effectiveness studies on education intervention programs (e.g., Bottge et al., 2015) and the data to be used in our empirical data analysis example have been collected with such design. Extensions could be made to accommodate other assessment designs such as parallel forms with anchor items and to scenarios with more than two time points.

Taking into account the between-person multiple strategies in longitudinal CDMs is advantageous as it potentially attenuates the undesirable effect of the Q-matrix misspecification induced by multiple strategies (de la Torre & Douglas, 2008). In particular, the Q-matrix misspecification can jeopardize the classification accuracy and item parameter estimates in CDMs (Rupp & Templin, 2008a). The between-person multiple strategies at each time point can be modelled by incorporating the MCDM into the longitudinal CDMs. A Longitudinal MCDM is yielded by extending the longitudinal DINA model (Zhan, Jiao, Liao, et al., 2019) to the LCDM case and incorporating a mixture-distribution structure. Specifically, in the Longitudinal MCDM, the marginal probability of the response pattern at time *t* is given as

$$P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}}) = \sum_{m=1}^{M} \pi_m \, P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}} \mid m), \tag{5}$$

and, assuming that individuals do not change their problem-solving strategies over time, the conditional probability of response pattern given strategy *m* at time *t* is:

$$P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}} \mid m) = \int \left[\prod_{k=1}^{K} P(\alpha_{jk}^{(t)} = 1 \mid \boldsymbol{\theta}) \prod_{i=1}^{I} P(Y_{ij}^{(t)} = y_{ij}^{(t)} \mid \boldsymbol{\alpha}_{\mathbf{c}}^{(\mathbf{t})}, \boldsymbol{\theta}, m) \right] f(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where is a vector of ability parameters at all the *t* time points, i.e., $\boldsymbol{\theta} = (\theta^{(1)}, ..., \theta^{(t)})'$. The measurement model is specified as

$$P(Y_{ij}^{(t)} = 1 \mid \boldsymbol{\alpha}_{\mathbf{c(j)}}^{(\mathbf{t})}, \boldsymbol{\theta}_{\mathbf{j}}, m) = \frac{\exp(\lambda_{i,0}^{(t)} + \sum_{k=1}^{K} \lambda_{i,1,(k)}^{(t)} \alpha_{jk}^{(t)} q_{ikm} + \sum_{k_1 < k_2} \lambda_{i,2,(k_1,k_2)}^{(t)} \alpha_{jk_1}^{(t)} \alpha_{jk_2}^{(t)} q_{ik_1m} q_{ik_2m} + ...)}{1 + \exp(\lambda_{i,0}^{(t)} + \sum_{k=1}^{K} \lambda_{i,1,(k)}^{(t)} \alpha_{jk}^{(t)} q_{ikm} + \sum_{k_1 < k_2} \lambda_{i,2,(k_1,k_2)}^{(t)} \alpha_{jk_1}^{(t)} \alpha_{jk_2}^{(t)} q_{ik_1m} q_{ik_2m} + ...)}. \tag{6}$$

Given that the item parameters have been assumed to be time-invariant in most existing studies on longitudinal CDMs (Cho et al., 2010; Kaya & Leite, 2017; S. Wang et al., 2018; Zhan, Jiao, Liao, et al., 2019) and that Madison and Bradshaw (2018) did not find evidence suggesting that the violation of measurement invariance due to IPD would diminish the attribute classification accuracy, this study specifies the item parameters to be invariant across time points, i.e., $\lambda_{i,0}^{(t)} = \lambda_{i,0}$, $\lambda_{i,1,(k)}^{(t)} = \lambda_{i,1,(k)}$, $\lambda_{i,2,(k_1,k_2)}^{(t)} = \lambda_{i,2,(k_1,k_2)}$.

The higher-order structure underlying the attributes is parameterized differently from that in Zhan, Jiao, Liao, et al. (2019). While Zhan, Jiao, Liao, et al. (2019) follows the Anderson-type parameterization (Andersen, 1985), specifying that the latent abilities at different time points follow a *T*-dimensional multivariate normal

distribution, this study follows the Embretson-type parameterization (Embretson, 1991) specifying a multivariate normal distribution for the initial ability and ability changes, as shown in Figure 5. Specifically, Embretson (1991) proposed the multidimensional Rasch model for learning and change (MRMLC) where the ability at the t^{th} time point, T_t (t>1), is written as a linear combination of the initial ability and ability changes:

$$\theta_j^{(T_t)} = \theta_j^{(T_1)} + \sum_{t=T_2}^{T_t} \Delta\theta_j^{(t)} , \qquad (7)$$

where $\Delta\theta_j^{(t)}$, referred to as "modifiabilities" by Embretson (1991), represents the change in ability from the (t-1)th time point to the t^{th} time point. The advantage of using the Embretson-type parameterization is that it enables the ability change to be directly estimated and the hypotheses about ability change to be tested. While the interpretation of the parameters are different between the Embretson-type parameterization and the Andersen-type parameterization, the two parameterizations are statistically equivalent: W.-C. Wang et al. (1998) have found that the two parameterizations yielded comparable model-data fit when fitted to an empirical dataset; and W.-C. Wang (2014) has established equations for converting the mean and variance of the ability parameters between the two parameterizations. See von Davier et al. (2011) and W.-C. Wang (2014) for more detailed contrasts between the Anderson-type and Embretson-type parameterizations. Another advantage of utilizing the Embretson-type parameterization is that it overcomes the paradoxical reliability issue, noted by Bereiter (1963), of measuring change using the observed change scores. Specifically, as explained by W.-C. Wang and Wu (2004), the paradoxical

reliability refers to the phenomenon where the higher correlation between the pretest and posttest scores is associated with the lower reliability of the change scores. The paradoxical reliability results from the fact that the measurements at different occasions are incorrectly assumed to be unidimensional, thus this issue is hard to be resolved in the classical test theory (CTT) framework. However, the paradoxical reliability issue can be resolved by formulating the initial ability and ability change as separate latent dimensions in the IRT framework as demonstrated by Embretson (1991).

This study, incorporated a 2PL version of the MRMLC (Embretson, 1997) into the higher-order structure. When only two time points, T_1 and T_2, are involved, the conditional probability of attribute mastery given the latent ability is specified as:

$$P(\alpha_{jk}^{(T_1)}=1|\,\boldsymbol{\theta}_{\mathbf{j}}^{*})=\frac{\exp(\xi_{k,1}^{(T_1)}\theta_j^{(T_1)}+\beta_k^{(T_1)})}{1+\exp(\xi_{k,1}^{(T_1)}\theta_j^{(T_1)}+\beta_k^{(T_1)})},$$

$$P(\alpha_{jk}^{(T_2)}=1|\,\boldsymbol{\theta}_{\mathbf{j}}^{*})=\frac{\exp(\xi_{k,1}^{(T_2)}\theta_j^{(T_1)}+\xi_{k,2}^{(T_2)}\Delta\theta_j+\beta_k^{(T_2)})}{1+\exp(\xi_{k,1}^{(T_2)}\theta_j^{(T_1)}+\xi_{k,2}^{(T_2)}\Delta\theta_j+\beta_k^{(T_2)})},\ \boldsymbol{\theta}_{\mathbf{j}}^{*}=(\theta_j^{(T_1)},\Delta\theta_j)' \quad (8)$$

where $\theta_j^{(T_1)}$ represents the initial ability; $\Delta\theta_j$ represents the ability change from time point T_1 to time point T_2. $\boldsymbol{\theta}_{\mathbf{j}}^{*}=(\theta_j^{(T_1)},\Delta\theta_j)'$ is specified to follow a multivariate normal distribution:

$$\begin{pmatrix}\theta_j^{(T_1)}\\ \Delta\theta_j\end{pmatrix}\sim MVN\left[\begin{pmatrix}\mu_{\theta^{(T_1)}}\\ \mu_{\Delta\theta}\end{pmatrix},\begin{pmatrix}\sigma^2_{\theta^{(T_1)}} & \\ \sigma_{\theta^{(T_1)}\Delta\theta} & \sigma^2_{\Delta\theta}\end{pmatrix}\right].$$

The mean and variance of $\theta_j^{(T_1)}$ are constrained at 0 and 1, respectively, for scale identification. The attribute discrimination and easiness parameters are set to be

invariant across time points, i.e., $\xi_{k,1}^{(T_1)} = \xi_{k,1}^{(T_2)} = \xi_{k,1}$, and $\beta_k^{(T_1)} = \beta_k^{(T_2)} = \beta_k$. When more than two time points are involved, the conditional probability of attribute mastery given the latent ability at the t^{th} time point, T_t (t>1), is specified as

$$P(\alpha_{jk}^{(T_t)} = 1 | \boldsymbol{\theta}_j^*) = \frac{\exp(\xi_{k,1}\theta_j^{(T_1)} + \sum_{t=T_2}^{T_t} \xi_{k,t}\Delta\theta_j^{(t)} + \beta_k)}{1 + \exp(\xi_{k,1}\theta_j^{(T_1)} + \sum_{t=T_2}^{T_t} \xi_{k,t}\Delta\theta_j^{(t)} + \beta_k)},$$

where $\Delta\theta_j^{(t)}$ is the ability change from the (t-1) th time point to the t^{th} time point.

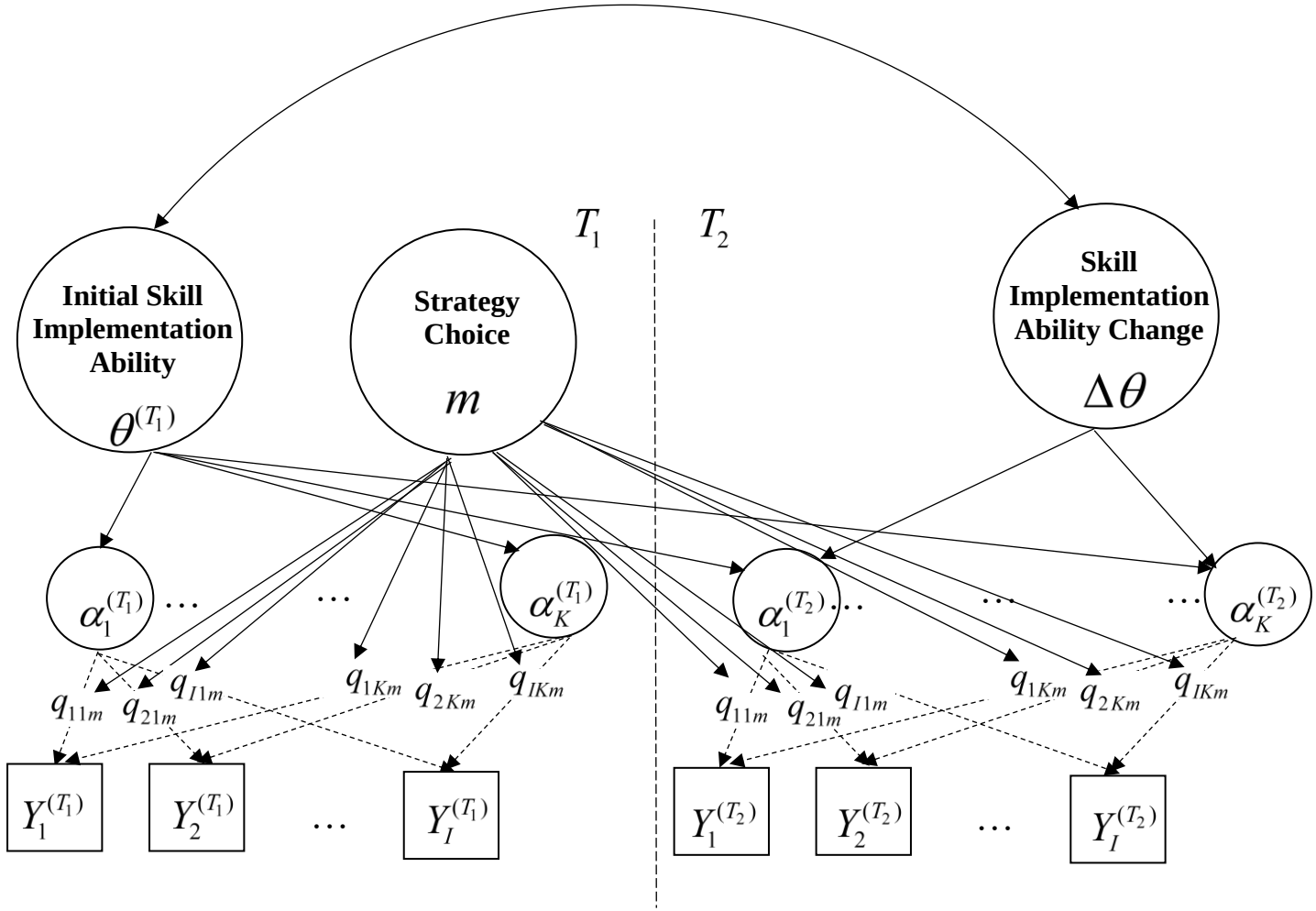


Figure 5. Model structure of the Longitudinal MCDM under a repeated-measure pretest-posttest design.

Measuring within-person strategy shift. A shift in problem-solving strategy could occur over time as a result of learning and education intervention (e.g., Cho et al., 2010; Siegler et al., 1981). From a modeling perspective, a strategy shift can be operationally defined by a shift in the strategy latent class membership, *m*, over time in the Longitudinal MCDM (equation 5), as shown in Figure 6. Thus, the LTA model can be incorporated into the Longitudinal MCDM to model the shift in the strategy choice over time. In the proposed LTA-longitudinal-MCDM, the marginal probability of the response pattern is given as

$$P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}}) = \sum_{m_1=1}^{M_1} \cdots \sum_{m_T=1}^{M_T} \pi_{m_1}^{(1)} \prod_{t'=2}^{T} \tau_{m_{t'}|m_{t'-1}}^{(t'-1)} P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}} \mid \mathbf{m}^*), \qquad (9)$$

where m_t indicates the strategy choice at time point *t*, $t = 1, 2, ..., T$; M_t having the subscript, *t*, implies that the options of strategies are allowed to vary across time points. $\pi_{m_1}^{(1)}$ represents the mixing proportion of respondents who choose strategy m_1 at the first time point. $\tau_{m_{t'}|m_{t'-1}}^{(t'-1)}$ is the latent transition probability from strategy $m_{t'-1}$ to $m_{t'}$ at time point $t'-1$, $t' = 2, ..., T$; in other words, for a respondent who chooses strategy $m_{t'-1}$ at $t'-1$, the probability of this respondent choosing strategy $m_{t'}$ at time point t' is $\tau_{m_{t'}|m_{t'-1}}^{(t'-1)}.\mathbf{m}^*$, represented as a pattern of chosen strategies over time, i.e., $(m_1, m_2, ..., m_T)$, is referred to as the "strategy choice trajectory". In the LTA-longitudinal-MCDM, each latent class is a strategy choice trajectory as opposed to a static strategy choice, and the mixing proportion of a strategy choice trajectory is $\pi_{m_1} \prod_{t'=2}^{T} \tau_{m_t|m_{t-1}}^{(t'-1)}$. The maximum number of strategy choice trajectories is $\prod_{t=1}^{T} M_t$.

The conditional probability of an observed response pattern given a strategy choice trajectory is written as

$$P(\mathbf{Y_j^{(t)}} = \mathbf{y_j^{(t)}} \mid \mathbf{m}^*) = \int \left[\prod_{k=1}^{K} P(\alpha_{jk}^{(t)} = 1 \mid \boldsymbol{\theta}) \prod_{i=1}^{I} P(Y_{ij}^{(t)} = y_{ij}^{(t)} \mid \boldsymbol{\alpha}_{\mathbf{c}}^{(\mathbf{t})}, \boldsymbol{\theta}, \mathbf{m}^*) \right] f(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where the measurement model is the LCDM:

$$P(Y_{ij}^{(t)} = 1 \mid \boldsymbol{\alpha}_{\mathbf{c(j)}}^{(\mathbf{t})}, \boldsymbol{\theta}_{\mathbf{j}}^{*}, \mathbf{m}^*) = \frac{\exp(\lambda_{i,0} + \sum_{k=1}^{K} \lambda_{i,1,(k)} \alpha_{jk}^{(t)} q_{ik\mathbf{m}^*}^{(t)} + \sum_{k_1<k_2} \lambda_{i,2,(k_1,k_2)} \alpha_{jk_1}^{(t)} \alpha_{jk_2}^{(t)} q_{ik_1\mathbf{m}^*}^{(t)} q_{ik_2\mathbf{m}^*}^{(t)} + \ldots)}{1 + \exp(\lambda_{i,0} + \sum_{k=1}^{K} \lambda_{i,1,(k)} \alpha_{jk}^{(t)} q_{ik\mathbf{m}^*}^{(t)} + \sum_{k_1<k_2} \lambda_{i,2,(k_1,k_2)} \alpha_{jk_1}^{(t)} \alpha_{jk_2}^{(t)} q_{ik_1\mathbf{m}^*}^{(t)} q_{ik_2\mathbf{m}^*}^{(t)} + \ldots)}. \quad (10)$$

The measurement model is similar to the one in the Longitudinal MCDM (equation 6) except the subscripts of *q*. Specifically, the q-entries are invariant across time points in the Longitudinal MCDM. In contrast, in the LTA-longitudinal-MCDM, $q_{ik\mathbf{m}^*}^{(t)}$ is specific to strategy choice trajectories; the values of $q_{ik\mathbf{m}^*}^{(t)}$ can vary across time points in the trajectories involving strategy shift. The higher-order attribute structure is specified in a similar way as that of the longitudinal-MCDM (equation 8).

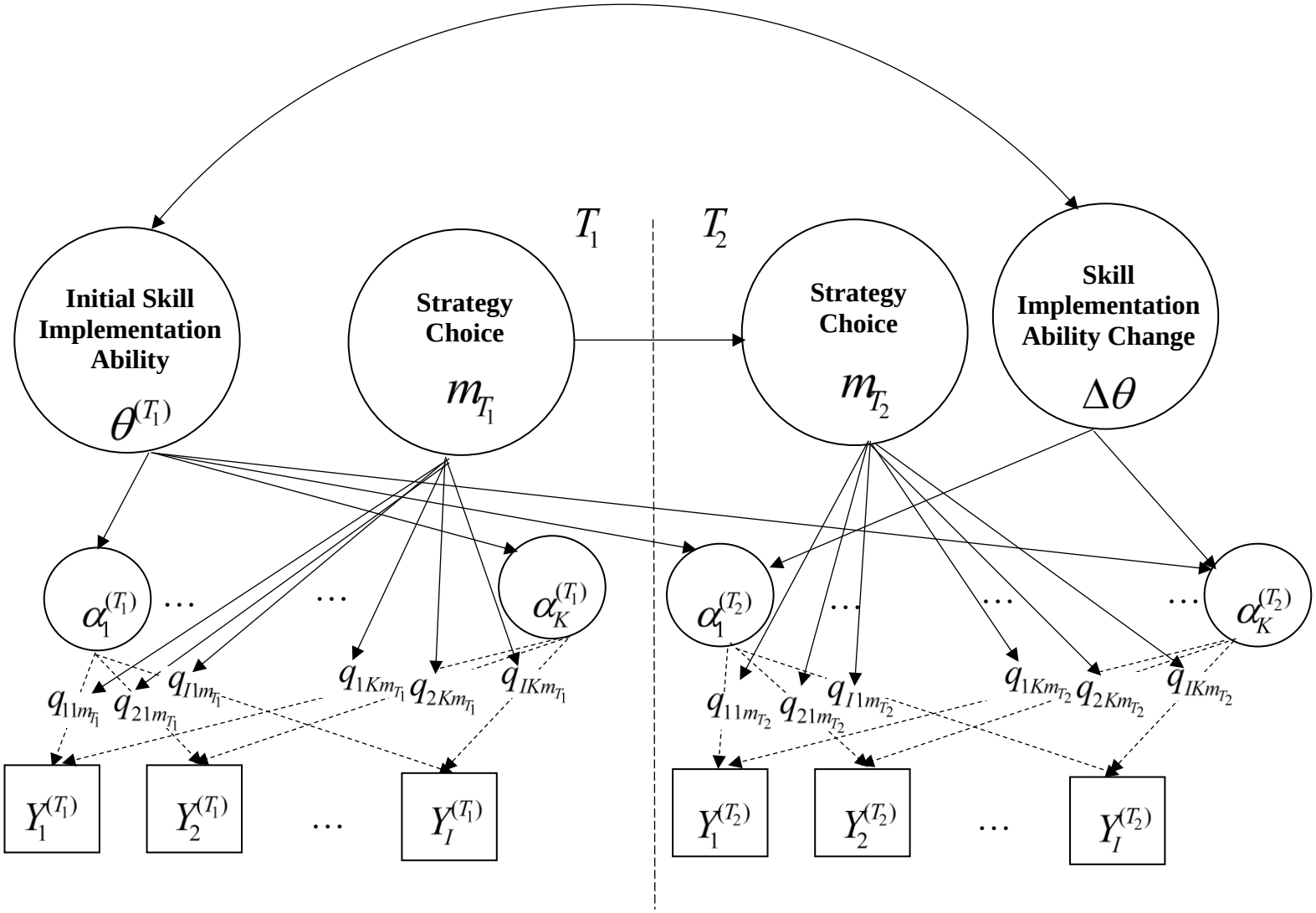


Figure 6. Model structure of the LTA-longitudinal-MCDM under a repeated-measure pretest-posttest design.

3.2 Model Parameter Estimation

Given that a higher-order structure is assumed underlying the attribute profiles and that the Bayesian MCMC estimation method, compared to the MLE, is more suitable for the higher-order structure as introduced in Section 2.6, the model parameters in this study are estimated using the Bayesian MCMC method. The model parameter estimation is implemented in JAGS (Plummer, 2015) with the default Gibbs sampler (Gelfand & Smith, 1990). The JAGS program is called from R 3.5.3 (R Development Core Team, 2013) with the R2jags package v0.5-7 (Su & Yajima, 2015).

Prior specifications. The priors of the model parameters are set based on previous studies that utilized the Bayesian MCMC method for longitudinal model parameter estimation (Kadengye et al., 2013; F. Li et al., 2016; Zhan, Jiao, Man, et al., 2019). As the item parameters and the higher-order structural parameters are specified as time- and strategy-invariant, the prior distributions of these parameters are the same between the single-time-point model (i.e., MCDM) and the multiple-timepoint models (i.e., Longitudinal MCDM and LTA-longitudinal-MCDM):

$$\lambda_{i,0} \sim Normal(-1.096, 4),$$
$$\lambda_{i,1,(k)} \sim Normal(0, 4)T(0, +\infty),$$
$$\beta_k \sim Normal(0, 4),$$
$$\xi_k \sim Normal(0, 4)T(0, +\infty),$$

where “Normal” indicates that a parameter follows a normal distribution; the two parameters specified in the parentheses following “Normal” denote the mean and variance of the distribution, respectively. For instance, the prior distribution of $\lambda_{i,0}$ is

a normal distribution with mean and variance being -1.096 and 4, respectively. "T" indicates that the distribution is truncated and the two elements in the parentheses following "T" denote the lower bound and upper bound of the truncated distribution, respectively. For example, $\lambda_{i,1,(k)} \sim Normal(0,4)T(0,+\infty)$ means that the main effect parameter, $\lambda_{i,1,(k)}$, is constrained to be positive which corresponds to the model's ordering constraints (e.g., Rupp et al., 2010). The prior distributions for the ability and attribute parameters are specified differently between the single-time-point and the multiple-time-point models. In the MCDM, the prior distributions of attributes and the underlying ability parameter are specified as:

$$\alpha_{jk} \mid \theta_j, \beta_k, \xi_k \sim Bernoulli\left(\frac{\exp(\xi_k \theta_j + \beta_k)}{1+\exp(\xi_k \theta_j + \beta_k)}\right),$$

$$\theta_j \sim Normal(0,1).$$

In the Longitudinal MCDM and LTA-longitudinal-MCDM, the prior distributions of attributes are specified as:

$$\alpha_{jk}^{(T_1)} \mid \theta_j^{(T_1)}, \beta_k, \xi_k \sim Bernoulli\left(\frac{\exp(\xi_k \theta_j^{(T_1)} + \beta_k)}{1+\exp(\xi_k \theta_j^{(T_1)} + \beta_k)}\right),$$

$$\alpha_{jk}^{(T_2)} \mid \theta_j^{(T_1)}, \Delta\theta_j, \beta_k, \xi_k \sim Bernoulli\left(\frac{\exp(\xi_k \theta_j^{(T_1)} + \xi_k \Delta\theta_j + \beta_k)}{1+\exp(\xi_k \theta_j^{(T_1)} + \xi_k \Delta\theta_j + \beta_k)}\right),$$

where the multidimensional ability parameters are drawn from a multivariate normal distribution:

$$\begin{pmatrix} \theta_j^{(T_1)} \\ \Delta\theta_j \end{pmatrix} \sim MVN\left[\begin{pmatrix} \mu_{\theta^{(T_1)}} \\ \mu_{\Delta\theta} \end{pmatrix}, \boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)}\right].$$

$\boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)} = \begin{pmatrix} \sigma^2_{\theta^{(T_1)}} & \\ \sigma_{\theta^{(T_1)}\Delta\theta} & \sigma^2_{\Delta\theta} \end{pmatrix}$ is the variance-covariance matrix of the ability parameters.

$\mu_{\theta^{(T_1)}}$ and $\sigma^2_{\theta^{(T_1)}}$ need to be set at 0 and 1, respectively, for scale identification. The mean of the ability change parameter is drawn from a univariate normal distribution, i.e., $\mu_{\Delta\theta} \sim Normal(0, 2)$. Since it is hard to impose the constraint of $\sigma^2_{\theta^{(T_1)}} = 1$ if $\boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)}$ were drawn from the commonly used Wishart distribution (Wishart, 1928), this study utilizes the prior configuration proposed by Azevedo et al. (2016), which has been based on the work of McCulloch et al. (2000), to draw $\boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)}$. Azevedo et al. (2016)'s prior specification for $\boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)}$ was found to be effective in handling the restricted variance-covariance matrix for identification purpose in the longitudinal IRT model and achieved good parameter recovery. Specifically, $\sigma^2_{\Delta\theta}$ in $\boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)}$ is reparameterized conditioning on $\sigma^2_{\theta^{(T_1)}} = 1$, i.e., $\sigma^2_{\Delta\theta} = \sigma^{2\,*}_{\Delta\theta} + \sigma^2_{\theta^{(T_1)}\Delta\theta} / \sigma^2_{\theta^{(T_1)}} = \sigma^{2\,*}_{\Delta\theta} + \sigma^2_{\theta^{(T_1)}\Delta\theta}$, where the priors of $\sigma^{2\,*}_{\Delta\theta}$ and $\sigma_{\theta^{(T_1)}\Delta\theta}$ are $\sigma^{2\,*}_{\Delta\theta} \sim InvGamma(1,1)$ and $\sigma_{\theta^{(T_1)}\Delta\theta} \sim Normal(0,1)$, respectively. According to Azevedo et al. (2016), such reparameterization of $\sigma^2_{\Delta\theta}$ is equivalent for $\Delta\theta_j$ to have a conditional distribution of $\Delta\theta_j \mid \theta_j^{(T_1)} \sim Normal(\mu^*_{\Delta\theta}, \sigma^{2\,*}_{\Delta\theta})$, where $\mu^*_{\Delta\theta} = \mu_{\Delta\theta} + \sigma_{\theta^{(T_1)}\Delta\theta}(\theta_j^{(T_1)} - \mu_{\theta^{(T_1)}})$ and $\sigma^{2\,*}_{\Delta\theta} = \sigma^2_{\Delta\theta} - \sigma^2_{\theta^{(T_1)}\Delta\theta} / \sigma^2_{\theta^{(T_1)}}$.

As for the mixing proportions of strategies, the prior distributions are Dirichlet distributions:

$$\boldsymbol{\pi} \sim Dirichlet(\kappa_1, ..., \kappa_M),$$

where κ_m is set at 1, for $m = 1, 2, ...M$, which satisfies the criteria of a sparse prior (i.e., $\kappa_m < d/2$, where *d* is the number of latent class-specific parameters). Such sparse priors have been found to possess a property: they can make the redundant latent classes empty when a mixture model is overfitted through the MCMC sampling process (Nasserinejad et al., 2017; Rousseau & Mengersen, 2011). The implication of such property is that if the number of strategy options are overspecified (e.g., experts identify two different strategies among the population but the respondents only choose to use one of them), the estimated mixing proportion of the unused strategy will be extremely small when a sparse prior is specified for $\boldsymbol{\pi}$. Thus, the estimates of the mixing proportions may serve as indicators to verify the theory about problem-solving strategies. The prior distribution of the latent transition probability in the LTA-longitudinal-CDM is specified as

$$\boldsymbol{\tau}_{m_t|m_{t-1}}^{(t-1)} = (\tau_{1|m_{t-1}}^{(t-1)}, ..., \tau_{M_t|m_{t-1}}^{(t-1)}) \sim Dirichlet(\omega_1, ..., \omega_{M_t}),$$

where $\omega_{m_{t-1}} = 1,$ for $m_{t-1} = 1, 2, ..., M_{t-1}$.

The number of MCMC chains, iterations and convergence check. Two MCMC chains are run. To ensure that the sampled iterations adequately represent the posterior distributions of interest, the convergence of the iterative parameter draws from the two MCMC chains is evaluated by inspecting the trace plots of the MCMC draws and calculating the potential scale reduction factor (PSRF; Gelman & Rubin, 1992). A trace plot is a time series plot that displays the parameter draws at each iteration of the MCMC chains; traces of draws from different chains are often displayed in different colors. Thus, one can observe the mixing of the MCMC chains in the trace plot and the trace plot can serve as a graphical diagnostic of convergence.

In particular, if the location and spread of the traces are stable and the traces of different chains converge to the same location, it will be a piece of evidence for convergence; if different parts of the traces are stuck around different locations, a lack of convergence is suggested. The PSRF, also denoted as $\hat{R}$, is the ratio of the estimated pooled posterior variance of the MCMC draws (i.e., a weighted mean of the estimated between-chain and within-chain variances of the MCMC draws) to the estimated within-chain variance of the MCMC draws. An $\hat{R}$ close to 1 indicates that the inferences drawn from different chains are close to each other, which is a sign of convergence; by contrast, an $\hat{R}$ much greater than 1 suggests the lack of convergence. Brooks and Gelman (1998) and Gelman and Rubin (1992) suggested using the criterion of $\hat{R} < 1.2$ for all the model parameters to determine the MCMC convergence; this study applies a more stringent and commonly-used criterion in practice, $\hat{R} < 1.1$.

Five thousand iterations are run for each MCMC chain, including 2,500 burn-in iterations. The chains are thinned by 2 iterations to reduce the autocorrelation of the draws before summarizing the parameter estimates. As a result, each parameter estimate is summarized based on a total of 2,500[2] iterations. The chain length is determined based on a pilot study where one replication was run for every simulated condition and the convergence has been achieved with 5,000 iterations each chain including 2,500 burn-in and a thinning of 2 for all the three data-fitting models under

[2] In JAGS, the number of iterations retained in each chain=(the total number of iterations of each chain – the number of burn-in iterations)/the number of thinning. Thus, the total number of iterations used to summarize each parameter estimate is calculated as: [(5000 total – 2500 burn-in)/2 thinning]*2 chains=2500

all the conditions. Since it is possible that, the number of MCMC iterations required for convergence varies across simulation study replications, $\hat{R}$ is monitored for all the parameters in all the replications in the full study to ensure model convergence. If one or more parameters have $\hat{R}$ greater than 1.1 after the first 5,000 iterations, additional iterations will be run until the $\hat{R}$ convergence criterion is met before summarizing the parameter estimates.

3.3 Simulation Study Design

A Monte Carlo simulation study is conducted to examine the parameter recovery of the LTA-longitudinal-MCDM under several simulated conditions and investigate the effects of ignoring multiple strategies and strategy shift on the classification accuracy and growth estimate.

Due to the complex nature of human cognition and problem-solving strategies, this study simulates a simplified scenario of an education intervention in order to make the simulation study manageable. The goals and effects of the hypothesized education intervention are specified in light of the Enhanced Anchored Instruction (EAI; Bottge, 2001) which aims at improving students' performance on problem-solving. Some instructions designed in EAI are strategy-oriented in the sense that they guide students to first solve a problem in the multimedia and then to generalize the problem-solving methods to relevant hands-on problems (Hickey et al., 2001). An effectiveness study has found that students who were under the EAI condition tend to perform better on problems requiring complex skills than those who were not (Bottge et al., 2007). Thus, this study simulates an intervention that consists of both skill-oriented and strategy-oriented instructions. The intervention is designed

to improve students' problem-solving performance by guiding them to choose a more complex strategy as well as enhancing their skill implementation ability. In particular, the more complex strategy tends to involve more difficult and more various skills to solve a problem. In the simulated scenario, two strategies are assumed, labelled as "Strategy A" and "Strategy B": Strategy A is the simpler strategy that involves easier and fewer skills to solve the problems; Strategy B is the more complex strategy that involves more difficult and more various skills. It should be emphasized that, in this study, the strategy choice is independent from the skill implementation ability or the attribute mastery status. That is, choosing to use the simpler strategy does not imply the lower ability to implement the skills required by the more complex strategy and vice versa. Furthermore, the strategy choice and skill implementation process are assumed to independently contribute to the correct response probability as shown in Figure 2.

Thus, to operationally define the desired effects of the hypothesized intervention, compared to the pretest, a larger proportion of students are expected to choose Strategy B in the posttest; a growth is expected to be in students' average skill implementation ability and the probabilities of attribute mastery.

3.3.1 Fixed factors

A repeated-measure pretest-posttest (i.e., two time points) study design is simulated, the configuration of which reflects the design of the EAI effectiveness study (Bottge et al., 2015). At each time point, two groups of simulees who choose either Strategy A or Strategy B are simulated. After specifying the group membership, ability parameters, attribute profiles and item parameters, the response

data of each group at each time point are generated using the LLM (i.e., equation 2 without the attribute interaction terms). To make the simulation study manageable, factors that have been studied by previous longitudinal modeling studies (e.g., Cho et al., 2010; Lee, 2017; Madison & Bradshaw, 2018; Zhan, Jiao, Liao, et al., 2019) or are not expected to affect the parameter recovery performance are fixed across conditions, as listed in Table 3. The fixed factors include item intercepts, main effects, attribute easiness and discrimination parameters, the Q-matrix design, the types of strategy choice trajectories and the distribution of latent abilities and attribute profiles.

Table 3
Fixed Factors in the Simulation Study

Factor	Value
Test length (I)	20
The number of attributes (K)	4
The number of time points (T)	2
Strategy types	Strategy A (M_A), Strategy B (M_B)
Strategy choice trajectory types	$M^*_{AA} = (M_A, M_A)$, $M^*_{AB} = (M_A, M_B)$, $M^*_{BB} = (M_B, M_B)$
Data generating Q-matrices	$\mathbf{Q_A}$, $\mathbf{Q_B}$
The proportion of simulees who switch from Strategy B to Strategy A ($p_{M_A \mid M_B}$)	$p_{M_A \mid M_B} = 0$
Item intercept parameter ($\lambda_{i,0}$)	$\lambda_{i,0} = -2.2$ (It corresponds to a correct item response probability of 0.1 when no attribute is mastered.)
Attribute interaction parameter ($\lambda_{i,2,(k_1,k_2)}$)	$\lambda_{i,2,(k_1,k_2)} = 0$
Attribute easiness parameter (β_k)	$\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_3 = -0.5$, $\beta_4 = -1$
Attribute discrimination parameter ($\xi_{k,1}$, $\xi_{k,2}$)	$\xi_{k,1} = \xi_{k,2} = 1$
Mean and variance of the initial ability ($\theta^{(T_1)}$)	$\mu_{\theta^{(T_1)}} = 0$, $\sigma^2_{\theta^{(T_1)}} = 1$
Mean and variance of the ability change ($\Delta\theta$)	$\mu_{\Delta\theta} = 0.5$, $\sigma^2_{\Delta\theta} = 1$

Item intercept parameters and higher-order structural parameters. Each test contains 20 items measuring 4 attributes. The first ten items are single-approach items while the others are multiple-approach items. To simplify the relations between the attributes and item response probabilities and render the data generation feasible, each item-solving approach requires no more than two attributes (i.e., the maximum number of required attributes in a q-vector is 2) and no interaction effect is assumed among the attributes on the correct response probabilities (i.e., $\lambda_{i,2,(k_1,k_2)} = 0$). The true

item intercepts are set at -2.2, which corresponds to a correct item response probability of 0.1 when no attribute is mastered. Among the four attributes, the first two attributes, α_1 and α_2, are relatively easy to master; the last two attributes, α_3 and α_4, are difficult to master. The easiness of the attributes is controlled by the attribute easiness parameter, β_k. The true values of β_k are set at 1, 0.5, -0.5 and -1 for $k = 1, 2, 3, 4$; these parameter values correspond to the attribute mastery probabilities of 0.73, 0.62, 0.38 and 0.27, respectively, when the latent ability, θ, is zero. The attribute discrimination parameters, $\xi_{k,1}$ and $\xi_{k,2}$, are set at 1, indicating that if curves of attribute mastery probability are drawn as a function of θ, these curves would not intersect with each other; the order of the attribute mastery probabilities remain uniform across different latent ability levels.

The Q-matrix design and attribute main effect parameters. Two Q-matrices, $\mathbf{Q_A}$ and $\mathbf{Q_B}$, are designed for Strategy A and Strategy B, respectively. As shown in Table 4, $\mathbf{Q_A}$ and $\mathbf{Q_B}$ only differ in the q-vectors of the multiple-approach items (i.e., Items 11-20). The q-vectors of the single-approach items are designed to ensure the completeness of the Q-matrix (Chiu et al., 2009), which is relevant to ensure the identifiability of the CDMs. For the multiple-approach items, Strategy A and Strategy B involve different sets of attributes (i.e., $\mathbf{Q_A}$ only involves α_1 and α_2, whereas $\mathbf{Q_B}$ only involves α_2, α_3 and α_4), which are designed based on the findings about multiple strategies involved in math multiple-approach items (Tatsuoka, 1987) and the Q-matrices designed for these items (Mislevy, 1996).

Based on the findings that students tended to perform better on items involving complex skills after the EAI treatment (Bottge et al., 2007) and that students tend to make fewer mistakes in reading tasks after some strategy-oriented instructions (Afflerbach et al., 2008), the conditional correct response probabilities of the multiple-approach items given the successful application of the more complex strategy (i.e., Strategy B) are set to be higher than that of the simpler strategy (i.e., Strategy A). To control the effects of strategies on the conditional correct item response probabilities, the true main effect parameters, $\lambda_{i,1,(k)}$, are set as Table 5; the q-vectors of each multiple-approach item are designed to satisfy one or both of the following rules: 1) more difficult attributes are required by Strategy B than Strategy A; 2) more attributes are needed to solve the item using Strategy B than Strategy A. Specifically, as shown in Table 4, Items 11-16, satisfying the first rule, can be solved with the same number of attributes using both strategies but require more difficult attribute(s) under Strategy B. The correct item response probabilities given the successful application of Strategy A and Strategy B are 0.8 and 0.9, respectively, for Items 11-16. Items 17-20, satisfying both rules, are solved by more difficult and more various attributes when using Strategy B than Strategy A. The correct item response probabilities given the successful application of Strategy A and Strategy B are 0.8 and 0.95, respectively, for Items 17-20. While the conditional correct response probabilities given the successful application of Strategy B are higher than those of Strategy A, it is more difficult to implement skills required by Strategy B than Strategy A. As shown in Table 5, the probabilities of mastering all the required attributes of Strategy B are lower than those of Strategy A when the latent ability is

zero. Such relations remain the same for other levels of ability given that the attribute discrimination parameters are uniform across attributes.

Table 4
Q-matrices for Data Generation

Item Type	Item	Strategy A ($\mathbf{Q_A}$)				Strategy B ($\mathbf{Q_B}$)			
		α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
Single-approach items	1	1				1			
	2	1				1			
	3		1				1		
	4		1				1		
	5			1				1	
	6			1				1	
	7				1				1
	8				1				1
	9	1				1			
	10		1				1		
Multiple-approach items	11	1						1	
	12	1						1	
	13		1						1
	14		1						1
	15	1	1				1		1
	16	1	1					1	1
	17	1						1	1
	18	1						1	1
	19		1				1		1
	20		1				1		1

Note. The "0" entries are omitted from the Q-matrices.

Table 5

Main Parameters of Multiple-Approach Items for Data Generation, Conditional Item Correct Response Probability Given Successful Strategy Application and Skill Implementation Difficulty

Item	Main effect parameters ($\lambda_{i,1,(k)}$)				Conditional probability of correct response given the successful strategy application		Probability of simulees with $\theta = 0$ mastering all the required attributes of a strategy	
	α_1	α_2	α_3	α_4	Strategy A	Strategy B	Strategy A	Strategy B
11	3.6		4.4		0.8	0.9	0.73	0.38
12	3.6		4.4		0.8	0.9	0.73	0.38
13		4.4		3.6	0.8	0.9	0.62	0.27
14		4.4		3.6	0.8	0.9	0.62	0.27
15	1.8	1.8		2.6	0.8	0.9	0.45	0.17
16	1.8	1.8	2.2	2.2	0.8	0.9	0.45	0.10
17	3.6		2.6	2.6	0.8	0.95	0.73	0.10
18	3.6		2.6	2.6	0.8	0.95	0.73	0.10
19		3.6		1.6	0.8	0.95	0.62	0.17
20		3.6		1.6	0.8	0.95	0.62	0.17

Note. A blank entry in the main effect parameters indicates that an attribute does not affect the correct item response probability in Strategy A or Strategy B as, based on the Q-matrices, the attribute is not required to solve the item by either strategy.

The types of strategy choice trajectories. An assumption made about the strategy shift is that respondents who choose the more complex strategy (i.e., Strategy B) at the first time point would not shift to the simpler strategy (i.e., Strategy A) at the second time point (i.e., $p_{M_A|M_B} = 0$ and $p_{M_B|M_B} = 1$). As a results, there are three unique strategy choice trajectories, including consistently choosing Strategy A ($\mathbf{M}^*_{\mathbf{AA}}$), consistently choosing Strategy B ($\mathbf{M}^*_{\mathbf{BB}}$) and shifting from Strategy A to Strategy B ($\mathbf{M}^*_{\mathbf{AB}}$). The underlying assumption about the strategy choice made by this study is that, for the simulees who have acquired both strategies, they would rationally choose the strategy that corresponds to higher correct response probabilities (i.e., Strategy B) over the other one (i.e., Strategy A). While not simulated in this study, scenarios

where simulees' strategy choices are affected by other factors, such as the mastery statuses of the skills required by the strategies, could be considered in future studies.

The distribution of latent abilities and attribute profiles. The distribution of the underlying abilities (i.e., initial ability and ability change parameters) is specified to follow a multivariate normal distribution. The means of the initial ability ($\theta_j^{(T_1)}$) and ability change ($\Delta\theta$) are set at 0 and 0.5, respectively, which correspond to attribute mastery probability changes of 0.09, 0.11, 0.12 and 0.11 of the four attributes, respectively. These changes in attribute mastery probabilities fall within the range of attribute mastery probability changes from the pretest to posttest reported by Madison and Bradshaw (2018) based on the data from the EAI effectiveness study (Bottge et al., 2015). The variances of the initial ability and ability change parameters are set at 1 based on the empirical analysis results that the variance estimates of the two parameters are close to each other (Embretson, 1991).

3.3.2 Manipulated factors

Four factors, the sample size (small, medium), the mixing proportions of strategies at the first time point (balanced, imbalanced), the proportions of simulees shifting from Strategy A to Strategy B (low, high) and the correlation between the initial ability and ability change (negative, none, positive) are manipulated and fully crossed, yielding a total of 2*2*2*3=24 conditions. The values set at each level of the manipulated factors are listed in Table 6. The manipulated factors and their values are chosen in light of the simulation study design and empirical data analysis findings from previous literature on longitudinal CDMs or problem-solving strategies (e.g., Bottge et al., 2015; Cho et al., 2010; Zhan, Jiao, Liao, et al., 2019).

Sample size. The sample size is manipulated to examine the effect of sample size on the parameter recovery. The parameter recovery under extremely small sample size is of special interest. While Lee (2017) and Zhan, Jiao, Liao, et al. (2019) have found in simulation studies that a smaller sample size is associated with less accurate model parameter estimates by manipulating the sample size at 200, 500 and/or 1000 simulees, no research has been done to study the parameter recovery under an extremely small sample size (i.e., smaller than 200). In fact, extremely small sample sizes (i.e., around 100) have been observed in empirical data analyses especially in those involve longitudinal data (e.g., Cho et al., 2010; Li et al., 2016). Thus, in this simulation study, two levels of sample size, 100 and 800, are used to represent the extremely small and medium levels of sample size. Sample sizes larger than 1000 are not considered in this simulation study as large sample sizes are rarely observed in the longitudinal diagnostic assessments (Bottge et al., 2015; F. Li et al., 2016; Zhan, Jiao, Liao, et al., 2019).

Initial mixing proportions of strategies. The mixing proportions of $\pi_A^{(1)} : \pi_B^{(1)}$ = 0.8:0.2 and 0.6:0.4 are used to represent the imbalanced and balanced mix of Strategy A and Strategy B at the first time point. The two levels of initial mixing proportions are chosen based on either theoretical assumptions or empirical observations. It is intuitive that the majority of the students who are enrolled into an education intervention program would choose to use the simpler strategy before the intervention. Cho et al. (2010) specified that the majority (i.e., with a proportion of around 0.8) of the simulees belong to the "low-ability" latent class, as opposed to the "high-ability" class, at the initial time point when they simulated the response data of

an education intervention program. However, the empirical data analysis results show a more balanced (i.e., around 0.6:0.4) mix of latent classes at the initial time point (Cho et al., 2010). Therefore, this study considers conditions with different mixing proportions of strategies at the initial time point.

Strategy transition probability. The high (i.e., 0.7) and low (i.e., 0.3) proportions of the simulees shifting from Strategy A to Strategy B are used to mimic the strategy transition with and without the strategy-oriented instructions. These values are set based on the finding from an effectiveness study of the EAI involving two pretests (i.e., pretest 1 and pretest 2) and a posttest (Cho et al., 2010). Cho et al. (2010) found that a larger proportion of students transitioning from latent class 1 to latent class 2 after the pretest 2 (around 0.82) than pretest 1 (around 0.45), implying that the implementation of the EAI induce a significant strategy shift. However, in Cho et al. (2010), latent classes are distinguished by different item parameters and different latent ability distributions, implying that an individual's transition of latent class membership over time could be a result of a strategy shift characterized by differential item parameters as well as a memory effect characterized by differential ability distributions. Therefore, it is expected that the proportions of individuals with strategy shift over time is lower than the latent class transition probabilities observed in Cho et al. (2010).

Correlation between the initial ability and ability change. Negative ($\rho_{\theta^{(T_1)}\Delta\theta}$ =-0.3), none ($\rho_{\theta^{(T_1)}\Delta\theta}$ =0) and positive ($\rho_{\theta^{(T_1)}\Delta\theta}$ =0.3) correlations are used to reflect different intrinsic relations between the initial ability and ability change. Although negative correlations between the initial ability and ability change are

reported in empirical studies (e.g., Alder et al., 1990), the observed correlation values should be taken with caution: The observed correlations between the initial ability and ability change are jointly affected by the measurement error and the intrinsic relations between the initial ability and ability change (Allison, 1990); it has been found that the measurement error can result in a negatively biased correlation estimate between the initial ability and ability change (e.g., Alder et al., 1990). Positive correlations between the initial ability and ability change have also been observed in a few studies (e.g., Thorndike, 1966). In fact, the sign of the observed correlations can vary when different time points are chosen as the initial time point and there is not a consistent rule for choosing the initial time point (Willett, 1997). The intrinsic association between the initial ability and the ability change is affected by a variety of factors such as the nature of the test and the social process (Kelly & Ye, 2017) and no consensus has been reached on the direction of the association. Therefore, different directions of the intrinsic association between the initial ability and ability change are considered in this simulation study.

The range of the true correlation between the initial ability and ability change (i.e., from -0.3 to 0.3) was chosen such that it falls within the range of the correlations between the initial ability and ability changes observed from the empirical data analyses (W.-C. Wang et al., 1998; W.-C. Wang & Wu, 2004). Correlations of -0.3, 0 and 0.3 between the ability and ability change correspond to correlations of 0.59, 0.71 and 0.81 between the abilities at the two timepoints, respectively, which are derived based on the true variances set for the initial ability and ability change parameters. In

other words, in the simulated conditions, the correlations between the abilities at the two timepoints range from medium to large.

Table 6
Manipulated Factors in Simulation Study

Factor	Levels
Sample size (J)	Small: 100; Medium: 800
The mixing proportions of strategies at the first time point ($\pi_{m_1}^{(1)}$)	Balanced: $\pi_{M_A}^{(1)} = 0.6$, $\pi_{M_B}^{(1)} = 0.4$; Imbalanced: $\pi_{M_A}^{(1)} = 0.8$, $\pi_{M_B}^{(1)} = 0.2$
The proportion of simulees shifting from Strategy A to Strategy B ($p_{M_B \mid M_A}$)	Low: 0.3; High: 0.7
The correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$)	Negative: -0.3 ($\rho_{\theta^{(T_1)}\theta^{(T_2)}} = 0.59$); None: 0 ($\rho_{\theta^{(T_1)}\theta^{(T_2)}} = 0.71$); Positive: 0.3 ($\rho_{\theta^{(T_1)}\theta^{(T_2)}} = 0.81$).

3.3.3 Data generating procedure

The data are simulated following the steps below:

1) Simulate the true item parameters and higher-order structural parameters as specified in Tables 3 and 5. Specifically, the item parameters include the item intercept ($\lambda_{i,0} = -2.2$) and main effect whose values are listed in Table 5. The higher-order structural parameters include the attribute easiness parameters ($\beta_1 = 1$, $\beta_2 = 0.5$, $\beta_3 = -0.5$, $\beta_4 = -1$) and the attribute discrimination parameters ($\xi_{k,1} = \xi_{k,2} = 1$).

2) Simulate the skill implementation ability parameters, $\boldsymbol{\theta}$. The initial ability ($\theta_j^{(T_1)}$) and ability change ($\Delta\theta$) parameters are simulated from a multivariate normal distribution, the mean vector of which

is set as $\begin{pmatrix} \mu_{\theta^{(T_1)}} \\ \mu_{\Delta\theta} \end{pmatrix} = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$ and the variance-covariance matrix of

which is $\boldsymbol{\Sigma}_{(\theta^*)} = \begin{pmatrix} 1 & \\ \sigma_{\theta^{(T_1)}\Delta\theta} & 1 \end{pmatrix}$, where $\sigma_{\theta^{(T_1)}\Delta\theta}$=-0.3, 0 or 0.3

depending on the simulated condition specification. Note that the variance-covariance matrix specified above refers to the empirical variance-covariance matrix in order to avoid the potential non-positive-definite variance-covariance matrix issue.

3) Simulate the attribute mastery status parameters, $\boldsymbol{\alpha}$. The attribute mastery status parameters are simulated from the Bernoulli distributions whose the probability parameters are simulated from equation 1 by plugging in the higher-order structural parameters generated in step 1) and the skill implementation ability parameters generated in step 2).

4) Simulate the strategy choice membership parameters, $\mathbf{m}^*$. At the initial time point, the simulees are randomly assigned to Strategy A or Strategy B latent class with the constraint that the empirical proportions of simulees choosing Strategy A and Strategy B at the initial time points are $\pi_A^{(1)}$ and $\pi_B^{(1)}$, respectively ($\pi_A^{(1)} : \pi_B^{(1)}$ = 0.8:0.2 or 0.6:0.4 depending on the simulated condition specification). As for at the second time point, a portion of the simulees who are assigned to Strategy A latent class at the initial time point are assigned to Strategy B latent class at the second time point, the

proportion of whom is either 0.7 or 0.3 depending on the simulated condition specification. The strategy choice latent classes of the remaining simulees remain unchanged over time.

5) Simulate the response data using equation 10 by plugging in the item parameters, higher-order structural parameters and person parameters generated in steps 1) to 4) and the Q-matrices specified in Table 4.

As 30 replications are run in this simulation study, the data generating steps above are repeated and performed once for each replication. As a result, a dataset containing responses of either 100 or 800 simulees to 20 items is simulated for each replication.

3.3.4 Data-fitting models

Three models are fitted to the simulated data, including the LTA-longitudinal-MCDM, Longitudinal MCDM and Longitudinal LLM. The LTA-longitudinal-MCDM takes into account both between-person multiple strategies and within-person strategy shift while the other two models ignore one or both of the multiple-strategy scenarios. In particular, the Longitudinal MCDM (equation 5), without a latent transition probability parameter, constrains the strategy to be the same overtime and thus ignores the within-person strategy shift. In both the LTA-longitudinal-MCDM and Longitudinal MCDM, the Q-matrices are correctly specified. Although the Q-matrix misspecification is an important issue to investigate in CDM studies, it is not considered in this simulation study as the focus of the study is on parameter recovery and model misspecification.

The Longitudinal LLM is an extension of longitudinal DINA model (Zhan, Jiao, Liao, et al., 2019) to the LLM with the Embretson-type growth parameterization (equation 7), which ignores both between-person multiple strategies and within-person strategy shift. Only one of the Q-matrices can be used in the Longitudinal LLM. Given that the education intervention programs are designed for the students who use the simpler strategies, the Strategy A Q-matrix, Q_A, is used in the Longitudinal LLM. In all the three data-fitting models, the mean and variance of the initial ability parameter are set at 0 and 1, respectively, for scale identification.

3.3.5 Outcome measures and analysis procedure

Outcome measures and statistical analyses are chosen to address the following three research questions:

1) How do the relative model fit indices perform in identifying the proposed model as the best-fitting model in the presence of both between-person multiple strategies and within-person strategy shift?
2) What is the impact of ignoring between-person multiple strategies and/or within-person strategy shift on the recovery of the model parameters of the longitudinal CDMs?
3) How is the parameter recovery of the proposed model affected by the manipulated factors?

Before addressing the research questions, the posterior predictive model check (Guttman, 1967; Rubin, 1981, 1984) is conducted to evaluate the absolute model-data fit of the LTA-longitudinal-MCDM. Although, theoretically speaking, the LTA-longitudinal-MCDM would fit all the simulated datasets adequately given that it is the

true data-generating model, the posterior predictive p-value (PPP) is assessed as the absolute fit index in each replication to empirically confirm that each simulated dataset possesses the expected characteristics. In this study, the PPP is determined based on the distribution of a discrepancy measure between the data and the model – the sum of squares of standardized residuals – rather than classical test statistics. The major difference between a classical test statistic and a discrepancy measure is that the former is only dependent on the data while the latter is dependent on both the data and unknown model parameters; the latter is more aligned with the Bayesian formulation (Gelman et al., 1996). Specifically, the following procedure is implemented for posterior predictive model check:

i) Simulate item responses (i.e., replicated data, y_{rep}) are generated using the sampled model parameters in each MCMC iteration;

ii) Calculate the discrepancy measures (i.e., $D(y_{rep};\Theta)$ and $D(y;\Theta)$) using the replicated data (y_{rep}) and observed data (y), respectively, i.e.,

$$D(y_{rep};\Theta)=\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{y_{rep,ij}^{(t)}-P_{ij}^{(t)}}{\sqrt{P_{ij}^{(t)}(1-P_{ij}^{(t)})}}\right)^2,$$

$$D(y;\Theta)=\sum_{t=1}^{T}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{y_{ij}^{(t)}-P_{ij}^{(t)}}{\sqrt{P_{ij}^{(t)}(1-P_{ij}^{(t)})}}\right)^2,$$

where Θ represents all the model parameters, and $P_{ij}^{(t)}=P(y_{ij}^{(t)}=1|\Theta)$.

iii) Calculate the PPP value, i.e., $p_b(y)=P[D(y_{rep};\Theta)\geq D(y;\Theta)|L,y]$, where L is the proposed model.

The PPP value in this study denotes the proportion of $D(y_{rep};\Theta)$ s that are greater than $D(y;\Theta)$. According to Gelman et al. (2003), a PPP value extremely close to 0 or 1 indicates a misfit between the model and the observed data. However, given that the sum of squares of standardized residuals is used as the discrepancy measure in this study, only extremely small PPP indicates a rejection of the data-fitting model. Given that no specific suggestion on the PPP cut-off value was found, this study rejects a model when the PPP value is lower than 0.05, as 0.05<PPP<0.95 was mentioned by Gelman et al. (2003) as a reasonable range to accept a model. If the LTA-longitudinal-MCDM is not rejected by the simulated data (i.e., having extremely small PPP value), one can be more confident that the simulated datasets possess the desired characteristics of the true data-generating model and can move on to address the research questions.

Assessing the performance of the model fit indices. To address the first research question, the performance of three relative model fit indices, including Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978) and deviance information criterion (DIC; Spiegelhalter et al., 2002), in correctly identifying the LTA-longitudinal-MCDM as the best-fitting model in the presence of both between-person multiple strategies and within-person strategy shift is investigated. The performance of a model fit index is operationally defined as the frequency (i.e., the number of replications) that the model fit index correctly selects the LTA-longitudinal-MCDM as the best-fitting model (i.e., the LTA-longitudinal-MCDM has the lowest model fit index among the three data-fitting models).

AIC and BIC have originally been designed for the maximum likelihood estimation. According to Congdon (2003), when used in the Bayesian MCMC estimation, AIC and BIC can be calculated as:

$$AIC = \overline{D} + p,$$

$$BIC = \overline{D} + (\log J - 1)p,$$

where $\overline{D}$ is the posterior mean of the deviance; p is the number of estimated parameters; J is the sample size. DIC has been designed for model selection in the Bayesian MCMC estimation, which is a generalization of AIC. DIC is calculated as

$$DIC = \overline{D} + p_e = \overline{D} + \text{var}(D)/2,$$

where p_e represents the effective number of parameters. p_e can be approximated by $\text{var}(D)/2$ (Gelman et al., 2003; Su & Yajima, 2015), where $\text{var}(D)$ is the posterior variance of the deviance.

The evidence ratio (Anderson, 2008) has been proposed to examine whether the difference in AIC between two models (among Z data-fitting models) is significant. To determine whether the discrepancies in the model fit indices among the models in comparison are significant, this study calculated the evidence ratio based on AIC and applied the evidence ratio calculation to the other information-based model fit indices that are evaluated in this study, including BIC and DIC. Specifically, the evidence ratio of model Ω_p to model Ω_q is calculated as

$$E_{p,q} = \frac{L(\Omega_p \mid y)}{L(\Omega_q \mid y)} = \frac{\omega_p}{\omega_q},$$

where $L(\Omega_p \mid y)$ and $L(\Omega_q \mid y)$ are the likelihood of models Ω_p and Ω_q, respectively, given the data *y*. ω_p and ω_q are the Akaike weights of evidence of models Ω_p and Ω_q, respectively, being the best fitting model in the set of models. Specifically, ω_p for *p*=1,2,…,*Z*, is calculated based on the difference between the information-based model fit indices:

$$\omega_p = \frac{\exp(-\Delta_p / 2)}{\sum_{z=1}^{Z} \exp(-\Delta_z / 2)},$$

$$\Delta_p = IC_p - IC_{min},$$

where *IC* is a specific type of information-based model fit index (i.e., AIC, BIC or DIC in this study). IC_p is the model fit index value of model Ω_p while IC_{min} is the minimum model fit index value among *Z* data-fitting models. An evidence ratio greater than 55 can serve as a piece of evidence for a significant difference in the model fit index between two models (Anderson, 2008).

Assessing the model parameter recovery. To assess the recovery of the continuous model parameters such as the skill implementation ability parameters, the bias, empirical standard error (SE) and root mean squared error (RMSE) of the parameter estimates are calculated. Specifically, the bias, SE and RMSE are calculated as

$$Bias(y) = \frac{1}{R}\sum_{r=1}^{R} \hat{y} - y_{true},$$

$$SE(y) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{y} - \frac{\sum_{r=1}^{R}\hat{y}}{R}\right)^2},$$

$$RMSE(y) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{y} - y_{true})^2},$$

where y is the parameter to be evaluated, y_{true} is the simulated true value of the parameter, $\hat{y}$ is the parameter estimate, and R is the total number of replications. The bias and SE quantify the systematic errors and random errors, respectively, of a parameter estimate. The RMSE quantifies both the systematic and random errors, as, for a particular parameter, the square of RMSE equals to the sum of squares of bias and SE, i.e.,

$$RMSE^2(y) = Bias^2(y) + SE^2(y).$$

As for the discrete model parameters, such as the strategy latent class membership and attribute mastery status, the classification accuracies of these parameters are assessed. The classification accuracy of the problem-solving strategy is evaluated using the proportion of simulees whose strategy trajectories or strategy latent class memberships at each timepoint are correctly identified. The classification accuracy of the attribute mastery status is quantified with the attribute profile correct classification rate (PCCR) and the attribute correct classification rate (ACCR) at each time point. The former is the proportion of consistency between the true and estimated attribute profiles (i.e., the proportion of simulees with all the attributes being correctly classified out of the whole simulated sample) while the latter is the proportion of consistency between the true and estimated values of a single attribute (i.e., the proportion of simulees being correctly classified in terms of a single attribute

out of the whole simulated sample). Specifically, within each replication, the PCCR and the ACCR of attribute *k* are calculated as

$$PCCR = \frac{1}{J}\sum_{j=1}^{J} I(\hat{\boldsymbol{\alpha}}_{\mathbf{j}} = \boldsymbol{\alpha}_{\mathbf{j}}),$$

$$ACCR_k = \frac{1}{J}\sum_{j=1}^{J} I(\hat{\alpha}_{jk} = \alpha_{jk}),$$

where *I(x)* is a binary indicator of whether the estimated and true attributes/attribute profiles are consistent with each other. *J* is the simulated sample size.

Analyses for assessing the impact of ignoring the multiple-strategy scenarios. To address the second research question, it is necessary to compare the recovery of the parameters of interest across different data-fitting models. This study focuses on examining the effect of ignoring between-person multiple strategies and/or within-person strategy shift on the recovery of the attribute mastery profile, skill implementation ability change and the strategy latent class membership, as these parameters are directly relevant to the diagnostic inferences drawn from the longitudinal CDMs.

As it is possible that the impact of ignoring the multiple-strategy scenarios on the model parameter recovery could vary across different simulated conditions (i.e., there are interaction effects between the data-fitting model type and the manipulated factors on the model parameter recovery), the outcome measures of each type of parameters are plotted against different levels of the manipulated factors and/or the data-fitting model type to help discover the possible interactions between the data-fitting model type and the manipulated factors [i.e., sample size (*J*), initial mixing proportions of the strategies ($\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$), the transition probability from Strategy A to

Strategy B ($p_{M_B|M_A}$) and the correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$)].

To further investigate the statistical and practical significance of the effects of the data-fitting model type and its interactions with the manipulated factors, the mixed-effect analyses of variance (ANOVAs) are planned to be conducted for the parameter types with sufficient sample sizes[3], including the initial ability ($\theta_j^{(T_1)}$, $j = 1, 2, ..., J$), ability change ($\Delta\theta_j$, $j = 1, 2, ..., J$), item intercept ($\lambda_{i,0}$, $i = 1, 2, ..., I$), and attribute main effects[4] ($\lambda_{i,1,(k)}$, $i = 1, 2, ..., I$, $k = 1, 2, ..., K$, $\lambda_{i,1,(k)} \neq 0$), provided that the required assumptions for the ANOVAs are met or the ANOVA inferences are robust to the violation of assumptions. The ANOVAs and the corresponding assumption checks are implemented with IBM SPSS Statistics 20 (IBM Corporation, 2011). Specifically, in the mixed-effect ANOVA, each parameter is treated as a "subject". Given that three models (i.e. Longitudinal LLM, Longitudinal MCDM and LTA-longitudinal-MCDM) are fitted to the same dataset in each replication, the data-fitting model type is treated as the repeated-measure factor (i.e., within-subject factor). The recovery outcome measures (i.e., bias/SE/RMSE) of the parameters are treated as repeated measurements (i.e., the dependent variable) taken on each subject (i.e., parameter). The manipulated factors are treated as between-subject factors. Thus, there are $3 \times 24 = 72$ cells of the design in the mixed-effect ANOVAs, and each

[3] The number of parameters of the parameter type is greater than 20.
[4] Twenty-two attribute main effect parameters are present in all the three data-fitting models, and thirteen attribute main effect parameters are present only in the Longitudinal MCDM and LTA-longitudinal-MCDM. The mixed-effect ANOVA was performed on the 22 attribute main effect parameters that are present in all the three models.

cell contains a parameter recovery outcome measure yielded from one of the three data-fitting models under one of the twenty-four simulated conditions.

Before conducting the mixed-effect ANOVAs, three assumptions of the ANOVAs are checked, including the normality of residuals, the homogeneity of residual variances and sphericity. The assumption of residual normality is assessed by testing whether the dependent variable in each cell is normally distributed with the Shapiro-Wilk test of normality (Shapiro & Wilk, 1965). Nevertheless, since the *F*-test in the ANOVA is considered to be fairly robust to the violation of the normality assumption (Pearson, 1931; Tiku, 1964), the ANOVA will be carried out in this study even if the normality assumption is violated, as long as the nonnormality is not extreme. The assumption of homogeneous residual variances means that the residual variances of the dependent variable are equal across groups of between-subject factors. This assumption is tested with the Levene's test of equality of error variances (Levene, 1960), the null hypothesis of which is that "the error variances of the dependent variable are equal across the groups". Thus, if the test statistics from the Levene's test is significantly different from zero, it can be inferred that the assumption of homogeneity is violated. The violation of the assumption of homogeneity may affect the Type I error rate of the ANOVA (Box, 1954; Horsnell, 1953). However, the ANOVA results were found to be robust to the violation of the assumption of homogeneity when the sample sizes are approximately equal across the groups (Kohr & Games, 1974), and it is also suggested to use equal sample size designs to protect against the violation of homogeneity assumption (Maxwell & Delaney, 1990). Therefore, given that the sample sizes of the item intercept ($\lambda_{i,0}$,

$i = 1, 2, ..., I$) and attribute main effect ($\lambda_{i,1,(k)}$, $i = 1, 2, ..., I$, $k = 1, 2, ..., K$, $\lambda_{i,1,(k)} \neq 0$) are equal across groups in the ANOVA, the ANOVA will be carried out for these parameters even if the assumption of homogeneity is violated. However, when the ANOVA is performed on the initial ability ($\theta_j^{(T_1)}$, $j = 1, 2, ..., J$) and ability change ($\Delta\theta_j$, $j = 1, 2, ..., J$) parameters, the group sample sizes would differ across different levels of the sample size factor (*J*); thus, if the assumption of homogeneity is violated, an alternative design of ANOVA will be carried out for the skill implementation ability parameters as elaborated in the following paragraph. The sphericity assumption means that the variances of the differences between the levels of the within-subject factor are equal, which can be tested with Mauchly's test of sphericity (Mauchly, 1940). If the sphericity assumption is violated, the *p*-values in the ANOVA results will be corrected by adjusting the degrees of freedom with the Greenhouse-Geisser procedure (Greenhouse & Geisser, 1959). In addition, observations are assumed to be randomly and independently sampled in ANOVA. The violation of the independence assumption could result in an inflated Type I error rate in the ANOVA (e.g., Kenny & Judd, 1986).

The mixed-effect ANOVA is conducted for each type of model parameters separately. If not otherwise specified, all the possible main effects and interactions among the data-fitting model type and the four manipulated factors are included in the mixed-effect ANOVA design. If the assumption of homogeneous residual variances is violated, the inferences about the effects of the factors on the recovery of the initial ability ($\theta_j^{(T_1)}$) and ability change ($\Delta\theta_j$) parameters based on the full

ANOVA design can be misleading due to the unequal sample size issue as mentioned in the last paragraph. Thus, if the assumption of homogeneous residual variances is violated, the mixed-effect ANOVAs will be performed separately for the small sample size (J=100) and large sample size (J=800) conditions to investigate the effects of the other three manipulated factors (i.e., $\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$, $p_{M_B|M_A}$ and $\rho_{\theta^{(T_1)}\Delta\theta}$) and data-fitting model type on the recoveries of $\theta_j^{(T_1)}$ and $\Delta\theta_j$.

A main or interaction effect with a p-value smaller than 0.05 is deemed statistically significant. In addition to the statistical significance, this study evaluated the effect size quantified by the partial η^2 (Cohen, 1965) for each effect. The effect size is used as a measure of practical significance of the results. Using the criterion suggested by Cohen (1988), i.e., $0.01 \leq$ partial $\eta^2 < 0.06$ for small effect, $0.06 \leq$ partial $\eta^2 < 0.14$ for medium effect, and partial $\eta^2 \geq 0.14$ for large effect, this study only reports and further investigates the effects that are both statistically significant and have at least a small effect size (i.e., partial $\eta^2 \geq 0.01$).

Analyses for assessing the effects of the manipulated factors on the parameter recovery of the proposed model. To address the third research question, effects of the manipulated factors on the LTA-longitudinal-MCDM parameter recovery need to be examined. As a graphical inspection, marginal means of the outcome measures of each type of parameters are plotted against different levels of the manipulated factors. Like in the second research question, the statistical and practical significance of the effects of the manipulated factors on the recovery of parameters of a sufficient sample size can be examined with the ANOVAs. In

particular, four-way ANOVAs, with each manipulated factor as a factor of the design, are planned to be conducted on the recovery outcome measures of the initial ability, ability change, item intercept and attribute main effect parameters of the LTA-longitudinal-MCDM. All the ANOVA assumptions described above, except the sphericity assumption which is not required by the four-way ANOVAs, are checked before conducting the four-way ANOVAs.

Checking the stability of the simulation results. Thirty replications are run. The number of replications is chosen based on the previous study on the longitudinal CDM (Zhan, Jiao, Liao, et al., 2019). To further justify the sufficiency of the number of replications, a pilot study was conducted where 70 replications were run in a selected simulated condition (J=100, $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.6 : 0.4$, $p_{M_B|M_A}$=0.3, $\rho_{\theta^{(T_1)}\Delta\theta}$=-0.3). The recovery outcome measures of all the parameters in the proposed model (i.e., the LTA-longitudinal-MCDM) are plotted against the number of replications (ranging from 2 to 70) to investigate the stability of the simulation study results at the replication number of 30. Figure 7 displays the classification accuracy results of the categorical model parameter estimates, including the attribute mastery status (α) and the strategy choice (m), whereas Figure 8 displays the bias, SE and RMSE of the estimates of the continuous model parameters, such as the initial ability ($\theta^{(T_1)}$), ability change ($\Delta\theta$), item intercept ($\lambda_{i,0}$), attribute main effect ($\lambda_{i,1,(k)}$) and strategy latent transition probability ($\tau^{(T_1)}_{M_B|M_A}$). Both Figures 7 and 8 demonstrate that the stability in recovery outcome measures of the LTA-longitudinal-MCDM parameters has been achieved at the 30th replication. The stabilities in the two more constrained alternative

models, the Longitudinal LLM and Longitudinal MCDM, have also been reached by the 30th replication although not presented here.

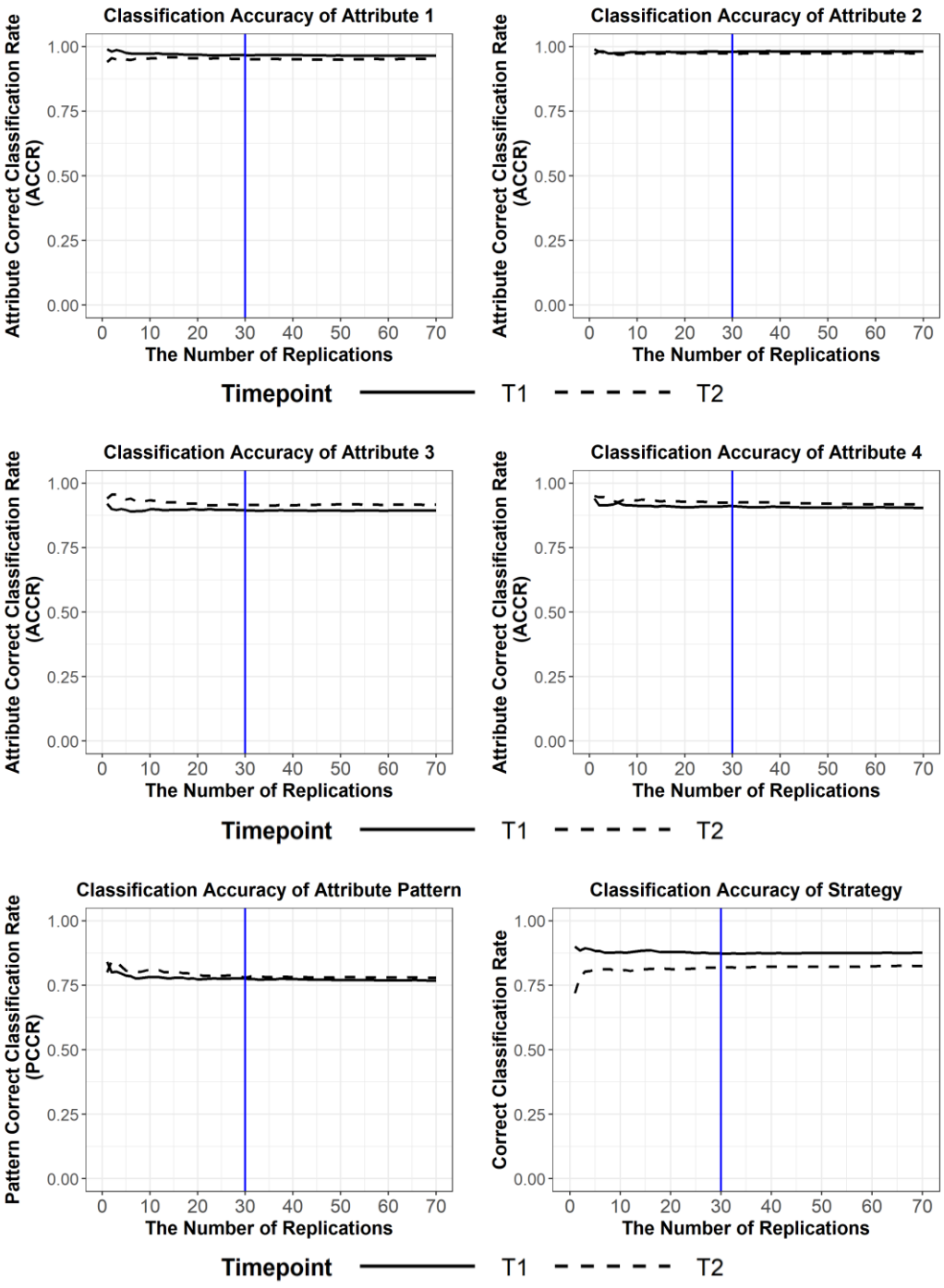
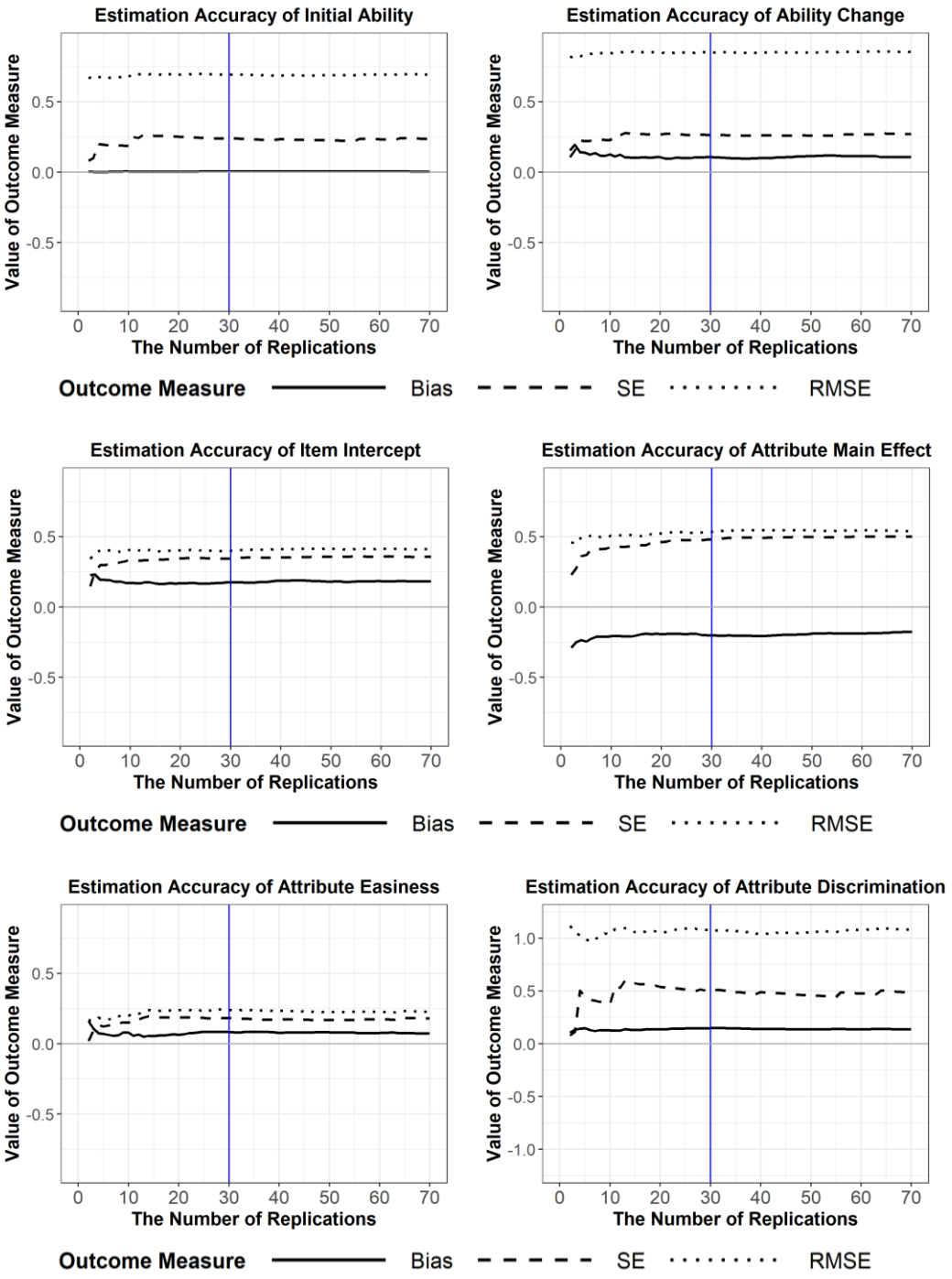
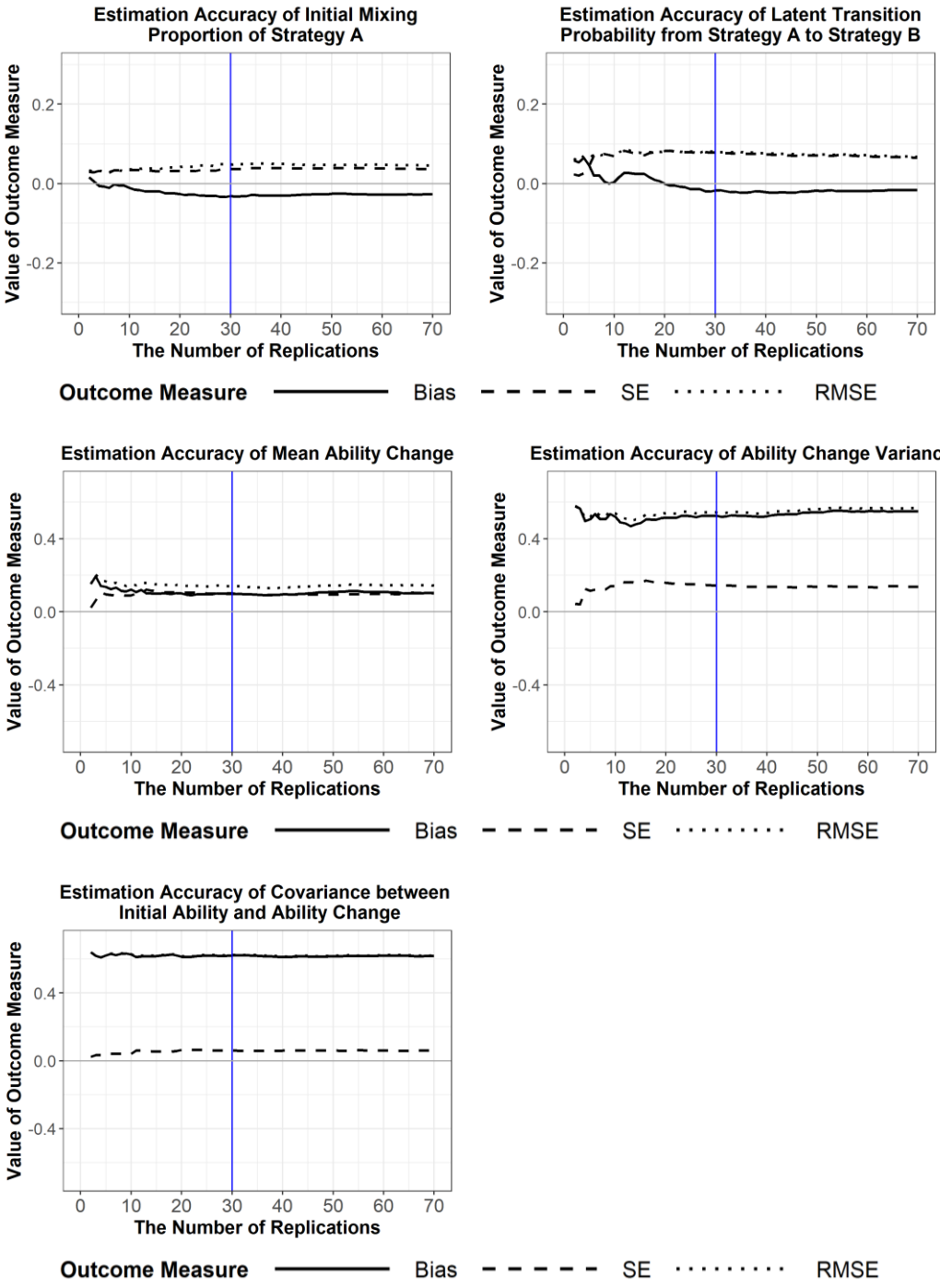


Figure 7. Correct classification rate of the categorical parameter estimates in the LTA-longitudinal-MCDM by the number of replications.

Figure 8. Bias, SE and RMSE of the continuous parameter estimates in the LTA-longitudinal-MCDM by the number of replications.


Estimation Accuracy of Initial Ability
Estimation Accuracy of Ability Change
Estimation Accuracy of Item Intercept
Estimation Accuracy of Attribute Main Effect
Estimation Accuracy of Attribute Easiness
Estimation Accuracy of Attribute Discrimination
Value of Outcome Measure
The Number of Replications
Outcome Measure
Bias
SE
RMSE


Estimation Accuracy of Initial Mixing Proportion of Strategy A
Estimation Accuracy of Latent Transition Probability from Strategy A to Strategy B
Estimation Accuracy of Mean Ability Change
Estimation Accuracy of Ability Change Variance
Estimation Accuracy of Covariance between Initial Ability and Ability Change
Value of Outcome Measure
The Number of Replications
Outcome Measure
Bias
SE
RMSE

3.4 Empirical Data Analyses

3.4.1 Data and research questions

The application of the proposed model is demonstrated with the empirical data from a study (Bottge et al., 2015) assessing the effectiveness of the Enhanced Anchored Instruction (EAI; Bottge, 2001) and comparing the effectiveness EAI to that of the business as usual (BAU). The study had a repeated-measure pretest-posttest design, and the participated schools were randomly assigned into the EAI or BAU condition (Bottge et al., 2015). Students in the BAU condition were given traditional mathematics instructions while students in the EAI condition were given instructions with realistic problems embedded in more interactive formats [See Bottge et al. (2014, 2015) for detailed instructional activities involved in the two conditions].

The empirical dataset contains item responses of 879 middle-school students (456 were in the BAU condition and 423 were in the EAI condition). Both the pretest and posttest contain 21 dichotomously scored items aiming at measuring students' mathematical problem-solving ability. The items were designed to measure four attributes, including 1) ratios and proportional relationships (RPR), 2) measurement and data (MD), 3) number system – fractions (NSF) and 4) geometry – graphing (GG). The expert-developed Q-matrix is shown in Table 7, which is referred to as the theoretical Q-matrix ($\mathbf{Q}_{\mathbf{T}}$).

The proposed LTA-longitudinal-MCDM is applied to the dataset to address two research questions:

1) How do students' strategy choice, overall skill implementation ability and attribute mastery status change from the pretest to the posttest?

2) Do EAI and BAU differ in terms of their effects on students' learning outcomes regarding the strategy choice, overall skill implementation ability and attribute mastery status?

Each of the research question has three perspectives, i.e., the strategy choice, overall skill implementation ability and attribute mastery status, each of which can be inferred from a type of parameters in the LTA-longitudinal-MCDM. Specifically, to answer the first research question, the strategy latent transition probability estimates ($\hat{\boldsymbol{\tau}}^{(\mathbf{T_1})}_{\mathbf{m}_{\mathbf{T_2}}|\mathbf{m}_{\mathbf{T_1}}}$) are reported to inform students' shift in the strategy choice over time; the mean of ability change estimate ($\hat{\mu}_{\Delta\theta}$) can provide inferences on students' skill implementation ability change; and the frequencies of attribute mastery status patterns ($\hat{\alpha}^{(t)}_{jk}$) are summarized to provide information on students' attribute mastery change over time.

In this study, the learning outcome regarding the strategy choice is operationally defined as the distribution of strategy choice trajectory; the learning outcome regarding the overall skill implementation ability is operationally defined as the ability change estimates of the individuals; the learning outcome regarding the attribute mastery status is operationally defined as the proportion of attribute non-mastery students at the pretest who are classified as attribute mastery at the posttest. Thus, to answer the second research question, the distribution of the strategy choice trajectory, average ability change estimates of the individuals, and the proportion of attribute non-mastery students at the pretest who are classified as attribute mastery at the posttest, are compared across the EAI and BAU groups.

3.4.2 Data analysis procedure

3.4.2.1 Data cleaning

Thirty out of the 879 students have missing item responses, and these students have missed 43% to 100% of the items in either the pretest or posttest. As the students with missing data only take up a small percentage of the sample (3.4%), these students are excluded from the analysis in this demonstration. As a result, the analytical sample contains 849 students (435 were in the BAU condition and 414 were in the EAI condition).

3.4.2.2 Empirical Q-matrix development

As improving students' problem solving is the main goal of either the EAI or BAU instructions (Bottge et al., 2014), which include teaching students to choose a more effective problem-solving strategies, this study assumes the existence of multiple Q-matrices representing different strategies before and after the instructional interventions. Given that little expert knowledge is available about alternative theoretical Q-matrices, a nonparametric Q-matrix refinement method (Chiu, 2013) is used to empirically construct two Q-matrices, one based on the pretest and the other based on posttest data. It is expected that the discrepancies in these two empirical Q-matrices can capture some differences in the problem-solving strategies before and after the interventions. An advantage of empirically constructing the Q-matrix is that it complements the expert knowledge and has the potential to discover strategies that are not expected by experts. In addition, the empirical Q-matrices are less prone to subjectivity of human judgment. In general, the nonparametric Q-matrix refinement method (Chiu, 2013) is implemented as follows: The theoretical Q-matrix is input as

a provisional Q-matrix to estimate the person attribute profiles for the initial iteration, and an iterative process is used to find the optimal Q-matrix that minimizes the residual sum of squares between the expected response and the observed response. The detailed empirical Q-matrix development algorithm can be found in Chiu (2013). Compared to the other empirical Q-matrix development methods, this nonparametric method is advantageous in that it does not rely on model-based assumptions about the observed item responses and it is computationally efficient (Chiu, 2013). As the focus of this empirical data analysis is to demonstrate the use of the LTA-longitudinal-MCDM, other Q-matrix construction methods are not considered. However, future research could investigate the effect of the different Q-matrix development methods on the results. The resulting Q-matrices from the empirical Q-matrix development method are labelled as empirical Q-matrices ($\mathbf{Q}_{\mathbf{E}}$) to be distinguished from the theoretical Q-matrix.

To attenuate the overfitting issue, the analytical dataset is randomly split into two sub-datasets, labelled as the "training set" (J=100; 53 in BAU and 47 in EAI) and "testing set" (J=749; 382 in BAU and 367 in EAI). The training set is used for the empirical Q-matrix development while the testing set is used for empirical Q-matrix validation and model fit.

The empirical Q-matrices developed with the nonparametric Q-matrix refinement method using the NPCD package (Zheng et al., 2014) are shown in Table 7. The empirical Q-matrices developed from the pretest ($\mathbf{Q}_{\mathbf{Epre}}$) and posttest ($\mathbf{Q}_{\mathbf{Epost}}$) have 16 and 4 different q-entries (shaded in yellow in Table 7) from the theoretical Q-matrix, respectively. For most of the items with discrepant theoretical and empirical

q-vectors, the empirical q-vectors involve more attributes than the theoretical q-vectors do.

Table 7
Q-matrices Used in the Empirical Data Analysis

Item	$\mathbf{Q}_{\mathbf{T}}$ (Theoretical Strategy)				$\mathbf{Q}_{\mathbf{Epre}}$ (Empirical Complex Strategy)				$\mathbf{Q}_{\mathbf{Epost}}$ (Empirical Simple Strategy)			
	RPR	MD	NSF	GG	RPR	MD	NSF	GG	RPR	MD	NSF	GG
1	1	0	0	0	1	1	0	0	1	0	0	0
2	0	1	0	0	0	1	0	0	0	1	0	0
3	0	1	0	0	1	1	1	1	0	1	1	0
4	0	0	1	0	0	1	1	0	0	0	1	0
5	0	0	1	0	0	1	1	0	0	0	1	0
6	0	0	1	0	0	0	1	0	0	0	1	0
7	0	0	1	0	1	0	1	1	0	0	1	0
8	0	0	1	0	0	0	1	1	0	0	1	0
9	0	1	0	0	0	1	0	0	0	1	0	0
10	0	1	0	0	0	1	0	0	0	1	0	0
11	0	1	0	0	0	1	0	0	0	1	0	0
12	0	1	0	0	0	1	0	0	0	1	0	0
13	1	0	0	0	0	1	0	0	1	0	0	0
14	1	0	0	0	0	1	1	0	1	0	0	0
15	0	0	0	1	0	0	0	1	0	0	0	1
16	0	0	0	1	0	0	0	1	0	0	0	1
17	1	0	0	0	1	0	0	0	1	0	1	1
18	0	0	0	1	0	0	0	1	0	0	0	1
19	0	0	0	1	0	0	0	1	0	0	0	1
20	0	0	0	1	1	0	0	1	0	0	0	1
21	0	0	0	1	1	0	0	1	0	1	0	1

Note. RPR=ratios and proportional relationships; MD=measurement and data; NSF=number system – fractions; GG=geometry – graphing. The empirical q-entries that are different from the theoretical q-entries are shaded in yellow. The item numbers of the multiple-approach items (different q-vectors between $\mathbf{Q}_{\mathbf{Epre}}$ and $\mathbf{Q}_{\mathbf{Epost}}$) are shaded in blue.

Eleven items have different q-vectors between $\mathbf{Q}_{\mathbf{Epre}}$ and $\mathbf{Q}_{\mathbf{Epost}}$ (The item numbers are shaded in blue in Table 7), which suggest that these items may be solved in a different approach in the posttest from that in the pretest. As eight out of the eleven multiple-approach items involve more attributes in $\mathbf{Q}_{\mathbf{Epre}}$ than $\mathbf{Q}_{\mathbf{Epost}}$, the

strategies associated with $\mathbf{Q}_{\mathbf{Epre}}$ than $\mathbf{Q}_{\mathbf{Epost}}$ are labelled as "empirical complex strategy ($M_{E,Complex}$)" and "empirical simple strategy ($M_{E,Simple}$)", respectively, in the following sections for the convenience of interpretation. Accordingly, the strategy associated with the theoretical Q-matrix is labelled as "theoretical strategy (M_T)".

The two empirical Q-matrices are further validated with the testing dataset by fitting the following models: Four single-time-point models (the models in Table 8) and four longitudinal models (the first four models in Table 9) with different Q-matrices are fitted to the testing dataset. The four single-time-point models include two LLMs with either the theoretical or one of the empirical Q-matrices (i.e., S-LLM-T and S-LLM-E), an MCDM with the theoretical and one of the empirical Q-matrices (S-MCDM-TE) and an MCDM with the two empirical Q-matrices (S-MCDM-EE). The four longitudinal models include three Longitudinal LLMs with either the theoretical or one of the empirical Q-matrices (i.e., L-LLM-T, L-LLM-E-pre and L-LLM-E-post) and a Longitudinal MCDM with a mixture of the two empirical Q-matrices (L-MCDM-EE). The relative model fit indices, AIC, BIC and DIC, were compared across the models. The S-MCDM-EE is identified as the best-fitting model among the four single-time-point models by both AIC and BIC in both the pretest and posttest; the L-MCDM-EE is identified as the best-fitting model among the four longitudinal models by both AIC and BIC (The detailed Q-matrix validation results will be presented in Tables 34 and 35 in Chapter 5). These model comparison results provide a justification for using the empirical Q-matrices, $\mathbf{Q}_{\mathbf{Epre}}$ and $\mathbf{Q}_{\mathbf{Epost}}$, in the subsequent longitudinal analyses involving the LTA-longitudinal-MCDM. Further, the absolute model-data fit is evaluated using the posterior predictive model check,

the procedure of which is the same as that carried out in the simulation study (See Section 3.3.5). An adequate model-data fit could also serve as evidence supporting the appropriateness of the empirical Q-matrices.

Table 8
Single-Time-Point Model Specifications for Empirical Q-matrix Validation

Model No.	Model Abbrevation	Model	Pretest Data			Posttest Data		
			$\mathbf{Q_T}$	$\mathbf{Q_{Epre}}$	$\mathbf{Q_{Epost}}$	$\mathbf{Q_T}$	$\mathbf{Q_{Epre}}$	$\mathbf{Q_{Epost}}$
1	S-LLM-T	LLM	√			√		
2	S-LLM-E	LLM		√				√
3	S-MCDM-TE	MCDM	√	√		√		√
4	S-MCDM-EE	MCDM		√	√		√	√

Note. $\boldsymbol{Q}_T$ =Theoretical Q-matrix; $\boldsymbol{Q}_{Epre}$ =Empirical Q-matrix based on the pretest; $\boldsymbol{Q}_{Epost}$ = Empirical Q-matrix based on the posttest.

Table 9
Longitudinal Model Specifications

Model No.	Model Abbrevation	Model	Q-matrix		
			$\mathbf{Q_T}$	$\mathbf{Q_{Epre}}$	$\mathbf{Q_{Epost}}$
1	L-LLM-T	Longitudinal LLM	√		
2	L-LLM-E-pre	Longitudinal LLM		√	
3	L-LLM-E-post	Longitudinal LLM			√
4	L-MCDM-EE	Longitudinal MCDM		√	√
5	LTA-L-MCDM-EE	LTA-longitudinal-MCDM		√	√

Note. $\boldsymbol{Q}_T$ =Theoretical Q-matrix; $\boldsymbol{Q}_{Epre}$ =Empirical Q-matrix based on the pretest; $\boldsymbol{Q}_{Epost}$ = Empirical Q-matrix based on the posttest.

3.4.2.3 LTA-longitudinal-MCDM analysis

In order to answer the research questions, the LTA-longitudinal-MCDM is fit to the testing dataset. $\mathbf{Q_{Epre}}$ and $\mathbf{Q_{Epost}}$, are used as the Q-matrices for both timepoints. Thus, there expected to be four unique strategy choice trajectories, i.e., $(M_{E,Complex}, M_{E,Complex})$, $(M_{E,Complex}, M_{E,Simple})$, $(M_{E,Simple}, M_{E,Complex})$ and

$(M_{E,Simple}, M_{E,Simple})$. As mentioned in Section 3.4.1, the estimates of the model parameters relevant to strategy choice, overall skill implementation ability change and attribute mastery status are summarized and reported to address the research questions.

In addition, the relative model fit indices (i.e., AIC, BIC and DIC) of the LTA-longitudinal-MCDM are compared to those of the other longitudinal models listed in Table 9, including the L-MCDM-EE that ignores within-person strategy shift and the L-LLMs that ignore both between-person multiple strategies and within-person strategy shift.

Chapter 4: Simulation Study Results

The simulation study was conducted to examine 1) the performance of AIC, BIC and DIC in correctly selecting the LTA-longitudinal-MCDM as the best-fitting model in the presence of between-person multiple strategies and within-person strategy shift; 2) the impact of ignoring the multiple-strategy scenarios in the model on the parameter recovery of the longitudinal CDMs; and 3) the effect of the manipulated factors on the parameter recovery of the LTA-longitudinal-MCDM.

In particular, the effects of ignoring between-person multiple strategies and within-person strategy shift on the parameter recovery were examined by comparing the parameter recovery outcome measures across three data-fitting models, including i) the LTA-longitudinal-MCDM that models both between-person multiple strategies and within-person strategy shift, ii) the Longitudinal-MCDM that ignores within-person strategy shift and iii) the Longitudinal LLM that ignores both between-person multiple strategies and within-person strategy shift. An overview of the model specification of the three data-fitting models is presented in Table 10.

Table 10
Overview of Model Specifications of the Data-Fitting Model in the Simulation Study

Model	The presence multiple-strategy scenario in the models	
	Between-person multiple strategies	Within-person strategy shift
Longitudinal LLM	×	×
Longitudinal MCDM	×	√
LTA-longitudinal-MCDM	√	√

Note. × represents absence; √ represents presence.

Four factors were manipulated (See Table 6 for detailed specifications of the manipulated factors) in the simulation study, including the sample size, the initial mixing proportions of the strategies, the strategy latent transition probability and the

correlation between the initial strategy and strategy change, resulting in a total of 24 simulated conditions. Thirty replications were run in each simulated condition, yielding a total of $30 \times 24 = 720$ replications. The model parameters were estimated with Bayesian MCMC method. Two MCMC chains were run, each of which contained 5,000 iterations including 2,500 iterations as burn-in and a thinning of 2. As a result, estimates of the model parameters were summarized based on a total of 2,500 iterations. Convergence was achieved for all the model parameters in all the replications and simulated conditions, according to the $\hat{R} < 1.1$ criterion and the trace plots. As for the computational efficiency, running the LTA-longitudinal-MCDM with two MCMC chains each containing 5,000 iterations took around 3 minutes and 90 minutes for the small sample size (J=100) and large sample size (J=800) conditions, respectively.[5]

In order to confirm that the simulated datasets possess the desired characteristics of the data-generating model (i.e., the LTA-longitudinal-MCDM), the absolute fit of the LTA-longitudinal-MCDM to each simulated dataset was inspected. Table 11 lists the summary statistics that reflect the distributions of posterior predictive p-value (PPP) across the 30 replications for each simulated condition. The smallest PPP value of the proposed model among all the replications is 0.47 which is not extremely close to 0. The PPP values range from 0.47 to 0.78, meaning that the proportion of the replicated data generated from the LTA-longitudinal-MCDM having a sum of squares of standardized residuals that are greater than that of the

[5] The computing time is based on the analyses run on a desktop with Intel Core i7 CPU and 3.2GHz processor. Multiple MCMC chains were run in parallel with multiple cores.

observed data ranges from 0.47 to 0.78. These PPP value results support that the observed datasets are likely to be seen in the replicated data if the LTA-longitudinal-MCDM is the true data-generating model. In other words, the simulated datasets possess the characteristics of the data-generating model from the perspective of the sum of squares of standardized residuals.

Table 11
Summary of the Posterior Predictive P-Values of the LTA-longitudinal-MCDM under the 24 Simulated Conditions

Condition No.	J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	PPP value			
					Min	Median	Mean	Max
1	100	0.6:0.4	0.3	-0.3	0.55	0.67	0.66	0.74
2				0	0.54	0.67	0.67	0.78
3				0.3	0.60	0.66	0.67	0.76
4			0.7	-0.3	0.53	0.66	0.65	0.71
5				0	0.52	0.65	0.65	0.75
6				0.3	0.58	0.64	0.66	0.76
7		0.8:0.2	0.3	-0.3	0.55	0.68	0.67	0.76
8				0	0.55	0.67	0.67	0.76
9				0.3	0.54	0.66	0.67	0.78
10			0.7	-0.3	0.50	0.66	0.66	0.73
11				0	0.58	0.64	0.66	0.77
12				0.3	0.54	0.64	0.65	0.76
13	800	0.6:0.4	0.3	-0.3	0.50	0.56	0.55	0.60
14				0	0.50	0.57	0.56	0.59
15				0.3	0.51	0.55	0.56	0.63
16			0.7	-0.3	0.49	0.54	0.55	0.61
17				0	0.49	0.56	0.56	0.62
18				0.3	0.47	0.53	0.54	0.60
19		0.8:0.2	0.3	-0.3	0.51	0.55	0.55	0.59
20				0	0.47	0.55	0.55	0.59
21				0.3	0.51	0.55	0.55	0.62
22			0.7	-0.3	0.50	0.56	0.56	0.61
23				0	0.49	0.55	0.56	0.62
24				0.3	0.49	0.54	0.55	0.63

Note. PPP=Posterior Predictive P-Values. J=Sample size; $\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$ =initial mixing proportions of the strategies; $p_{M_B|M_A}$ =transition probability from Strategy A to Strategy B; $\rho_{\theta^{(T_1)}\Delta\theta}$ =correlation between the initial ability and ability change.

The remaining simulation study results are arranged and presented by the type of outcomes or model parameters. Table 12 provides an overview of the model parameters evaluated in the simulation study. Since some parameters (e.g., the strategy choice parameters) are absent in certain data-fitting models, the recovery of these parameters was only compared across the models that contain them. If a parameter only exists in one model (e.g., the strategy latent transition probability parameter only exists in the LTA-longitudinal-MCDM), only the descriptive statistics of the parameter recovery outcome measures will be presented. Specifically, the remaining part of this chapter is divided into four major sections: (a) the performance of the model fit indices, (b) the recovery of the person parameters, (c) the recovery of the item parameters and (d) the recovery of the higher-order structural parameters.

Table 12

Overview of the Model Parameters Evaluated in the Simulation Study

Parameter type	Parameter	Description	The presence of parameter in the models		
			L-LLM	L-MCDM	LTA-L-MCDM
Person parameter	$\alpha_{jk}^{(t)}$	Attribute mastery status	√	√	√
	$\theta^{(T_1)}$	Initial skill implementation ability	√	√	√
	$\Delta\theta$	Skill implementation ability change	√	√	√
	$\mu_{\Delta\theta}$	Mean of the skill implementation ability change	√	√	√
	$\sigma_{\Delta\theta}^2$	Variance of the skill implementation ability change	√	√	√
	$\sigma_{\theta^{(T_1)}\Delta\theta}$	Covariance between the initial skill implementation ability and ability change	√	√	√
	m	Strategy choice membership	×	√	√
	$\boldsymbol{\pi}_m^{(T_1)}$	Initial mixing proportion of the strategy	×	√	√
	$\boldsymbol{\tau}_{m_{T_2}\|m_{T_1}}^{(T_1)}$	Strategy latent transition probability (from Timepoint 1 to Timepoint 2)	×	×	√
Item parameter	$\lambda_{i,0}$	Item intercept	√	√	√
	$\lambda_{i,1,(k)}$	Attribute main effect	√	√	√
Higher-order structural parameter	β_k	Attribute easiness	√	√	√
	ξ_k	Attribute discrimination	√	√	√

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal-MCDM. × represents absence; √ represents presence.

4.1 Performance of the Model Fit Indices

To evaluate the performance of AIC, BIC and DIC in correctly selecting the LTA-longitudinal-MCDM as the best-fitting model, the number of replications of each data-fitting model (Longitudinal LLM, Longitudinal MCDM and LTA-longitudinal-MCDM) being identified as the best-fitting model by each relative fit index are reported in Table 13. Specifically, the model with the smallest model fit index was labeled as the best-fitting model. The LTA-longitudinal-MCDM (i.e., the data-generating model) was correctly identified as the best-fitting model by AIC and BIC in nearly all the replications (i.e., 29 to 30 out of 30) under all the simulated conditions. The performance of DIC varies across conditions. In particular, DIC correctly identified the LTA-longitudinal-MCDM as the best-fitting model in nearly all the replications (29 to 30 out of 30) under the conditions with a high transition probability from Strategy A to Strategy B ($p_{M_B|M_A} = 0.7$). However, DIC incorrectly favored the Longitudinal MCDM which only takes into account between-person multiple strategies in a small proportion of replications (7 to 10 out of 30) under the conditions with a low transition probability from Strategy A to Strategy B, imbalanced initial mixing proportions of the strategies and a small sample size, simultaneously (i.e., $p_{M_B|M_A} = 0.3$; $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.8 : 0.2$; $J = 100$). Further, under the conditions with both a low transition probability from Strategy A to Strategy B and balanced initial mixing proportions of the strategies ($p_{M_B|M_A} = 0.3$; $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.6 : 0.4$), DIC tended to incorrectly identify the Longitudinal MCDM as the best-fitting model in most of the replications. In sum, the ability of DIC to correctly select the proposed model as the best-fitting model is diminished when the

true latent transition probability from Strategy A to Strategy B is low ($p_{M_B|M_A} = 0.3$), and such diminishment is more severe under the conditions with more balanced initial mixing proportions of strategies ($\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)} = 0.6 : 0.4$) .

While it is convenient to identify the best-fitting model by directly comparing the model fit indices, it remains unknown whether the discrepancies in fit indices between the LTA-longitudinal-MCDM as the best-fitting model and the alternative models are significantly large. Therefore, to examine the significance in discrepancies, the evidence ratios of the LTA-longitudinal-MCDM to the alternative models were calculated when the LTA-longitudinal-MCDM has the smallest fit index among the three data-fitting models. In this study, the evidence ratio being greater than 55 was used as a criterion to determine the significant difference in the model fit index between two models (Anderson, 2008). Table 14 reports the frequency of the evidence ratio of the LTA-longitudinal-MCDM to the alternative models being greater than 55 (among the replications where the LTA-longitudinal-MCDM was identified as the best-fitting model).

Table 13

The Number of Replications of Each Model Identified as the Best-Fitting Model in the Simulation Study

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	AIC			BIC			DIC		
				L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0	0	**30**	0	0	**30**	0	**22**	8
			0	0	0	**30**	0	1	**29**	0	**25**	5
			0.3	0	0	**30**	0	0	**30**	0	**23**	7
		0.7	-0.3	0	0	**30**	0	0	**30**	0	0	**30**
			0	0	0	**30**	0	0	**30**	0	0	**30**
			0.3	0	0	**30**	0	0	**30**	0	0	**30**
	0.8:0.2	0.3	-0.3	0	0	**30**	0	0	**30**	0	9	**21**
			0	0	0	**30**	0	0	**30**	0	10	**20**
			0.3	0	0	**30**	0	0	**30**	0	7	**23**
		0.7	-0.3	0	0	**30**	0	0	**30**	0	1	**29**
			0	0	0	**30**	0	0	**30**	0	0	**30**
			0.3	0	0	**30**	0	0	**30**	0	1	**29**
800	0.6:0.4	0.3	-0.3	0	0	**30**	0	0	**30**	0	13	**17**
			0	0	0	**30**	0	0	**30**	0	**21**	9
			0.3	0	0	**30**	0	0	**30**	0	**23**	7
		0.7	-0.3	0	0	**30**	0	0	**30**	0	0	**30**
			0	0	0	**30**	0	0	**30**	0	0	**30**
			0.3	0	0	**30**	0	0	**30**	0	0	**30**
	0.8:0.2	0.3	-0.3	0	0	**30**	0	0	**30**	0	1	**29**
			0	0	0	**30**	0	0	**30**	0	0	**30**
			0.3	0	0	**30**	0	0	**30**	0	1	**29**
		0.7	-0.3	0	0	**30**	0	0	**30**	0	0	**30**
			0	0	0	**30**	0	0	**30**	0	0	**30**
			0.3	0	0	**30**	0	0	**30**	0	0	**30**

Note. The largest numbers of replications among the three model under each condition are bolded. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal-MCDM.

The "Total" columns in Table 14 list the frequencies of the LTA-longitudinal-MCDM having the smallest fit indices among the three models, the values of which match those in the "LTA-L-MCDM" columns in Table 13. It can be seen that, among the replications where the LTA-longitudinal-MCDM have the smallest AIC, BIC or DIC, the differences in the fit indices between the LTA-longitudinal-MCDM and the Longitudinal LLM are significant under all the conditions. The differences between the LTA-longitudinal-MCDM and the Longitudinal MCDM in terms of AIC and BIC are significant in nearly all the replications (with at most one exception) under all the conditions. Nevertheless, insignificant DIC discrepancies between the LTA-longitudinal-MCDM and the Longitudinal MCDM are observed in some of the conditions with both a low transition probability from Strategy A to Strategy B and balanced initial mixing proportions of the strategies ($p_{M_B|M_A} = 0.3$; $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.6 : 0.4$).

Further, for the replications where the Longitudinal MCDM has the smallest relative fit index among the three models, the evidence ratios of the Longitudinal MCDM to the LTA-longitudinal-MCDM were calculated. None of these evidence ratios was greater than 55. The small evidence ratios of the Longitudinal MCDM to the LTA-longitudinal-MCDM are a result of the small magnitude of difference in DIC between the two models, suggesting that, even if the Longitudinal MCDM had a smaller fit index than the LTA-longitudinal-MCDM in certain conditions, no evidence was found to support a significant discrepancy in model fit between the two models. Such results are expected since the LTA-longitudinal-MCDM is the true data-generating model, while the Longitudinal MCDM is an under-specified data-

fitting model. Even though DIC may overly punish the LTA-longitudinal-MCDM for its model complexity in certain conditions, resulting in a higher DIC for the LTA-longitudinal-MCDM than the Longitudinal MCDM, a strong evidence favoring the under-specified model to the true data-generating model is not expected.

In summary, AIC and BIC had a satisfying performance under the simulated conditions – they correctly identified the proposed model as the best-fitting model such that the proposed model display significant discrepancies against the alternative models that ignore the multiple-strategy scenarios. The performance of DIC was sensitive to some of the manipulated factors, such as the true initial mixing proportions of strategies and the latent transition probability from Strategy A to Strategy B. DIC tended to select the Longitudinal MCDM which ignores the within-person strategy shift as the best-fitting model when the initial mixing proportions of strategies was balanced and the latent transition probability from Strategy A to Strategy B was low. Even so, the discrepancies in DIC between the Longitudinal MCDM as the best-fitting model and the proposed model as the second-best-fitting model were not adequately large.

Table 14

The Number of Replications of the Evidence Ratio of the Proposed Model to Each Alternative Model Being Greater than 55

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	AIC			BIC			DIC		
				Total	Vs L-LLM	Vs L-MCDM	Total	Vs L-LLM	Vs L-MCDM	Total	Vs L-LLM	Vs L-MCDM
100	0.6:0.4	0.3	-0.3	30	30	30	30	30	30	8	8	4
			0	30	30	29	29	29	29	5	5	5
			0.3	30	30	30	30	30	30	7	7	3
		0.7	-0.3	30	30	30	30	30	30	30	30	29
			0	30	30	30	30	30	30	30	30	30
			0.3	30	30	30	30	30	30	30	30	30
	0.8:0.2	0.3	-0.3	30	30	30	30	30	30	21	21	17
			0	30	30	30	30	30	30	20	20	18
			0.3	30	30	30	30	30	29	23	23	23
		0.7	-0.3	30	30	30	30	30	30	29	29	29
			0	30	30	30	30	30	30	30	30	30
			0.3	30	30	30	30	30	30	29	29	29
800	0.6:0.4	0.3	-0.3	30	30	30	30	30	30	17	17	16
			0	30	30	30	30	30	30	9	9	8
			0.3	30	30	30	30	30	30	7	7	6
		0.7	-0.3	30	30	30	30	30	30	30	30	30
			0	30	30	30	30	30	30	30	30	30
			0.3	30	30	30	30	30	30	30	30	30
	0.8:0.2	0.3	-0.3	30	30	30	30	30	30	29	29	29
			0	30	30	30	30	30	30	30	30	30
			0.3	30	30	30	30	30	30	29	29	29
		0.7	-0.3	30	30	30	30	30	30	30	30	30
			0	30	30	30	30	30	30	30	30	30
			0.3	30	30	30	30	30	30	30	30	30

Total=Total frequency of the LTA-longitudinal-MCDM having the smallest fit index among the three data-fitting models; Vs L-LLM=LTA-longitudinal-MCDM versus Longitudinal LLM; Vs L-MCDM= LTA-longitudinal-MCDM versus Longitudinal MCDM.

4.2 Recovery of the Person Parameters

The person parameters allow one to draw inferences about each individual person or the population and, thus, are directly related to the diagnostic information provided by the proposed model. In general, three categories of person parameters are examined, including the attribute mastery status (all the parameters relevant to α), skill implementation ability (all the parameters relevant to θ) and strategy choice (all the parameters relevant to *m*), as listed in Table 12.

4.2.1 Attribute mastery status

The recovery of the attribute mastery status ($\alpha_{jk}^{(t)}$) is evaluated with the attribute correct classification rate (ACCR) and profile correct classification rate (PCCR). Given that $\alpha_{jk}^{(t)}$ are estimated in all the three data-fitting models, the ACCRs and PCCRs are compared across the three models under all the simulated conditions in order to investigate the effects of ignoring the multiple-strategy scenarios on the attribute (profile) classification accuracy. As shown in Figure 9, the marginal mean ACCRs of LTA-longitudinal-MCDM by the levels of each manipulated factors are higher than 0.85, indicating that, on average, the attribute mastery status of over 85% of the simulees were correctly classified for each attribute using the proposed model. The average-across-replications ACCRs for each of the 24 simulated conditions are supplied in Appendix A. According to Figure 9, the marginal means of ACCR of Attribute 1 is lower for the Longitudinal LLM which ignore both between-person multiple strategies and within-person strategy shift than the other two models.

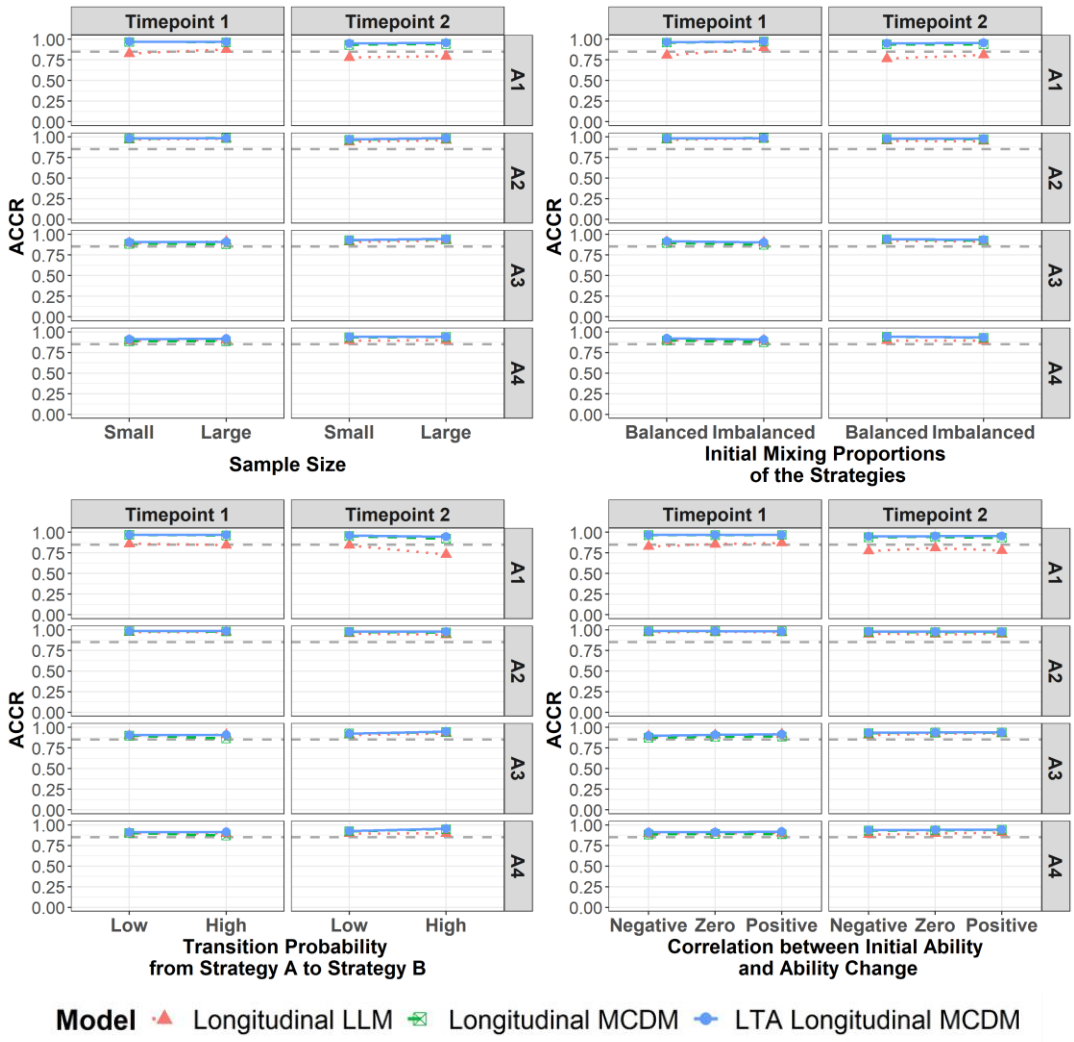


Figure 9. Marginal mean attribute correct classification rates (ACCRs) at each level of the manipulated factors. A1 to A4 indicate Attribute 1 to Attribute 4.

In terms of the attribute profile classification accuracy, the average-across-replications PCCRs of the LTA-longitudinal-MCDM are over 0.75 in all the simulated conditions (See Table 15), meaning that, on average, over 75% of the simulees' attribute profile (i.e., attribute mastery status patterns) were successfully recovered with the proposed model. In addition, as shown in Table 15, the LTA-longitudinal-MCDM demonstrates the highest PCCR among the three models under all the conditions.

Table 15

Attribute Profile Correct Classification Rate

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)} \Delta \theta}$	PCCR (Timepoint 1)			PCCR (Timepoint 2)		
				L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.533	0.763	**0.777**	0.560	0.777	**0.781**
			0	0.592	0.783	**0.793**	0.652	0.788	**0.801**
			0.3	0.643	0.787	**0.798**	0.620	0.794	**0.801**
		0.7	-0.3	0.532	0.721	**0.781**	0.484	0.790	**0.812**
			0	0.597	0.750	**0.802**	0.584	0.792	**0.834**
			0.3	0.642	0.745	**0.810**	0.536	0.787	**0.839**
	0.8:0.2	0.3	-0.3	0.640	0.761	**0.763**	0.634	0.757	**0.783**
			0	0.691	0.776	**0.788**	0.676	0.777	**0.796**
			0.3	0.702	0.764	**0.791**	0.626	0.769	**0.787**
		0.7	-0.3	0.623	0.700	**0.763**	0.516	0.773	**0.821**
			0	0.683	0.733	**0.793**	0.591	0.770	**0.821**
			0.3	0.706	0.734	**0.802**	0.548	0.748	**0.808**
800	0.6:0.4	0.3	-0.3	0.658	0.781	**0.805**	0.630	0.815	**0.821**
			0	0.670	0.785	**0.802**	0.650	0.819	**0.824**
			0.3	0.675	0.795	**0.811**	0.664	0.832	**0.836**
		0.7	-0.3	0.654	0.726	**0.804**	0.546	0.829	**0.852**
			0	0.667	0.731	**0.803**	0.568	0.832	**0.857**
			0.3	0.675	0.740	**0.813**	0.587	0.841	**0.865**
	0.8:0.2	0.3	-0.3	0.732	0.760	**0.788**	0.669	0.790	**0.810**
			0	0.737	0.761	**0.786**	0.690	0.798	**0.817**
			0.3	0.744	0.772	**0.797**	0.702	0.807	**0.828**
		0.7	-0.3	0.731	0.698	**0.788**	0.589	0.802	**0.845**
			0	0.739	0.705	**0.787**	0.605	0.802	**0.847**
			0.3	0.745	0.713	**0.800**	0.608	0.804	**0.857**

Note. PCCR=Profile Correct Classification Rate; L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM. The highest PCCR among the three data-fitting models is bolded under each condition at each time point.

To examine the effects of the manipulated factors on the attribute profile classification accuracy of the proposed model, the marginal mean PCCRs yielded from the proposed model are plotted against different levels of the manipulated factors, as shown in Figure 10. The slopes of the solid lines in Figure 10 are close to 0, implying that the manipulated factors have little effect on the attribute profile classification accuracy of the LTA-longitudinal-MCDM.

To sum up, the ACCR and PCCR results above suggest that ignoring between-person and/or within-person multiple strategies does correspond to a diminished accuracy in the recovery of the attribute (profile) classification. Nevertheless, on average, the attribute (profile) classification accuracy of the proposed model tends to be similar across different levels of the manipulated factors.

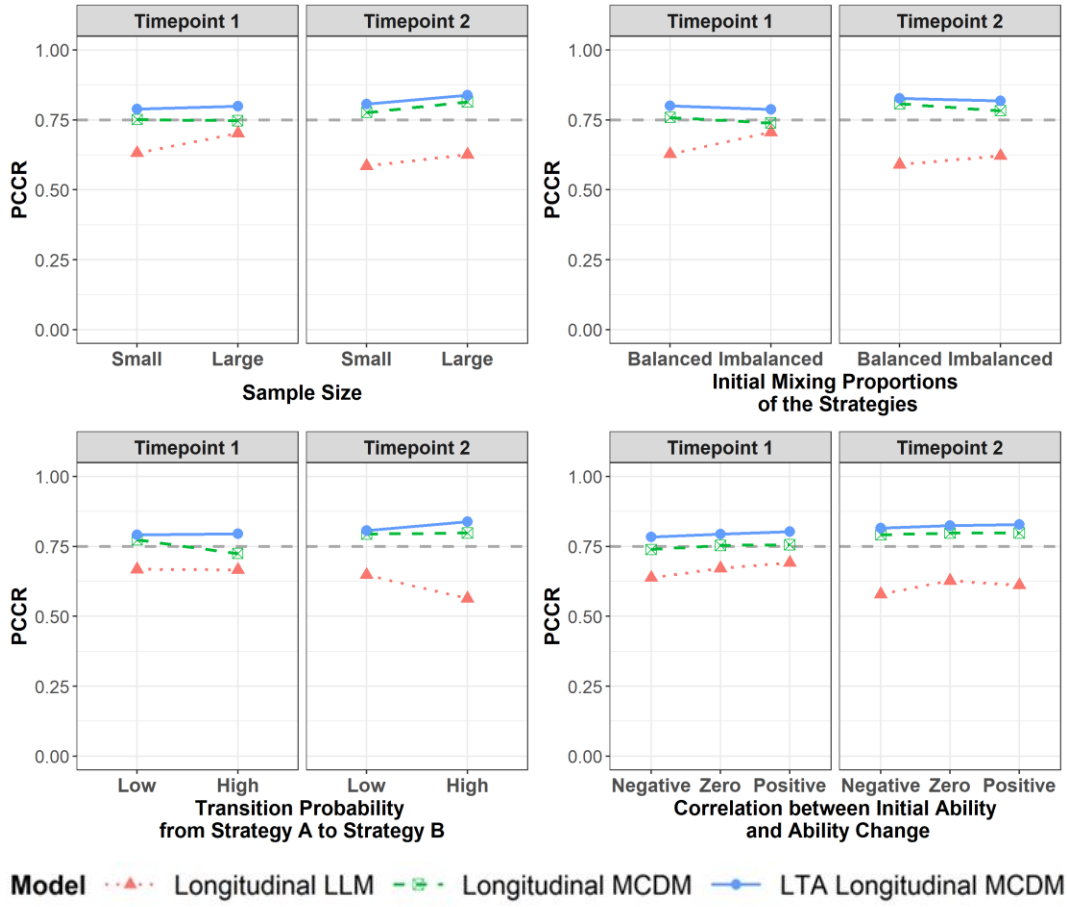


Figure 10. Marginal mean attribute profile correct classification rates (PCCRs) at each level of the manipulated factors.

4.2.2 Skill implementation ability

The skill implementation ability parameters include the initial ability parameter ($\theta_j^{(T_1)}$) and ability change parameter ($\Delta\theta_j$) and their corresponding mean vectors, $(\mu_{\theta^{(T_1)}} \quad \mu_{\Delta\theta})^T$, and variance-covariance matrix ($\boldsymbol{\Sigma}_{(\theta^*)}$). This study classifies the skill implementation ability parameters into the first-level and second-level parameters. The first-level parameters, including $\theta_j^{(T_1)}$ and $\Delta\theta_j$, are individual-specific. In contrast, the second-level parameters, including $(\mu_{\theta^{(T_1)}} \quad \mu_{\Delta\theta})^T$ and $\boldsymbol{\Sigma}_{(\theta^*)}$, delineate the distributions of the first-level parameters and, thus, are population-

specific. The bias, SE and RMSE are evaluated to examine the recovery of the ability parameters[6].

As described in Section 3.3.5, the mixed-effect ANOVAs were conducted to examine the effects of the data-fitting model type and the manipulated factors on the recovery of $\theta_j^{(T_1)}$ and $\Delta\theta_j$. In particular, investigating the effects of the data-fitting model type allow one to draw inferences about the impact of ignoring the multiple-strategy scenarios in the models on the model parameter recovery. Given that the mixed-effect ANOVAs' assumption of homogenous residual variances was violated but that the mixed-effect ANOVA results from groups with nearly equal sample sizes are robust to such assumption violation, mixed-effect ANOVAs were performed separately for the small sample size (J=100) and large sample size (J=800) conditions to investigate the effects of the data-fitting model type and its interaction with the other three manipulated factors (i.e., $\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$, $p_{M_B|M_A}$ and $\rho_{\theta^{(T_1)}\Delta\theta}$) on the recovery of $\theta_j^{(T_1)}$ and $\Delta\theta_j$. Specifically, the mixed-effect ANOVAs were set up as follows within each level of sample size: each individual ability parameter was treated as a subject, the bias/SE/RMSE of the ability parameter was treated as the measurement (i.e., dependent variable) taken on each subject, the data-fitting model type was treated as the repeated-measure factor (i.e., within-subject factor) and the three manipulated factors (i.e., the initial mixing proportions of the strategies, the latent

[6] The recoveries of the mean and variance parameters of the initial ability are not examined, as these parameters have been constrained at 0 and 1, respectively, for scale identification. For the same reason, only the SE and RMSE are examined for the initial ability parameter estimates.

transition probability from Strategy A to Strategy B, and the correlation between the initial ability and ability change) were treated as the between-subject factors.

The statistically significant effects (i.e., *p*-value <0.05) with at least a small effect size (i.e., partial $\eta^2 \geq 0.01$) are reported in the following sections. Since the interpretation of the lower-order interactions or main effects would be misleading if a higher-order interaction were significant, the following sections only visualize and elaborate the highest-order significant effects. Table 16 provides an overview of the highest-order significant effects found in the mixed-effect ANOVAs along with the effect sizes. In general, the significant two-way or three way interactions among the correlation between the initial ability and ability change (CORR), latent transition probability from Strategy A to Strategy B (TR_Prob) and the data-fitting model type (MODEL) are found on the recovery outcome measures of $\theta_j^{(T_1)}$ and $\Delta\theta_j$.

In addition, the three-way ANOVAs were performed on the recovery outcome measures of $\theta_j^{(T_1)}$ and $\Delta\theta_j$ of the LTA-longitudinal-MCDM to investigate the effects of the manipulated factors on the first-level ability parameter recovery of the proposed model. The three-way ANOVA results regarding $\theta_j^{(T_1)}$ and $\Delta\theta_j$ are presented at the end of Sections 4.2.2.1 and 4.2.2.2, respectively.

As for the recovery of the second-level person parameters, i.e., $(\mu_{\theta^{(T_1)}} \quad \mu_{\Delta\theta})^T$ and $\boldsymbol{\Sigma}_{(\boldsymbol{\theta}^*)}$, to which ANOVA was not applicable, the marginal means of the recovery outcome measures of these parameters by the levels of each manipulated factor are summarized and compared across the models. As the mean ($\mu_{\theta^{(T_1)}}$) and variance (

$\sigma^2_{\theta^{(T_1)}}$) of the initial ability parameter were constrained to be 0 and 1, respectively, the estimated second-level person parameters only include the mean ($\mu_{\Delta\theta}$) and variance ($\sigma^2_{\Delta\theta}$) of the ability change, and the covariance between the initial ability and ability change ($\sigma_{\theta^{(T_1)}\Delta\theta}$). Biases, SEs and RMSEs of the estimates of all the assessed parameters under all the simulated conditions are tabulated in Appendix A.

Table 16

Summary of Effect Sizes of the Highest-Order Significant Effects from the Mixed-Effect ANOVA on the Skill Implementation Ability Parameter Recovery

J	Effect	Initial Ability Parameter ($\theta_j^{(T_1)}$)	Ability Change Parameter ($\Delta\theta_j$)		
		SE	Bias	SE	RMSE
100	TR_Prob*MODEL		0.010		0.061
	CORR*TR_Prob*MODEL	0.014		0.058	
800	TR_Prob*MODEL		0.012	0.259	0.040
	CORR*MODEL	0.024			

Effect Size	Small (0.01≤partial η^2 <0.06)	Medium (0.06≤partial η^2 <0.14)	Large (partial η^2 ≥0.14)

Note. *J*=Sample size; CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type. The values in the cells are partial η^2.

4.2.2.1 Initial ability estimates

As an overview, the data-fitting model type interact with one or multiple manipulated factors to affect the random error of $\hat{\theta}_j^{(T_1)}$ quantified by SE in both the small sample size (*J*=100) and large sample size (*J*=800) conditions, according to the "Initial Ability Parameter" column in Table 16. In particular, when the sample size is small (*J*=100), a significant three-way interaction among CORR, TR_Prob and MODEL on the SE of $\hat{\theta}_j^{(T_1)}$ was found with a small effect size (*F*=8.54, *p*<0.001,

partial η^2=0.014), as shown in Table 17. Moreover, the mixed-effect ANOVA results indicate a large two-way interaction effect of CORR*MODEL (F=197.23, p<0.001, partial η^2=0.249), a medium main effect of MODEL (F=116.48, p<0.001, partial η^2 =0.089) and a small two-way interaction of TR_Prob*MODEL (F=13.81, p<0.001, partial η^2=0.011) on the SE of $\hat{\theta}_j^{(T_1)}$. The patterns of the highest-order significant effect, the three-way interaction of CORR*TR_Prob*MODEL on the SE of $\hat{\theta}_j^{(T_1)}$, are visualized in Figure 11. Specifically, the left, middle and right panels of Figure 11 display the interactions of MODEL*TR_Prob on the SE of $\hat{\theta}_j^{(T_1)}$ under the conditions with negative ($\rho_{\theta^{(T_1)}\Delta\theta} = -0.3$), zero ($\rho_{\theta^{(T_1)}\Delta\theta} = 0$) and positive ($\rho_{\theta^{(T_1)}\Delta\theta} = 0.3$) true correlations between the initial ability and ability change, respectively. Different line patterns in Figure 11 represent different data-fitting models. When the true correlation between the initial ability and ability change is negative ($\rho_{\theta^{(T_1)}\Delta\theta} = -0.3$), the Longitudinal LLM produces the highest mean SE of $\hat{\theta}_j^{(T_1)}$ among the three data-fitting models, followed by the LTA-longitudinal-MCDM, regardless of the levels of the strategy transition probability (i.e., either $p_{M_B|M_A} = 0.3$ or $p_{M_B|M_A} = 0.7$); the discrepancies in the mean SE of $\hat{\theta}_j^{(T_1)}$ among the three models vary across the levels of the strategy transition probability. In contrast, when the true correlation between the initial ability and ability change is positive ($\rho_{\theta^{(T_1)}\Delta\theta} = 0.3$), the Longitudinal LLM produces the lowest mean SE of $\hat{\theta}_j^{(T_1)}$ among the three models. When there is no

correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta} = 0$), the three models yield similar mean SEs of $\hat{\theta}_j^{(T_1)}$.

Table 17
Significant Effects in the Mixed-Effect ANOVA Results of the SE of the Initial Ability Estimates (J=100)

Source	SE of $\hat{\theta}_j^{(T_1)}$		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)			
MODEL	116.48	<0.001	0.089
TR_Prob*MODEL	13.81	<0.001	0.011
CORR*MODEL	197.23	<0.001	0.249
CORR*TR_Prob*MODEL	8.54	<0.001	0.014

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.

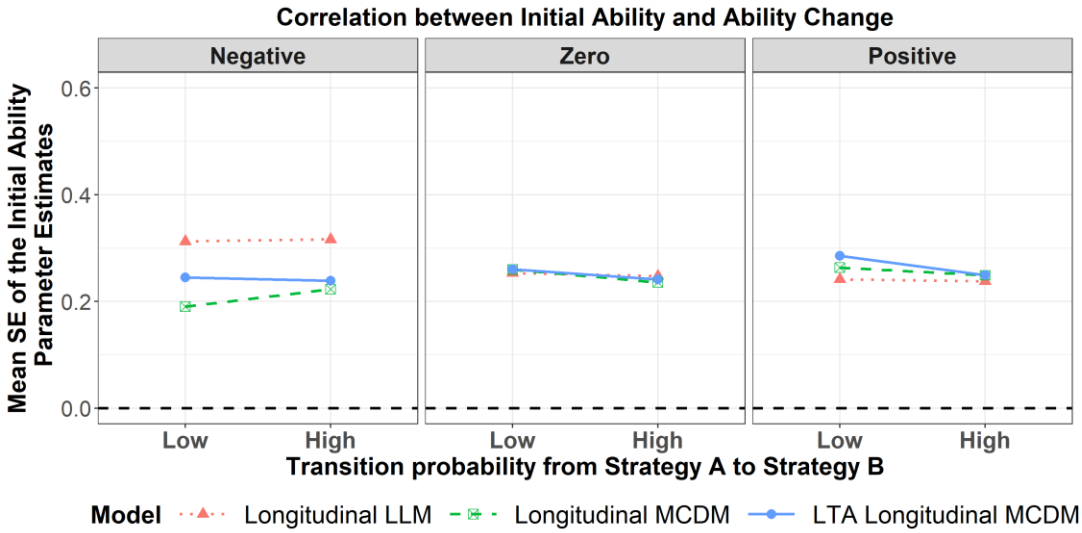


Figure 11. Significant three-way interaction of CORR*TR_Prob*MODEL on the SE of the initial ability parameter estimates, $\hat{\theta}_j^{(T_1)}$, in the conditions of small sample size (*J*=100). [Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.]

When the sample size is large (J=800), there is a significant CORR*MODEL interaction on the SE of $\hat{\theta}_j^{(T_1)}$ with a small effect size (F=115.45, p<0.001, partial η^2 =0.024), as shown in Table 18. In addition, a large main effect of MODEL (F=4625.72, p<0.001, partial η^2=0.325) is found on the SE of $\hat{\theta}_j^{(T_1)}$. The significant two-way interaction of CORR*MODEL on the SE of $\hat{\theta}_j^{(T_1)}$ is plotted in Figure 12. It can be seen that, at each level of the true correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$ = -0.3, 0 or 0.3), the LTA-longitudinal-MCDM produces the lowest mean SE of $\hat{\theta}_j^{(T_1)}$ among the three data-fitting models, followed by the Longitudinal MCDM. However, the magnitude of differences in the SE of $\hat{\theta}_j^{(T_1)}$ among the three models varies across the levels of the true correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$).

Table 18
Significant Effects in the Mixed-Effect ANOVA Results of the SE of the Initial Ability Estimates (J=800)

Source	SE of $\hat{\theta}_j^{(T_1)}$		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)			
MODEL	4625.72	<0.001	0.325
CORR*MODEL	115.45	<0.001	0.024

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); MODEL=Data-fitting model type.

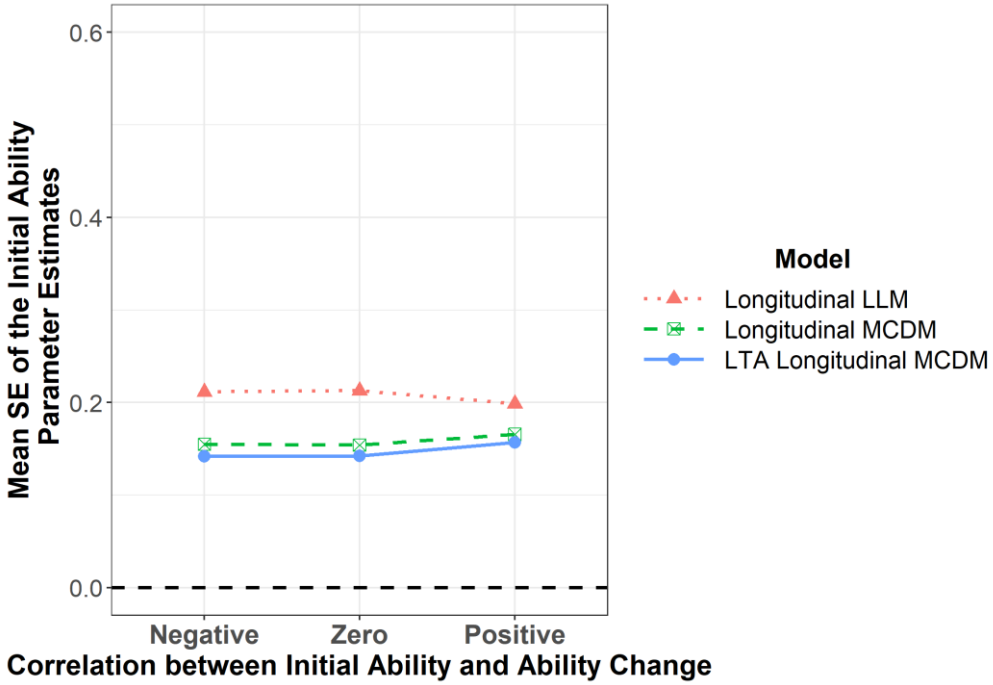


Figure 12. Significant two-way interaction of CORR*MODEL on the SE of the initial ability parameter estimates, $\hat{\theta}_j^{(T_1)}$, in the conditions of large sample size (J=800). [Note. MODEL=Data-fitting model type; CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$).]

The three-way ANOVAs were performed to investigate the effects of the manipulated factors on the initial ability parameter recovery from the proposed model. As shown in Table 19, only significant effects with small effect sizes are found of the correlation between the initial ability and ability change (CORR) and the transition probability from Strategy A to Strategy B (TR_Prob) on the SE of $\hat{\theta}_j^{(T_1)}$ of the proposed model. An inspection of the marginal means found that, when the sample size is small, the mean SE of $\hat{\theta}_j^{(T_1)}$ is higher in the lower strategy transition probability conditions ($p_{M_B|M_A}$=0.3). No significant effect is found of the manipulated factors on the RMSE of $\hat{\theta}_j^{(T_1)}$ of the proposed model.

Table 19

Significant Effects in the Three-Way ANOVA Results of the Recovery of the Initial Ability Parameter from the LTA-longitudinal-MCDM

J	Source	SE of $\hat{\theta}_j^{(T_1)}$	
		p-value	Partial η^2
100	CORR	<0.001	0.013
	TR_Prob	<0.001	0.014
800	CORR	<0.001	0.016

Effect Size	Small (0.01≤partial η^2 <0.06)	Medium (0.06≤partial η^2 <0.14)	Large (partial η^2 ≥0.14)

Note. *J*=Sample size; CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$).

4.2.2.2 Ability change estimates

Overall, as shown by the columns under "Ability Change Parameter" in Table 16, the data-fitting model interacted with one of the manipulated factors, the latent transition probability from Strategy A to Strategy B (TR_Prob), to affect the bias and RMSE of $\Delta\hat{\theta}_j$ in both the small sample size (*J*=100) and large sample size (*J*=800) conditions. However, under conditions of different sample sizes, the data-fitting model interacted with different sets of manipulated factors to affect the SE of $\Delta\hat{\theta}_j$.

In particular, when the sample size is small (*J*=100), significant TR_Prob*MODEL interactions are found on both the bias (*F*=12.30, *p*<0.001, partial η^2 =0.010) and RMSE (*F*=77.13, *p*<0.001, partial η^2 =0.061) of $\Delta\hat{\theta}_j$, according to Table 20. In addition, MODEL is found to have medium main effects on both the bias (*F*=85.51, *p*<0.001, partial η^2 =0.067) and RMSE (*F*=126.37, *p*<0.001, partial η^2 =0.096) of $\Delta\hat{\theta}_j$. CORR has small main effects on both bias (*F*=10.73, *p*<0.001, partial η^2 =0.018) and RMSE (*F*=11.24, *p*<0.001, partial η^2 =0.019) of $\Delta\hat{\theta}_j$. Figure

13 displays the patterns of the significant two-way interactions of TR_Prob and MODEL on the bias and RMSE of $\Delta\hat{\theta}_j$. The Longitudinal LLM that ignores both between-person multiple strategies and within-person strategy shift yield the highest absolute mean bias and RMSE of $\Delta\hat{\theta}_j$ at each level of the strategy transition probability (i.e., either $p_{M_B|M_A} = 0.3$ or $p_{M_B|M_A} = 0.7$). Nevertheless, the model discrepancies in terms of the mean bias and RMSE of $\Delta\hat{\theta}_j$ appear to be larger in the higher strategy transition probability conditions ($p_{M_B|M_A} = 0.7$).

Table 20
Significant Effects in the Mixed-Effect ANOVA Results of the Bias and RMSE of the Ability Change Estimates (J=100)

Source	Bias of $\Delta\hat{\theta}_j$			RMSE of $\Delta\hat{\theta}_j$		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)						
MODEL	85.51	<0.001	0.067	126.37	<0.001	0.096
TR_Prob*MODEL	12.30	<0.001	0.010	77.13	<0.001	0.061
Between-Subject Effects						
CORR	10.73	<0.001	0.018	11.24	<0.001	0.019

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$);
TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.

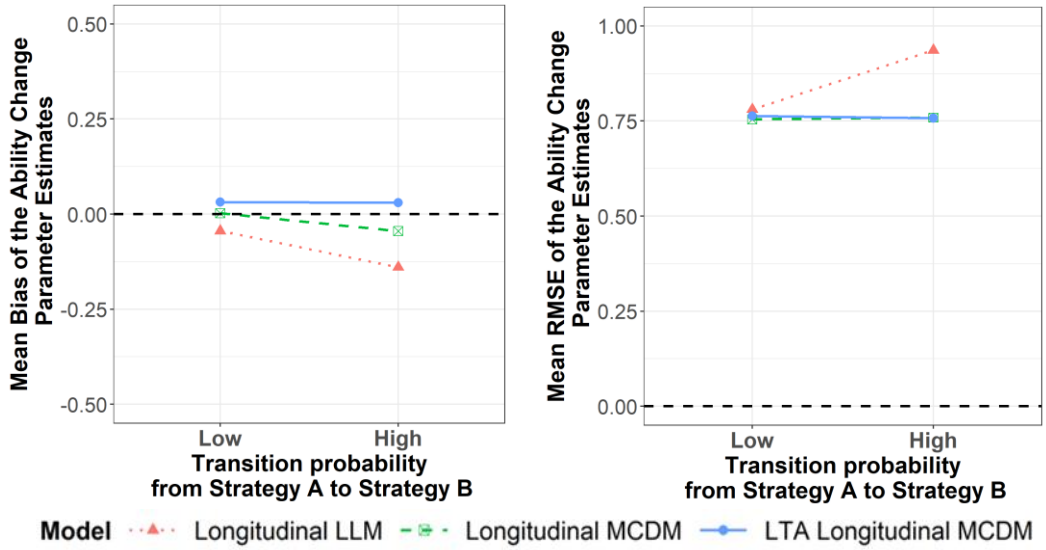


Figure 13. Significant two-way interactions of TR_Prob*MODEL on the bias and RMSE of the ability change parameter estimates, $\Delta\hat{\theta}_j$, in the conditions of small sample size (J=100). [Note. TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.]

As for the random errors of $\Delta\hat{\theta}_j$ in the small sample size conditions (J=100), a significant three-way interaction is found among CORR, TR_Prob and MODEL on the SE of $\Delta\hat{\theta}_j$ (F=36.48, p<0.001, partial η^2=0.058) as shown in Table 21. The lower-order interactions and main effects of these three factors (i.e., CORR, TR_Prob and MODEL) on the SE of $\Delta\hat{\theta}_j$ are also significant according to Table 21. The patterns of the highest-order significant interaction, i.e., the three-way interaction of CORR*TR_Prob*MODEL on the SE of $\Delta\hat{\theta}_j$, are displayed in Figure 14. The Longitudinal LLM tends to produce higher mean SE of $\Delta\hat{\theta}_j$ than the other two models except in the conditions with a positive correlation between the initial ability and ability change and a low strategy transition probability (i.e., $\rho_{\theta^{(T_1)}\Delta\theta} = 0.3$ and $p_{M_B|M_A} = 0.3$).

Table 21

Significant Effects in the Mixed-Effect ANOVA Results of the SE of the Ability Change Estimates (J=100)

Source	SE of $\Delta\hat{\theta}_j$		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)			
MODEL	1386.62	<0.001	0.325
CORR*MODEL	342.75	<0.001	0.366
TR_Prob*MODEL	347.21	<0.001	0.226
CORR*TR_Prob*MODEL	36.48	<0.001	0.058
Between-Subject Effects			
CORR	99.48	<0.001	0.143
TR_Prob	32.13	<0.001	0.026
CORR*TR_Prob	19.29	<0.001	0.031

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.

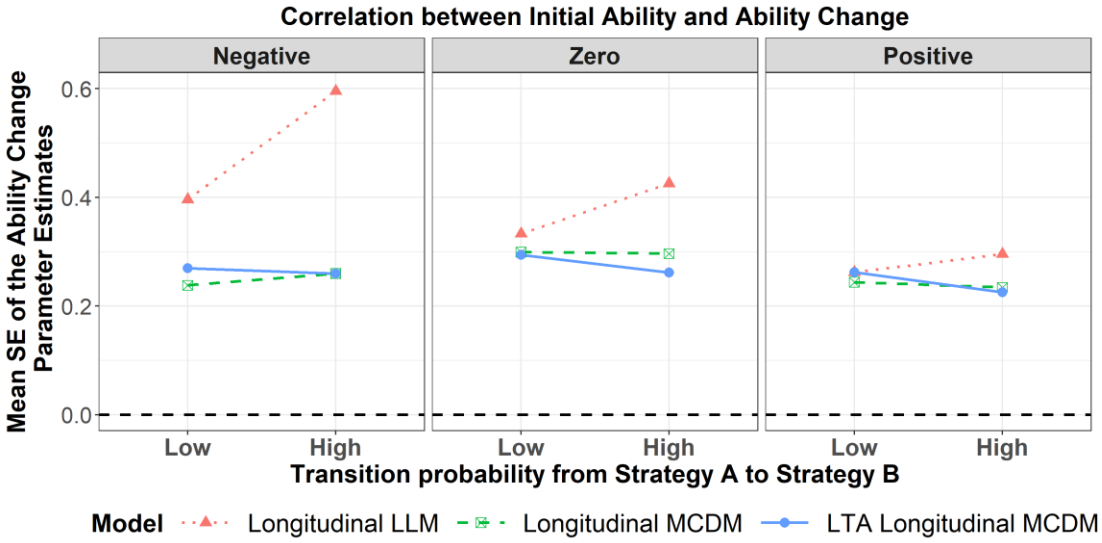


Figure 14. Significant three-way interactions of CORR*TR_Prob*MODEL on the SE of the ability change parameter estimates, $\Delta\hat{\theta}_j$, in the conditions of small sample size (*J*=100). [Note: MODEL=Data-fitting model type; TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$).]

In the large sample size conditions (J=800), significant two-way interactions are found of TR_Prob and MODEL on the bias (F=119.73, p<0.001, partial η^2 =0.012), SE (F=3355.56, p<0.001, partial η^2=0.259) and RMSE (F=403.31, p<0.001, partial η^2=0.040) of $\Delta\hat{\theta}$, according to Tables 22 and 23. Moreover, MODEL has a small main effect on RMSE (F=539.48, p<0.001, partial η^2=0.053), a medium main effect on bias (F=796.46, p<0.001, partial η^2=0.077) and a large main effect on SE (F=11254.17, p<0.001, partial η^2=0.540) of $\Delta\hat{\theta}$. Figure 15 plots the significant two-way interactions of TR_Prob*MODEL on the bias, SE and RMSE of $\Delta\hat{\theta}_j$. The upper left panel of Figure 15 shows that the mean biases of $\Delta\hat{\theta}_j$ produced by the LTA-longitudinal-MCDM are close to 0 regardless of the strategy transition probability (i.e., either $p_{M_B|M_A}=0.3$ or $p_{M_B|M_A}=0.7$), while those produced by the Longitudinal LLM and Longitudinal MCDM are negative, the magnitudes of which increase as the strategy transition probability increases. Such results imply that $\Delta\theta_j$ tends to be underestimated by the models that ignore multiple-strategy scenarios. The upper right panel of Figure 15 indicates that the Longitudinal LLM produces higher mean SEs of $\Delta\hat{\theta}_j$ than the other two models do at both levels of strategy transition probability (i.e., either $p_{M_B|M_A}=0.3$ or $p_{M_B|M_A}=0.7$), implying that ignoring within-person strategy shift may result in an increase in the random errors of $\Delta\hat{\theta}_j$. Nevertheless, the magnitude of the difference in the mean SE of $\Delta\hat{\theta}_j$ between the Longitudinal LLM and the other two models appears to be larger in the higher

strategy transition probability conditions ($p_{M_B|M_A} = 0.7$). The lower left panel of Figure 15 shows that the Longitudinal LLM yields higher mean RMSE of $\Delta\hat{\theta}_j$ than the other two models in the high strategy transition probability conditions ($p_{M_B|M_A} = 0.7$), while the three models yield similar mean RMSEs of $\Delta\hat{\theta}_j$ in the low strategy transition conditions ($p_{M_B|M_A} = 0.3$).

Table 22
Significant Effects in the Mixed-Effect ANOVA Results of the Bias and RMSE of the Ability Change Estimates (J=800)

Source	Bias of $\Delta\hat{\theta}_j$			RMSE of $\Delta\hat{\theta}_j$		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)						
MODEL	796.46	<0.001	0.077	539.48	<0.001	0.053
TR_Prob*MODEL	119.72	<0.001	0.012	403.31	<0.001	0.040

Note. TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.

Table 23
Significant Effects in the Mixed-Effect ANOVA Results of the SE of the Ability Change Estimates (J=800)

Source	SE of $\Delta\hat{\theta}_j$		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)			
MODEL	11254.17	<0.001	0.540
TR_Prob*MODEL	3355.56	<0.001	0.259
Between-Subject Effects			
CORR	99.36	<0.001	0.020
TR_Prob	1066.11	<0.001	0.100

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type.

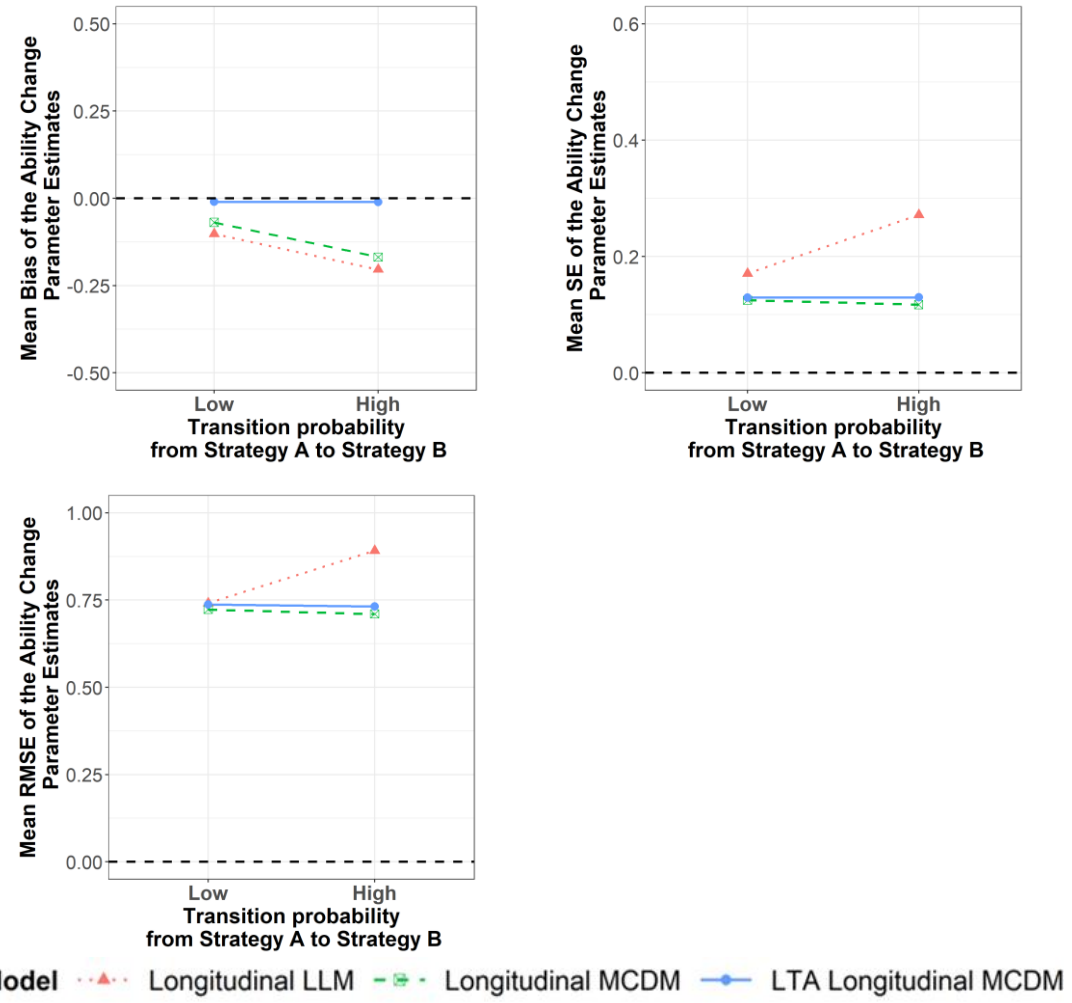


Figure 15. Significant two-way interactions of TR_Prob*MODEL on the bias, SE and RMSE of the ability change parameter estimates, $\Delta\hat{\theta}_j$, in the conditions of large sample size (J=800). [Note: MODEL=Data-fitting model type; TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$).]

Similar to the findings about the initial ability parameter recovery of the proposed model, the three-way ANOVA results on the ability change parameter indicate that the correlation between the initial ability and ability change (CORR) and transition probability from Strategy A to Strategy B (TR_Prob) have significant effects on the SE of $\Delta\hat{\theta}_j$ from the LTA-longitudinal-MCDM (See Table 24). An inspection in the marginal mean SEs indicated that, when the sample size is small, the

mean SE of $\Delta\hat{\theta}_j$ is higher in the lower strategy transition probability conditions ($p_{M_B|M_A}$ =0.3). Additionally, in the small sample size conditions, the correlation between the initial ability and ability change (CORR) has significant effects on the bias and RMSE of $\Delta\hat{\theta}_j$ from the proposed model.

Table 24
Significant Effects in the Three-Way ANOVA Results of the Recovery of the Ability Change Parameter from the LTA-longitudinal-MCDM

J	Source	Bias of $\Delta\hat{\theta}_j$		SE of $\Delta\hat{\theta}_j$		RMSE of $\Delta\hat{\theta}_j$	
		p-value	Partial η^2	p-value	Partial η^2	p-value	Partial η^2
100	CORR	<0.001	0.019	<0.001	0.025	<0.001	0.017
	TR_Prob			<0.001	0.028		
800	CORR			<0.001	0.021		

Effect Size	Small (0.01≤partial η^2 <0.06)	Medium (0.06≤partial η^2 <0.14)	Large (partial η^2 ≥0.14)

Note. J=Sample size; CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$).

4.2.2.3 The mean estimate of the ability change

To examine how the under-specification of the multiple-strategy scenarios in the model affect the recovery of the mean ability change parameter, $\mu_{\Delta\theta}$, under various simulated conditions, the mean bias, SE and RMSE of $\hat{\mu}_{\Delta\theta}$ are plotted against each level of the manipulated factors for each data-fitting model, as shown in Figures 16, 17 and 18. The impact of ignoring the multiple-strategy scenarios on the recovery of $\mu_{\Delta\theta}$ can be inferred by comparing the recovery outcome measures of $\mu_{\Delta\theta}$ across different data-fitting model types which are manifested as different line patterns in Figures 16, 17 and 18. Overall, the systematic errors of the $\hat{\mu}_{\Delta\theta}$ quantified by bias are

sensitive to the data-fitting model type, compared to its random errors quantified by SE that are relatively invariant across different models. To be specific, Figure 16 indicate that, the marginal mean biases of $\hat{\mu}_{\Delta\theta}$ from the LTA-longitudinal-MCDM are closer to 0 than those from the other two models, when averaged by sample size (J), initial mixing proportions of the strategies ($\pi_{m_1}^{(1)}$) or the transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$). Further, all the marginal mean biases of $\hat{\mu}_{\Delta\theta}$ estimated from the Longitudinal LLM and Longitudinal MCDM are negative, implying that, on average, $\mu_{\Delta\theta}$ tends to be underestimated by the models that ignore the multiple-strategy scenarios in the simulated conditions. However, the pattern of the marginal biases of $\hat{\mu}_{\Delta\theta}$ estimated from the LTA-longitudinal-MCDM is less consistent when averaged by different levels of correlation between the initial ability and ability change: On average, when the initial ability and ability change are positively correlated ($\rho_{\theta^{(T_1)}\Delta\theta} = 0.3$), $\mu_{\Delta\theta}$ tends to be underestimated by the LTA-longitudinal-MCDM; when the correlation between the initial ability and ability change is negative ($\rho_{\theta^{(T_1)}\Delta\theta} = -0.3$) or zero ($\rho_{\theta^{(T_1)}\Delta\theta} = 0$), $\mu_{\Delta\theta}$ tends to be overestimated by the LTA-longitudinal-MCDM. In contrast, the mean SEs of $\hat{\mu}_{\Delta\theta}$, averaged by any one of the four manipulated factors, are similar across the three data-fitting models according to Figure 17. The mean RMSEs of $\hat{\mu}_{\Delta\theta}$ displayed in Figure 18 reflect the magnitudes of systematic and random errors, as a whole, of $\hat{\mu}_{\Delta\theta}$. The marginal mean RMSEs of $\hat{\mu}_{\Delta\theta}$ estimated from the LTA-longitudinal-MCDM are the smallest among the three data-fitting models, which are followed by those from the

Longitudinal MCDM, at all the levels of the manipulated factors except at the small sample size level (J=100).

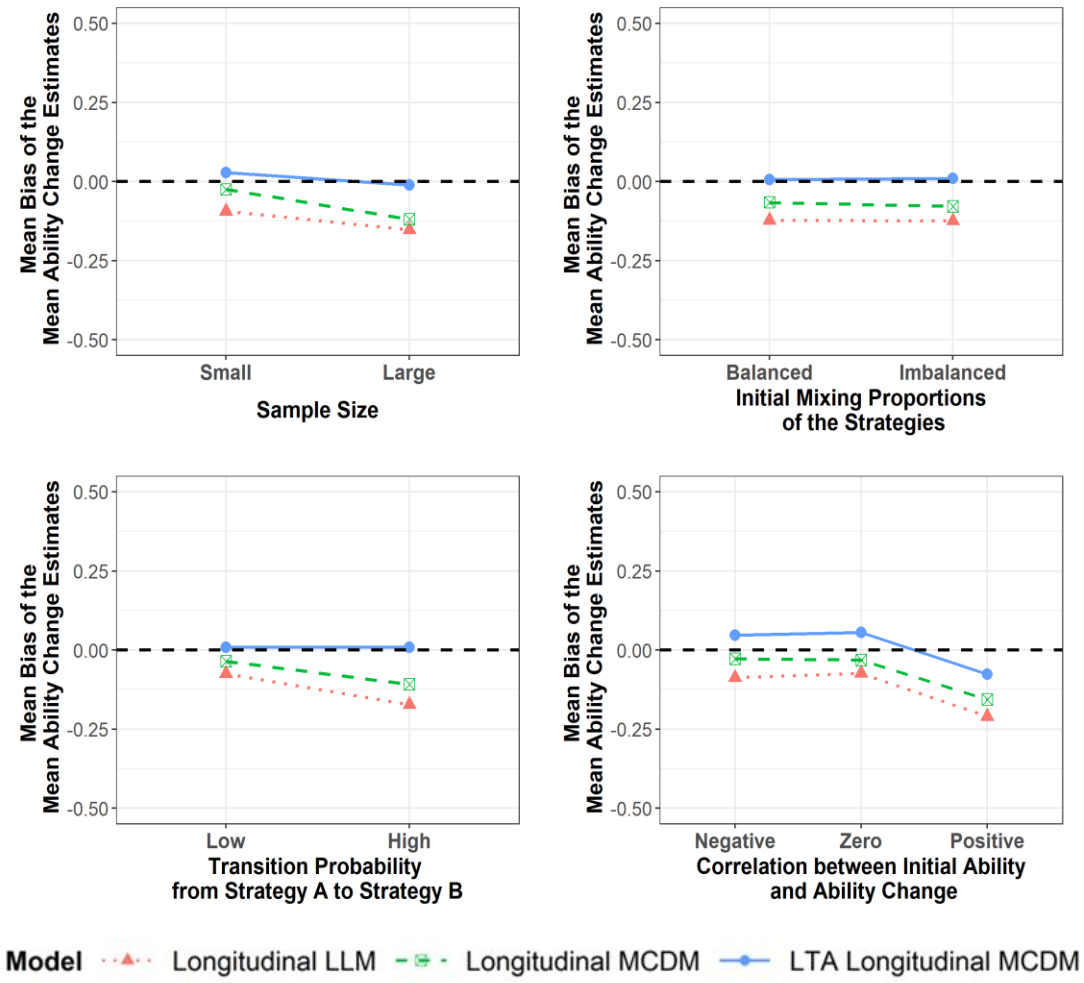


Figure 16. Marginal mean bias of the mean ability change parameter estimates, $\hat{\mu}_{\Delta\theta}$, at each level of the manipulated factors.

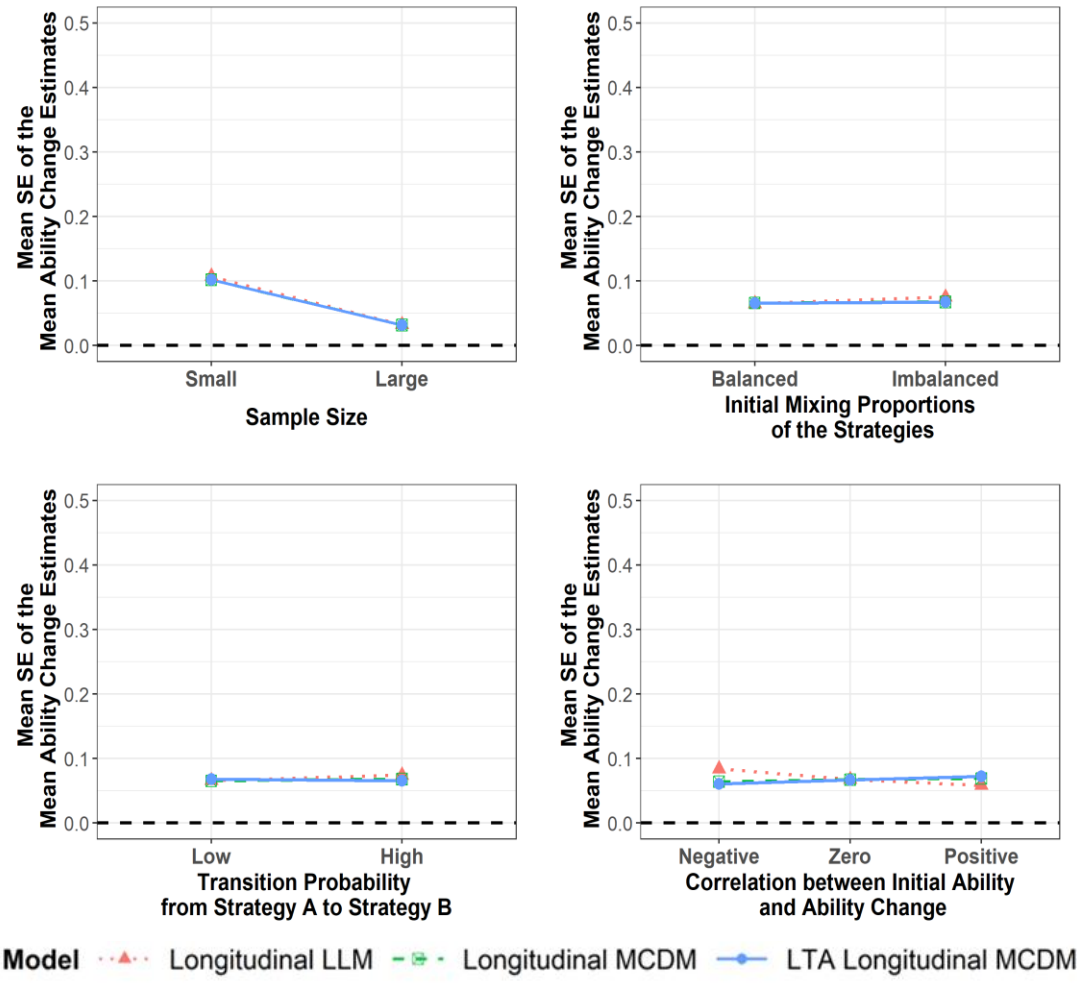


Figure 17. Marginal mean SE of the mean ability change parameter estimates, $\hat{\mu}_{\Delta\theta}$, at each level of the manipulated factors.

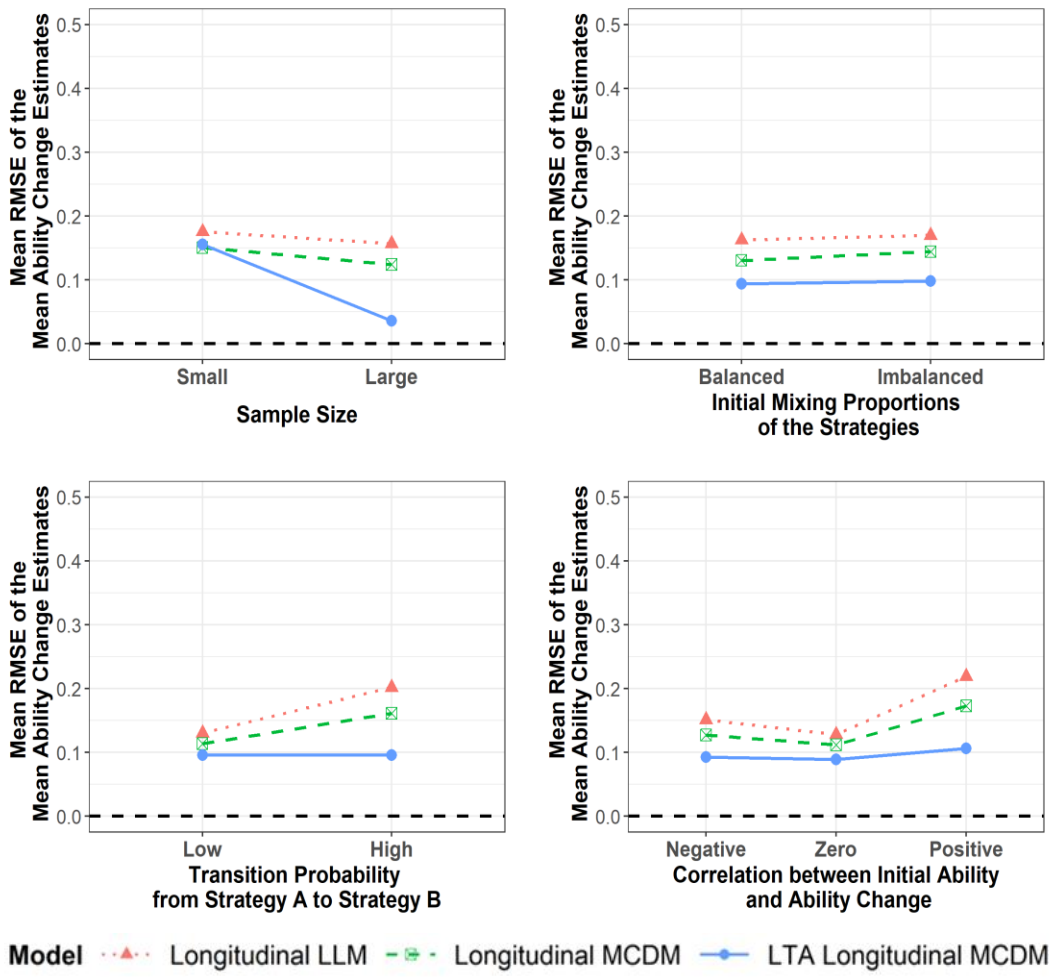


Figure 18. Marginal mean RMSE of the mean ability change parameter estimates, $\hat{\mu}_{\Delta\theta}$, at each level of the manipulated factors.

To examine how the manipulated factors affect the recovery of $\mu_{\Delta\theta}$ from the proposed model, the LTA-longitudinal-MCDM, the trends of the solid lines in Figures 16, 17 and 18 are further inspected. While the absolute value of the mean bias of $\hat{\mu}_{\Delta\theta}$ from the proposed model does not appear to differ significantly between different levels of sample size (See Figure 16), the mean SE and RMSE of $\hat{\mu}_{\Delta\theta}$ of the proposed model are notably higher in the smaller sample size conditions (See Figures 17 and 18).

4.2.2.4 The variance estimate of the ability change

As for the variance estimate of the ability change, $\hat{\sigma}^2_{\Delta\theta}$, both its systematic errors quantified by bias and the random errors quantified by SE are sensitive to the data-fitting model type. In particular, Figures 19 and 20 show that the absolute values of all the marginal mean biases and SEs of $\hat{\sigma}^2_{\Delta\theta}$ estimated from the Longitudinal LLM that ignores between-person multiple strategies are greater than those estimated from the models that consider between-person multiple strategies (i.e., the LTA-longitudinal-MCDM and Longitudinal MCDM). This observation implies that both the average systematic errors and random errors of $\hat{\sigma}^2_{\Delta\theta}$ increase when between-person multiple strategies are ignored in the model. Further, on average, $\sigma^2_{\Delta\theta}$ tends to be overestimated when between-person multiple strategies are ignored, as indicated by the positive marginal mean bias yielded by the Longitudinal LLM shown in Figure 19. Nevertheless, the mean biases and SEs of $\hat{\sigma}^2_{\Delta\theta}$ from the LTA-longitudinal-MCDM are similar to those from the Longitudinal MCDM, suggesting that the average systematic errors and random errors of $\hat{\sigma}^2_{\Delta\theta}$ are fairly robust to the neglect of the within-person strategy shift in the simulated conditions. A possible reason for such robustness of $\hat{\sigma}^2_{\Delta\theta}$ could be that, in the simulated conditions, the extent of Q-matrix misspecification resulted from ignoring the within-person strategy shift is not large enough to significantly affect the bias or SE of $\hat{\sigma}^2_{\Delta\theta}$. However, future study is needed to verify this hypothesis, as little study has been done currently to investigate the effects of Q-matrix misspecification on the recovery of the variance parameter of the higher-order latent trait in a higher-order CDM. In terms of the mean RMSEs of

$\hat{\sigma}^2_{\Delta\theta}$, those of the LTA-longitudinal-MCDM are the smallest among the three data-fitting models at all the levels of all the manipulated factors, according to Figure 21.

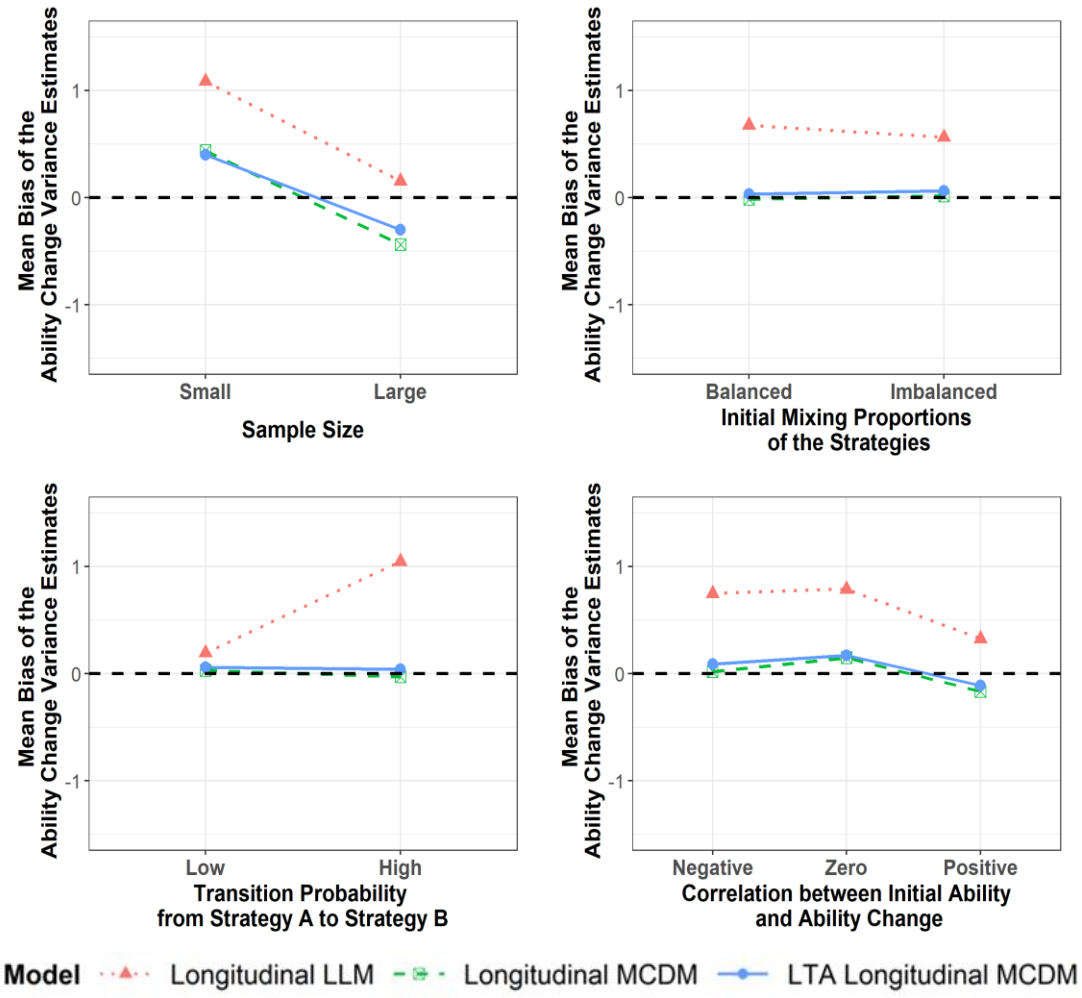


Figure 19. Marginal mean bias of the variance estimates of the ability change, $\hat{\sigma}^2_{\Delta\theta}$, at each level of the manipulated factors.

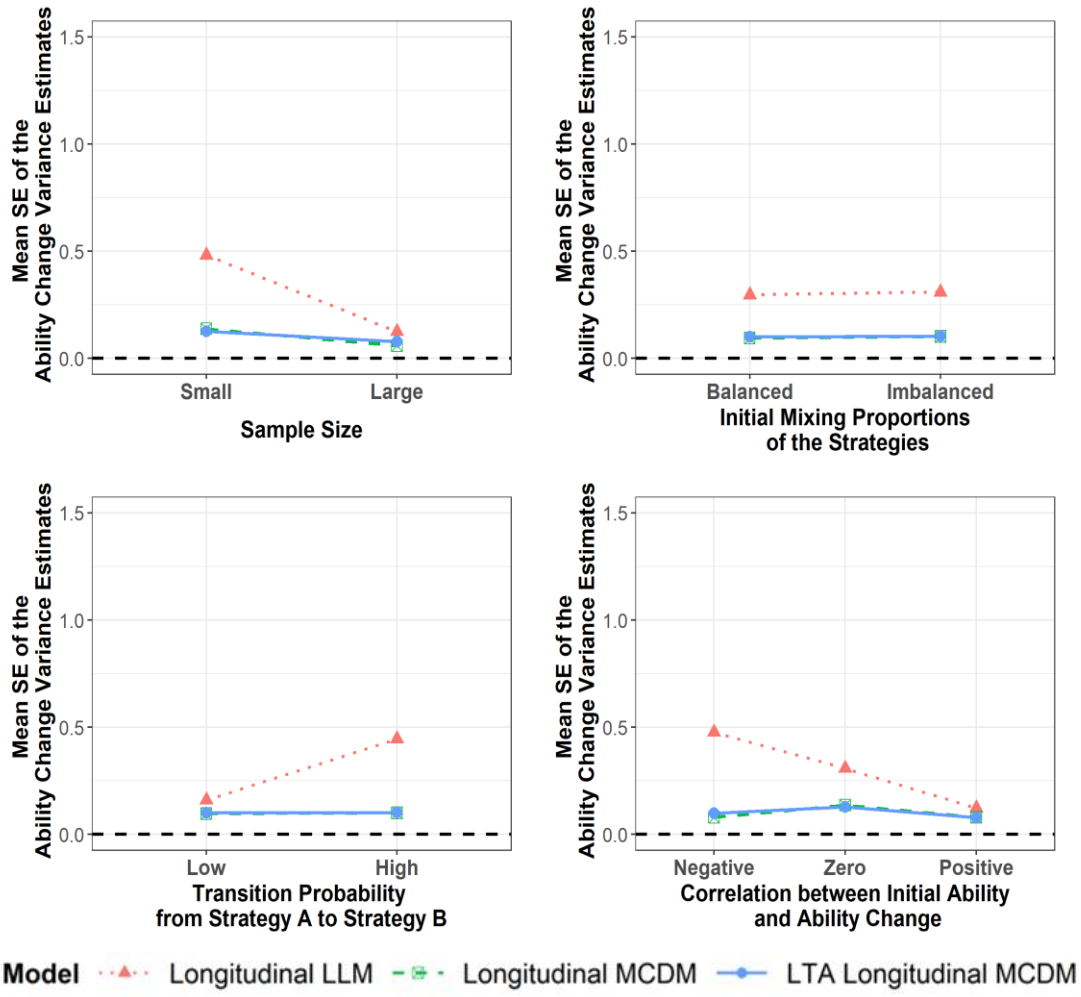


Figure 20. Marginal mean SE of the variance estimates of the ability change, $\hat{\sigma}^2_{\Delta\theta}$, at each level of the manipulated factors.

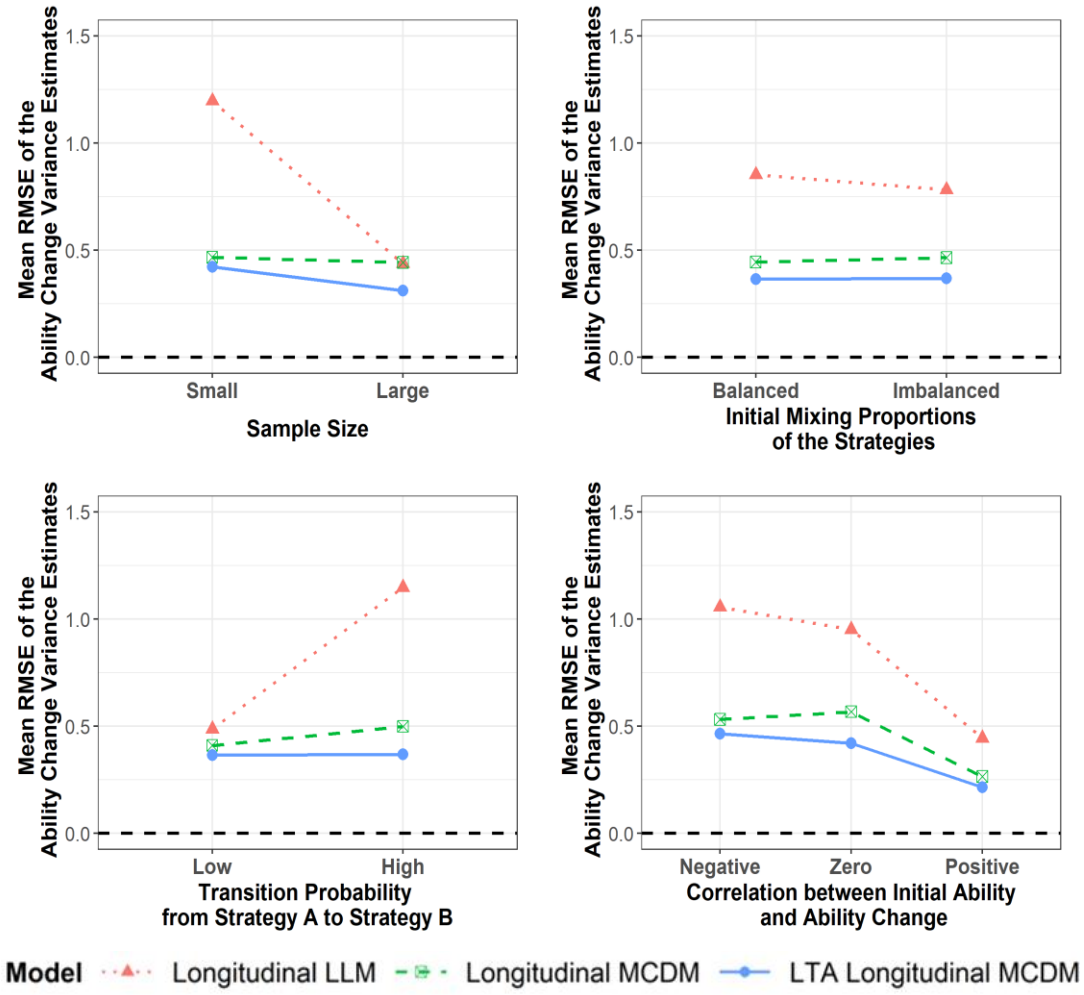


Figure 21. Marginal mean RMSE of the variance estimates of the ability change, $\hat{\sigma}^2_{\Delta\theta}$, at each level of the manipulated factors.

The trends of the solid lines in Figures 19, 20 and 21 indicate that, compared to the other manipulated factors, the sample size and the correlation between the initial ability and ability change have more significant effects on the bias, SE and RMSE of $\hat{\sigma}^2_{\Delta\theta}$ from the LTA-longitudinal-MCDM. In particular, the absolute value of the mean bias, SE and RMSE of $\hat{\sigma}^2_{\Delta\theta}$ from the LTA-longitudinal-MCDM is lower in the conditions with a larger sample size and a positive true correlation between the initial ability and ability change.

4.2.2.5 The covariance estimate between the initial ability and ability change

On average, the covariance between and ability and ability change, $\sigma_{\theta^{(T_1)}\Delta\theta}$, tends to be overestimated by all the three data-fitting models, according to the positive mean biases of $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$ shown in Figure 22. The marginal SEs of $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$ are similar across the three data-fitting models at all the levels of the manipulated factors (see Figure 23), while the marginal RMSEs of $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$ are relatively less consistent across the data-fitting models (See Figure 24). Obtaining an unbiased estimate of the covariance between the initial ability and ability change has been found to be challenging by previous longitudinal studies (e.g., Embretson, 1991), and results from this study indicate that the estimation of $\sigma_{\theta^{(T_1)}\Delta\theta}$ in the proposed model is not an exception. Therefore, cautions should be taken when drawing any diagnostic inferences from $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$. However, no evidence has been found in this study showing that the inaccurate $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$ hinders the accuracy the other person parameter estimates.

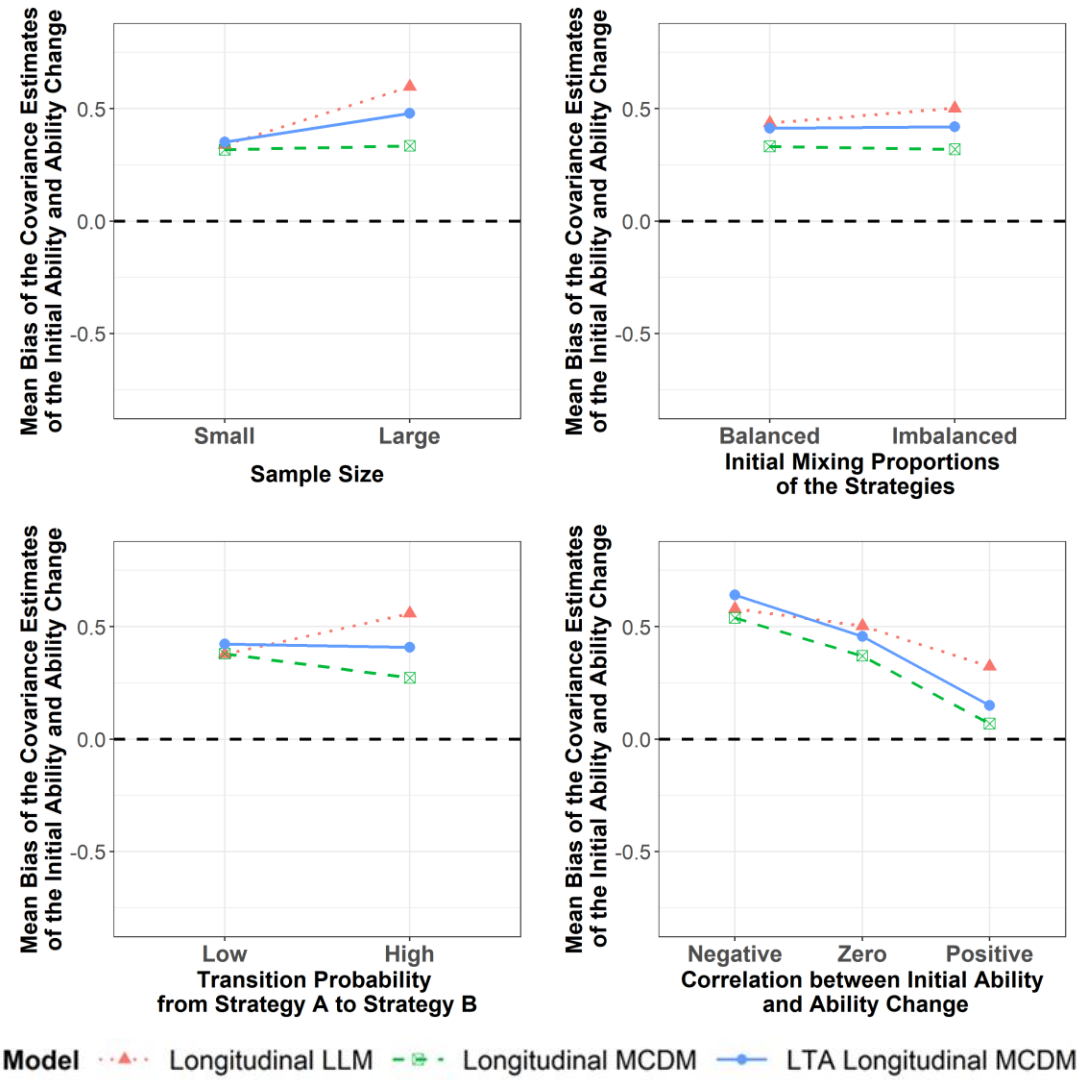


Figure 22. Marginal mean bias of the covariance estimates between the initial ability and ability change, $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$, at each level of the manipulated factors.

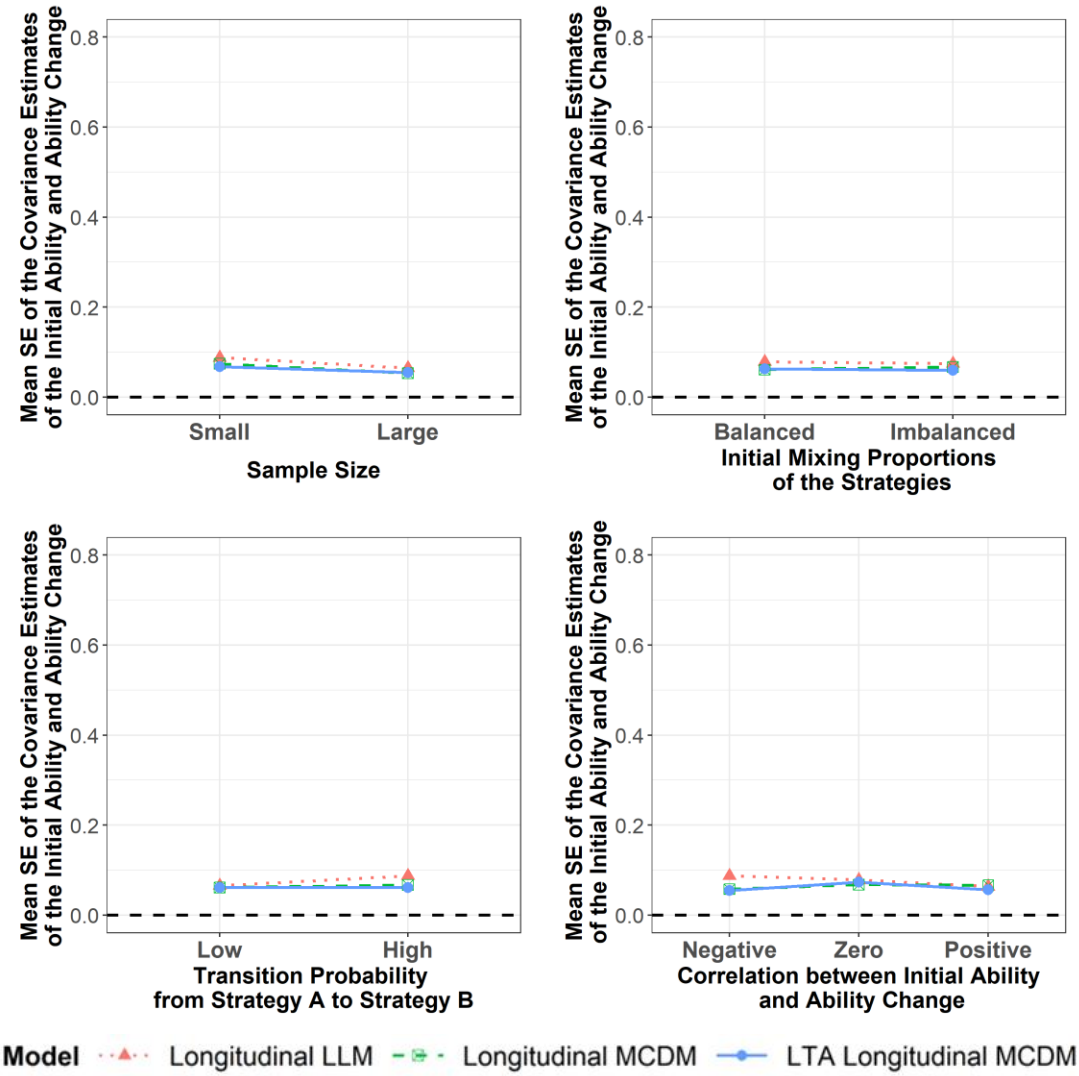


Figure 23. Marginal mean SE of the covariance estimate between the initial ability and ability change, $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$, at each level of the manipulated factors.

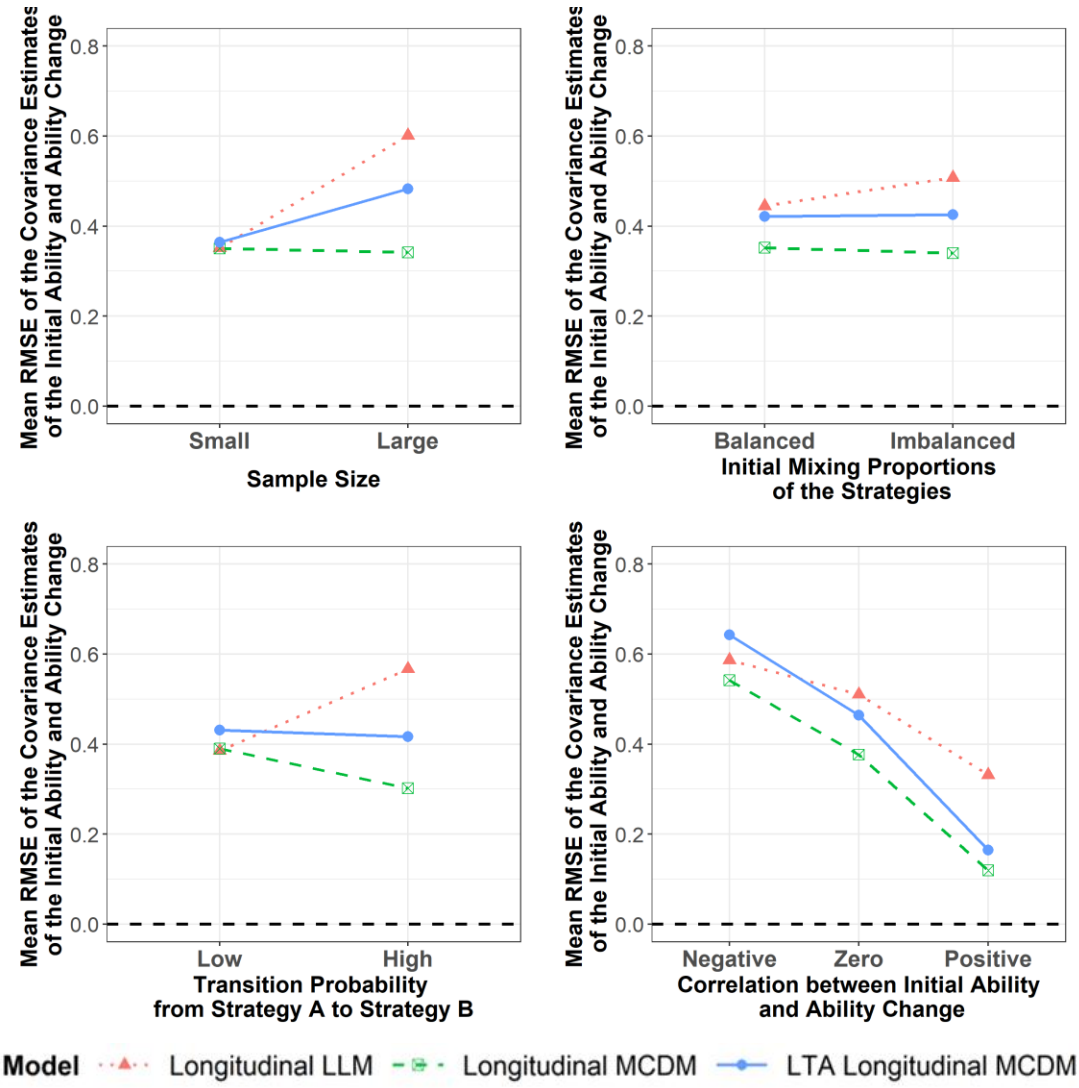


Figure 24. Marginal mean RMSE of the covariance estimate between the initial ability and ability change, $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$, at each level of the manipulated factors.

4.2.3 Strategy choice

Like the skill implementation ability parameters, the strategy choice parameters are also categorized into the first-level and second-level parameters. The first-level parameters include the strategy choice membership classifications (m) at each timepoint, while the second-level parameters include the initial mixing proportions of the strategies ($\pi_m^{(T_1)}$) and the strategy latent transition probability

parameter ($\tau^{(t-1)}_{m_t|m_{t-1}}$). Given that the strategy parameters are not present in the Longitudinal LLM, the recovery of the strategy choice parameters (except the latent transition probability parameter that is unique to the LTA-longitudinal-MCDM) are only compared across the LTA-longitudinal-MCDM and the Longitudinal MCDM. Hence, the results would shed light on the effects of ignoring within-person strategy shift on the strategy choice parameter estimates. Since the strategy latent transition probability parameter ($\tau^{(t-1)}_{m_t|m_{t-1}}$) is unique to the LTA-longitudinal-MCDM, the recovery of this parameter is only reported for the LTA-longitudinal-MCDM.

4.2.3.1 Strategy choice membership classifications

The recovery of the first-level strategy choice parameters is quantified by the correct classification rate of the strategy choice at each time point as well as the correct classification rate of the strategy choice trajectory. According to Figures 25 and 26, the marginal mean correct classification rates of the strategy choice at each time point as well as the the strategy choice trajectory are higher for the LTA-longitudinal-MCDM than those for the Longitudinal MCDM except at the low level of strategy transition probability ($p_{M_B|M_A} = 0.3$). The average-across-replication strategy choice (trajectory) classification accuracies under all the simulated conditions are listed in Tables 25 and 26 where the higher accuracy among the two models is bolded under each condition. The LTA-longitudinal-MCDM is higher in strategy choice trajectory classification accuracy than the Longitudinal-MCDM by a maximum of 0.407, except in certain conditions with balanced initial mixing proportions of the strategies and a low strategy transition probability (i.e., $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.6{:}0.4$ and $p_{M_B|M_A} = 0.3$), In the conditions with balanced initial mixing

proportions of the strategies and a low transition probability from strategy A to strategy B (i.e., $\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)} = 0.6{:}0.4$ and $p_{M_B|M_A} = 0.3$), the Longitudinal-MCDM is slightly higher than the LTA-longitudinal-MCDM in the strategy choice trajectory classification accuracy, by a maximum of 0.061.

Moreover, Tables 25 and 26 present the proportion of replications, among the 30 total replications, where the strategy choice (trajectory) classification accuracies are higher for the LTA-longitudinal-MCDM than the Longitudinal MCDM, which is an additional piece of information of the comparative performance of the LTA-longitudinal-MCDM and the Longitudinal MCDM in terms of the strategy choice classification. When the true latent transition probability from Strategy A to Strategy B is high ($p_{M_B|M_A} = 0.7$), the LTA-longitudinal-MCDM has higher classification accuracies of the strategy choice (trajectory) than the Longitudinal MCDM in 97% to 100% of the replications. When the true latent transition probability from Strategy A to Strategy B is low ($p_{M_B|M_A} = 0.3$), the relative performance of the LTA-longitudinal-MCDM to the Longitudinal MCDM on the strategy choice (trajectory) classification accuracy is diminished. The diminishment is particularly severe at the second time point under the conditions with balanced initial mixing proportions of strategies ($\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)} = 0.6{:}0.4$) and a small sample size (J=100).

To examine the effects of the manipulated factors on the strategy choice (trajectory) classification accuracy of the proposed model, trends of the solid lines in Figures 25 and 26 are inspected. The differential slopes of the solid lines across timepoints in Figure 25 imply that the effects of the manipulated factors on the strategy choice classification accuracy tend to be inconsistent across timepoints. For

instance, as shown in the lower-left panel of Figure 25, the marginal mean strategy choice classification accuracy is similar across the two levels of true latent transition probability from Strategy A to Strategy B at Timepoint 1, but the marginal mean strategy choice classification accuracy is higher in the conditions with higher strategy transition probability at Timepoint 2.

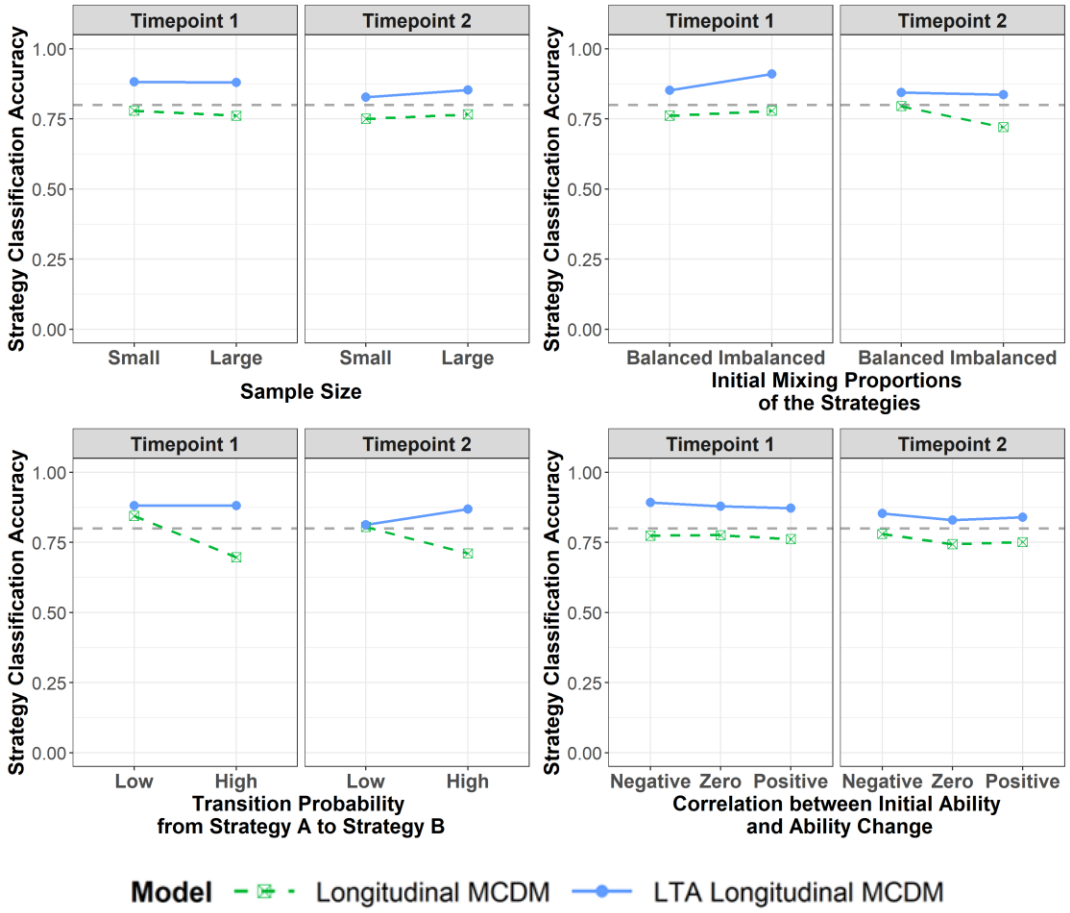


Figure 25. Marginal mean strategy classification accuracy at each level of the manipulated factors.

Table 25

Classification Accuracy of Strategy Choice at Each Timepoint

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Strategy Choice Classification Accuracy at Timepoint 1			Strategy Choice Classification Accuracy at Timepoint 2		
				L-MCDM	LTA-L-MCDM	Prop. of Rep LTA>L	L-MCDM	LTA-L-MCDM	Prop. of Rep LTA>L
100	0.6:0.4	0.3	-0.3	0.859	**0.873**	0.53	**0.843**	0.818	0.17
			0	0.836	**0.847**	0.63	**0.800**	0.786	0.40
			0.3	0.813	**0.827**	0.60	0.798	**0.810**	0.60
		0.7	-0.3	0.711	**0.875**	1.00	0.793	**0.874**	0.97
			0	0.721	**0.848**	1.00	0.728	**0.843**	1.00
			0.3	0.670	**0.827**	1.00	0.757	**0.874**	0.97
	0.8:0.2	0.3	-0.3	0.885	**0.930**	0.87	0.807	**0.818**	0.63
			0	0.856	**0.912**	0.97	**0.772**	0.766	0.47
			0.3	0.845	**0.907**	0.90	0.789	**0.801**	0.57
		0.7	-0.3	0.717	**0.928**	1.00	0.671	**0.862**	1.00
			0	0.738	**0.911**	1.00	0.606	**0.830**	1.00
			0.3	0.702	**0.905**	1.00	0.642	**0.857**	1.00
800	0.6:0.4	0.3	-0.3	0.828	**0.857**	1.00	**0.841**	0.832	0.33
			0	0.830	**0.852**	0.97	0.821	**0.825**	0.67
			0.3	0.830	**0.854**	0.93	**0.816**	0.814	0.37
		0.7	-0.3	0.674	**0.858**	1.00	0.799	**0.897**	1.00
			0	0.686	**0.854**	1.00	0.780	**0.886**	1.00
			0.3	0.680	**0.853**	1.00	0.774	**0.882**	1.00
	0.8:0.2	0.3	-0.3	0.844	**0.908**	1.00	0.808	**0.837**	1.00
			0	0.846	**0.903**	1.00	0.791	**0.828**	1.00
			0.3	0.858	**0.904**	1.00	0.784	**0.820**	0.97
		0.7	-0.3	0.672	**0.908**	1.00	0.679	**0.887**	1.00
			0	0.691	**0.903**	1.00	0.652	**0.873**	1.00
			0.3	0.694	**0.904**	1.00	0.649	**0.863**	1.00

Note. L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal-MCDM. Prop. of Rep LTA>L=The proportion of replications (out of 30 replications) with strategy choice classification accuracy higher for the LTA-longitudinal-MCDM than the Longitudinal MCDM. The higher classification accuracy among the two models are bolded under each condition at each time point.

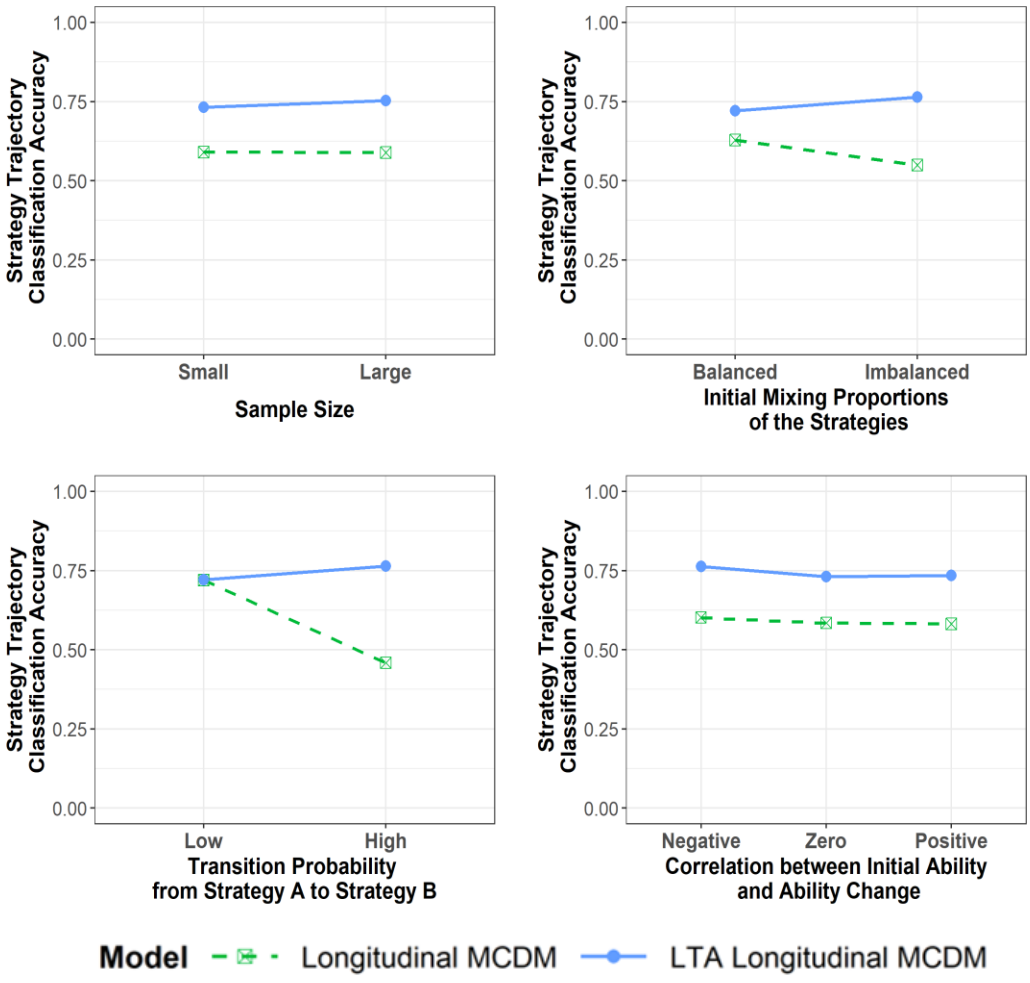


Figure 26. Marginal mean strategy trajectory classification accuracy at each level of the manipulated factors.

Table 26
Classification Accuracy of Strategy Choice Trajectory

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Strategy Choice Trajectory Classification Accuracy		
				L-MCDM	LTA-L-MCDM	Prop. of Rep LTA>L
100	0.6:0.4	0.3	-0.3	**0.761**	0.716	0.10
			0	**0.728**	0.667	0.07
			0.3	**0.715**	0.674	0.17
		0.7	-0.3	0.542	**0.764**	1.00
			0	0.514	**0.719**	1.00
			0.3	0.504	**0.730**	1.00
	0.8:0.2	0.3	-0.3	0.726	**0.761**	0.83
			0	0.694	**0.703**	0.53
			0.3	0.697	**0.73**	0.80
		0.7	-0.3	0.414	**0.797**	1.00
			0	0.392	**0.754**	1.00
			0.3	0.392	**0.771**	1.00
800	0.6:0.4	0.3	-0.3	**0.744**	0.715	0.03
			0	**0.735**	0.703	0.00
			0.3	**0.733**	0.693	0.00
		0.7	-0.3	0.527	**0.768**	1.00
			0	0.523	**0.753**	1.00
			0.3	0.517	**0.747**	1.00
	0.8:0.2	0.3	-0.3	0.706	**0.775**	1.00
			0	0.698	**0.764**	1.00
			0.3	0.701	**0.754**	1.00
		0.7	-0.3	0.396	**0.803**	1.00
			0	0.392	**0.784**	1.00
			0.3	0.391	**0.774**	1.00

Note. L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal-MCDM. Prop. of Rep LTA>L=The proportion of replications (out of 30 replications) with strategy choice classification accuracy higher for the LTA-longitudinal-MCDM than the Longitudinal MCDM. The higher classification accuracy among the two models are bolded under each condition.

4.2.3.2 Initial strategy mixing proportion estimates

To investigate the impact of ignoring the multiple-strategy scenarios on the estimates of the initial mixing proportion parameter, $\boldsymbol{\pi}_m^{(T_1)}$, under various simulated conditions, the marginal mean bias, SE and RMSE of the initial mixing proportion

estimates of Strategy A, $\hat{\pi}_{M_A}^{(T_1)}$, are plotted against each level of each manipulated factor for the LTA-longitudinal-MCDM and Longitudinal MCDM, as shown in Figures 27, 28 and 29. Given that the mixing proportion estimate of Strategy B, $\hat{\pi}_{M_B}^{(T_1)}$, can be directly calculated from that of Strategy A as the two mixing proportions sum up to 1 and, thus, the recoveries are similar between the two initial mixing proportion parameters, the outcome measures of $\hat{\pi}_{M_B}^{(T_1)}$ are not plotted to avoid redundancy. The impact of ignoring the multiple-strategy scenarios on the recovery of $\boldsymbol{\pi}_m^{(T_1)}$ can be examined by comparing the recovery outcome measures across different data-fitting model types manifested as different line patterns in Figures 27, 28 and 29. In general, the absolute values of the marginal mean biases of $\hat{\pi}_{M_A}^{(T_1)}$ from the LTA-longitudinal-MCDM are lower than those from the Longitudinal MCDM (See Figure 27), while the marginal mean SEs of $\hat{\pi}_{M_A}^{(T_1)}$ are similar across the two models (See Figure 28). Such patterns imply that, on average, the systematic errors of $\hat{\pi}_{M_A}^{(T_1)}$ are sensitive to the neglect of the within-person strategy shift in the model, but the random errors of $\hat{\pi}_{M_A}^{(T_1)}$ are relatively insensitive. As for the RMSE that quantifies the magnitudes of systematic and random errors of $\hat{\pi}_{M_A}^{(T_1)}$ as a whole, the marginal mean RMSEs from the LTA-longitudinal-MCDM are smaller than those from the Longitudinal-MCDM at all the levels of all the manipulated factors (See Figure 29).

To examine how the recovery of $\boldsymbol{\pi}_m^{(T_1)}$ from the LTA-longitudinal-MCDM is affected by the manipulated factors, the trends of the solid lines in Figures 27, 28 and 29 are inspected. The absolute values of the mean bias, SE and RMSE $\hat{\pi}_{M_A}^{(T_1)}$ are lower

in the large sample size conditions (J=800) than in the small sample size conditions (J=100). Nevertheless, the recovery outcome measures of $\boldsymbol{\pi}_m^{(T_1)}$ do not appear to differ significantly across the levels of the other manipulated factors.

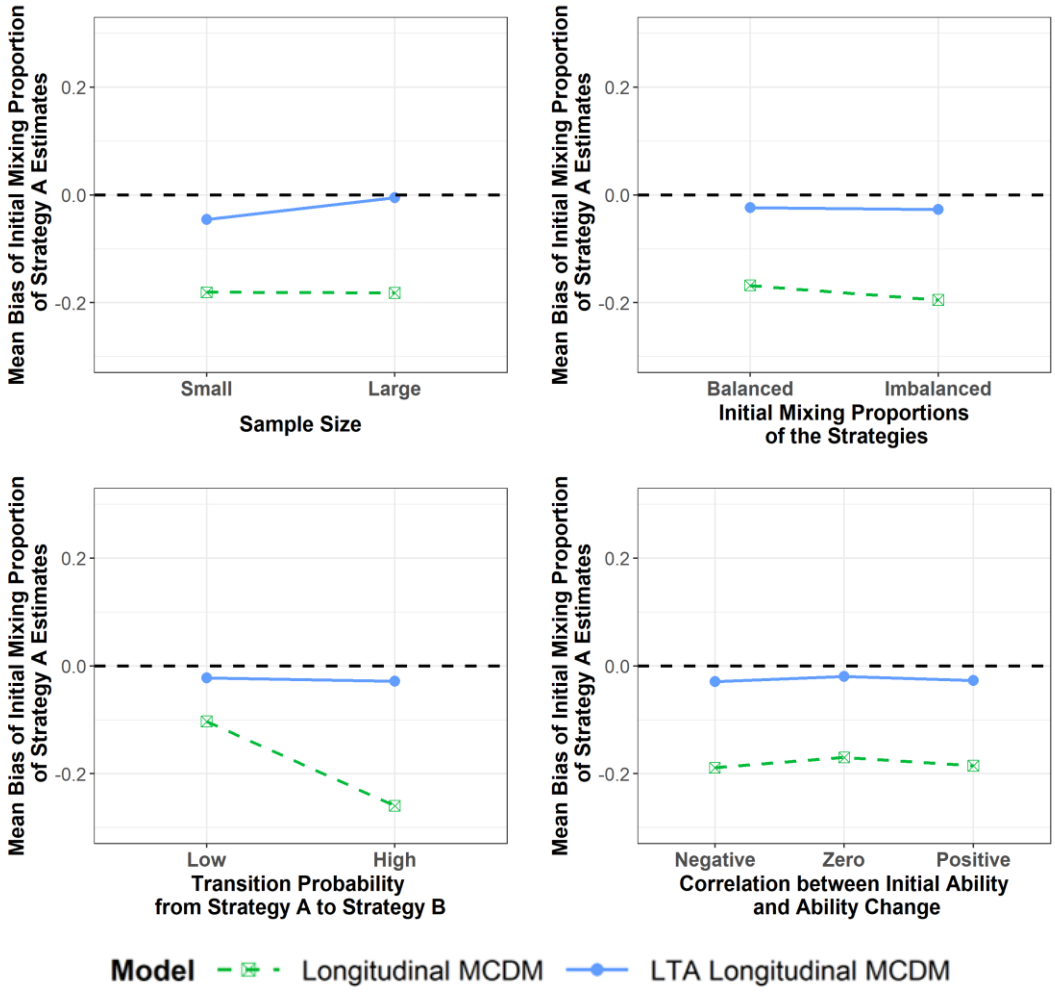


Figure 27. Marginal mean bias of the initial mixing proportion estimates of Strategy A, $\hat{\pi}_{M_A}^{(T_1)}$, at each level of the manipulated factors.

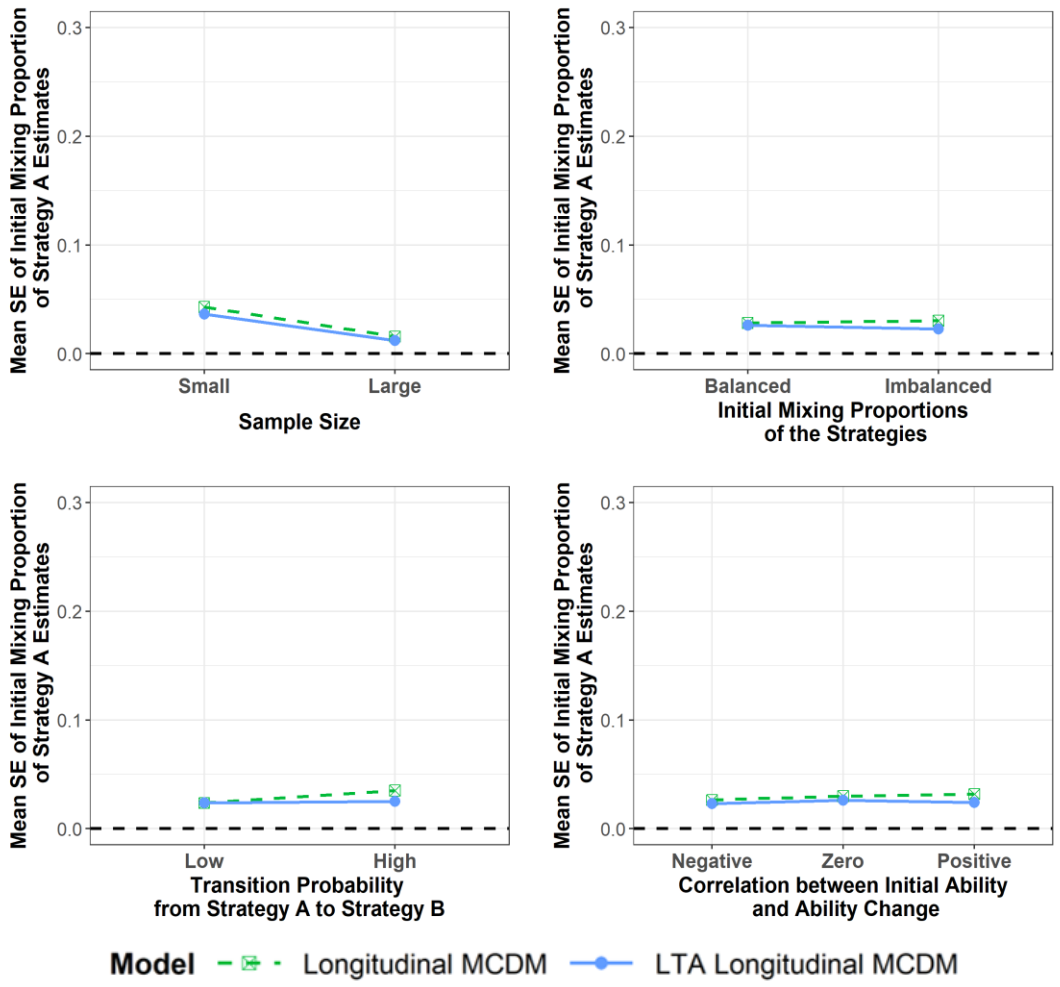


Figure 28. Marginal mean SE of the initial mixing proportion estimates of Strategy A, $\hat{\pi}_{M_A}^{(T_1)}$, at each level of the manipulated factors.

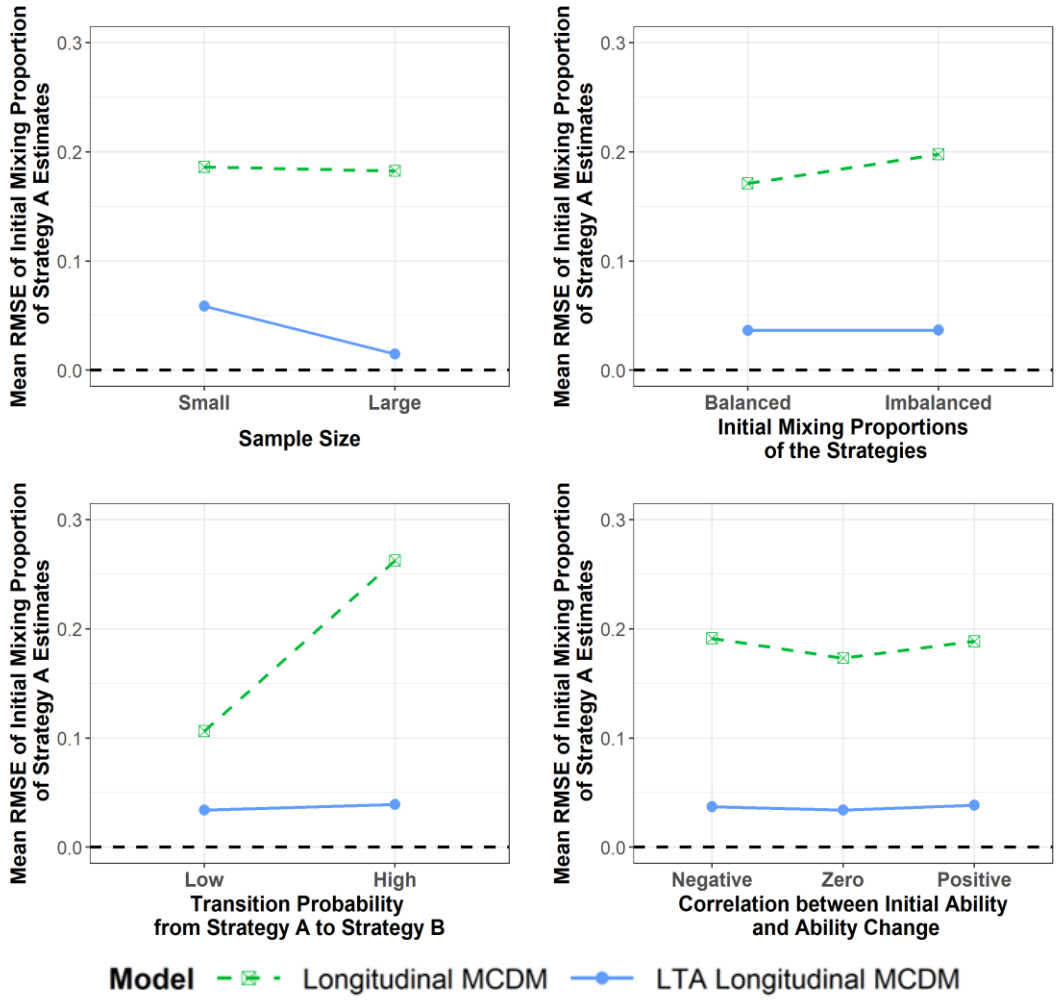


Figure 29. Marginal mean RMSE of the initial mixing proportion estimates of Strategy A, $\hat{\pi}_{M_A}^{(T_1)}$, at each level of the manipulated factors.

4.2.3.3 Strategy choice latent transition probability estimates

Given that the simulees are assumed to choose either Strategy A or Strategy B at each time point and that the transition probability from Strategy B to Strategy A has been constrained at zero, only one strategy choice latent transition probability parameter (i.e., the transition probability from Strategy A to Strategy B, $\tau_{M_B|M_A}$) is freely estimated. Figures 30, 31 and 32 display the biases, SEs and RMSEs of $\hat{\tau}_{M_B|M_A}^{(T_1)}$ from the LTA-longitudinal-MCDM in all the 24 simulated conditions. Each bar in

Figures 30, 31 and 32 represents the outcome measure of $\hat{\tau}^{(T_1)}_{M_B|M_A}$ in one condition, and comparing the bars from different perspectives can reveal the effects of each manipulated factor, controlling for the other three factors, on the recovery of $\tau^{(T_1)}_{M_B|M_A}$: First, one can observe the effects of sample size by comparing the adjacent bars of different colors. For instance, according to Figures 31 and 32, the SEs and RMSEs of $\hat{\tau}^{(T_1)}_{M_B|M_A}$ are lower in the large sample size conditions (J=800) than in the small sample size conditions (J=100), controlling for the other manipulated factors. Nevertheless, according to Figure 30, the effects of sample size on the bias of $\hat{\tau}^{(T_1)}_{M_B|M_A}$ are inconsistent across different levels of the other three manipulated factors. Second, comparing bars in the left panels to the bars at the corresponding position in the right panels allows one to see the effects of the initial mixing proportions of the strategies ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B}$). For example, according to Figure 31, when the true transition probability from Strategy A to Strategy B is high ($p_{M_B|M_A} = 0.7$), the SEs of $\hat{\tau}^{(T_1)}_{M_B|M_A}$ are lower in the conditions with more imbalanced initial mixing proportions of the strategies ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.8 : 0.2$), controlling for the other factors. Third, comparing bars in the upper panels to the bars at the corresponding position in the lower panels allows one to observe the effects of the true transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$). Last, comparing the bars of the same color across the x-axis within a panel enables one view the effects the true correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$). No consistent effect of the initial mixing proportions of the strategies ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B}$), true transition probability from Strategy A to Strategy B (

$p_{M_B|M_A}$) or the correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$) on the bias, SE or RMSE of $\hat{\tau}^{(T_1)}_{M_B|M_A}$ is observed across different levels of the other manipulated factors.

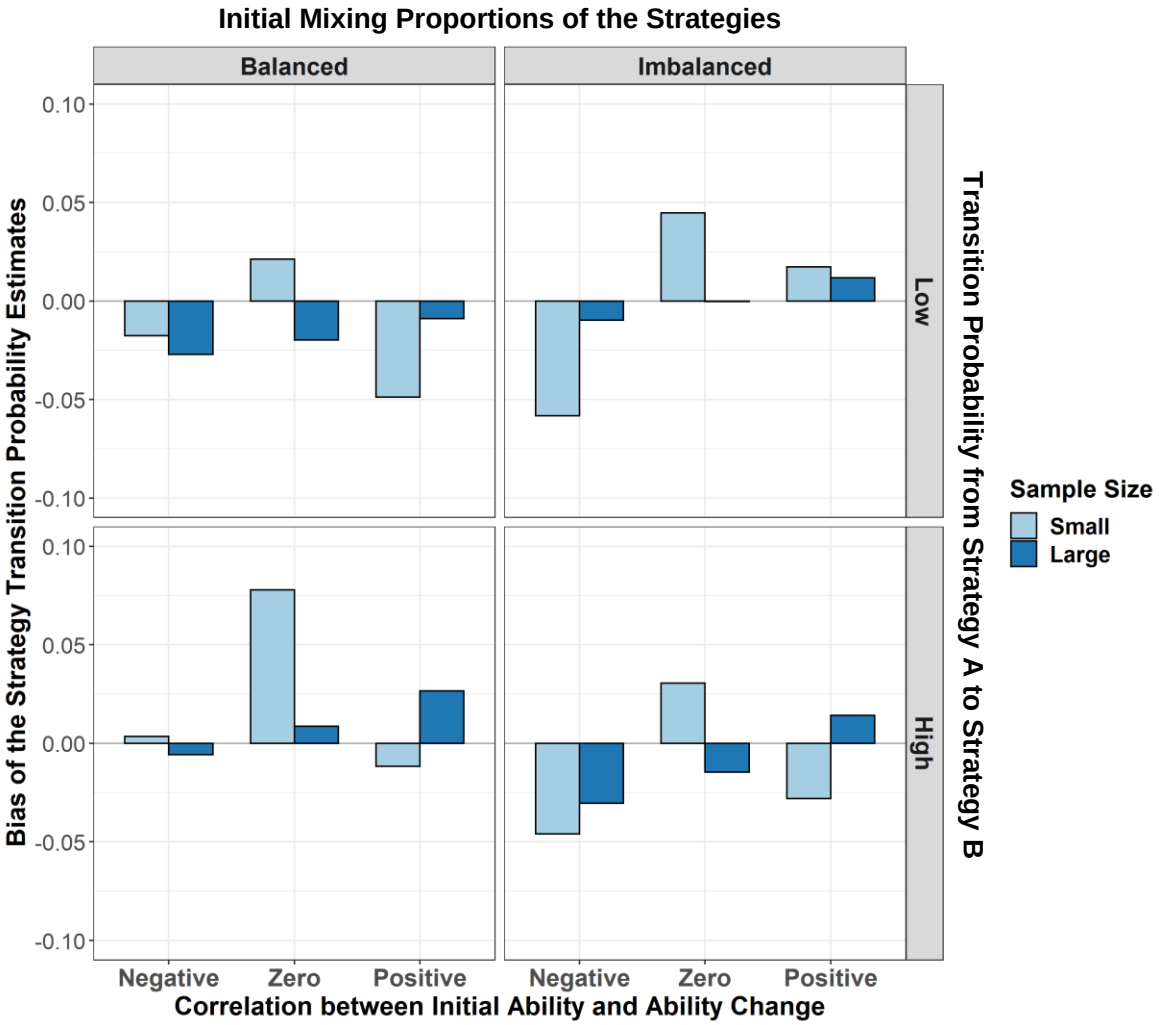


Figure 30. Bias of the latent transition probability estimate from Strategy A to Strategy B, $\hat{\tau}^{(T_1)}_{M_B|M_A}$, based on the LTA-longitudinal-MCDM at each simulated condition.

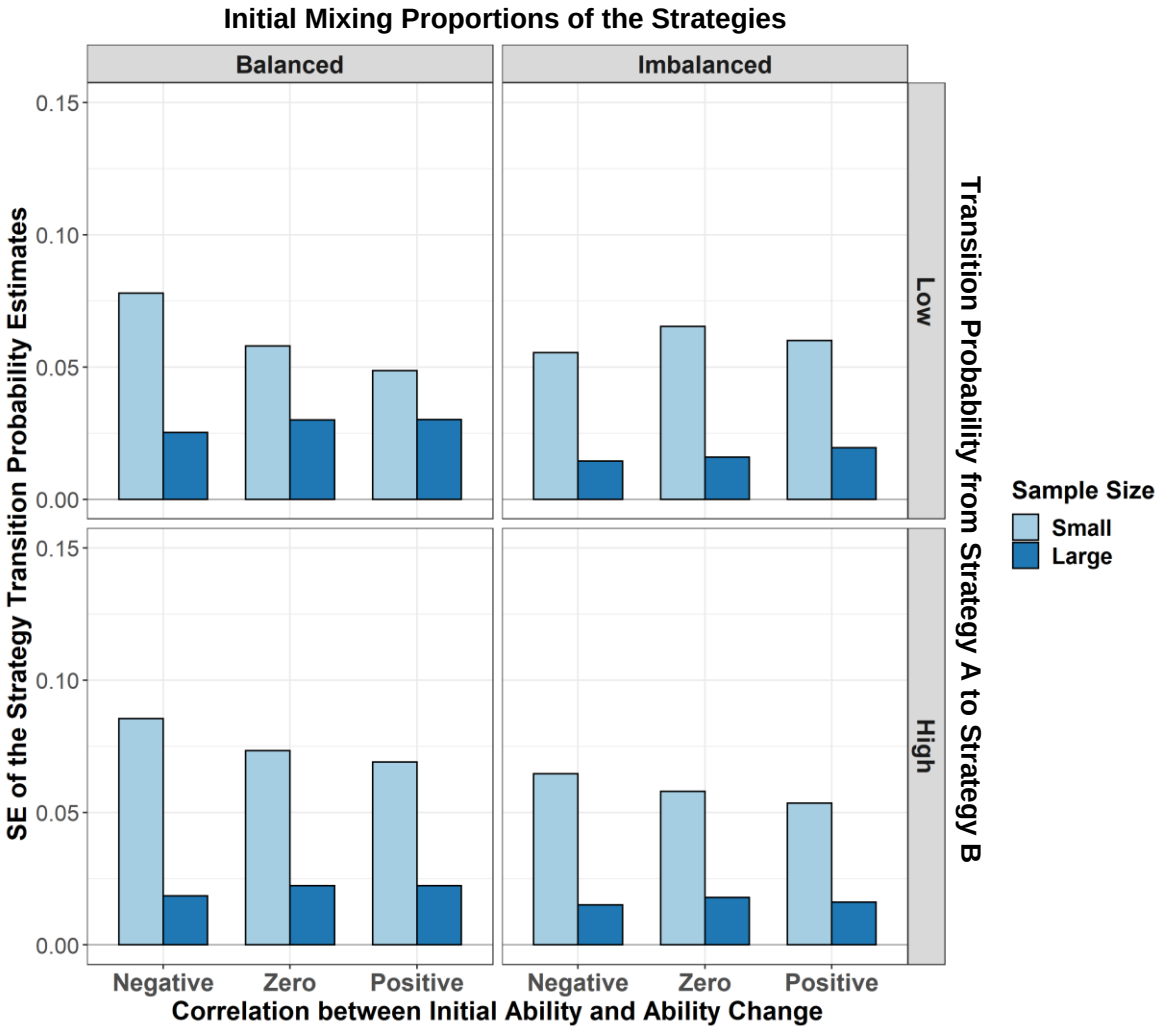


Figure 31. SE of the latent transition probability estimate from Strategy A to Strategy B, $\hat{\tau}^{(T_1)}_{M_B|M_A}$, based on the LTA-longitudinal-MCDM at each simulated condition.

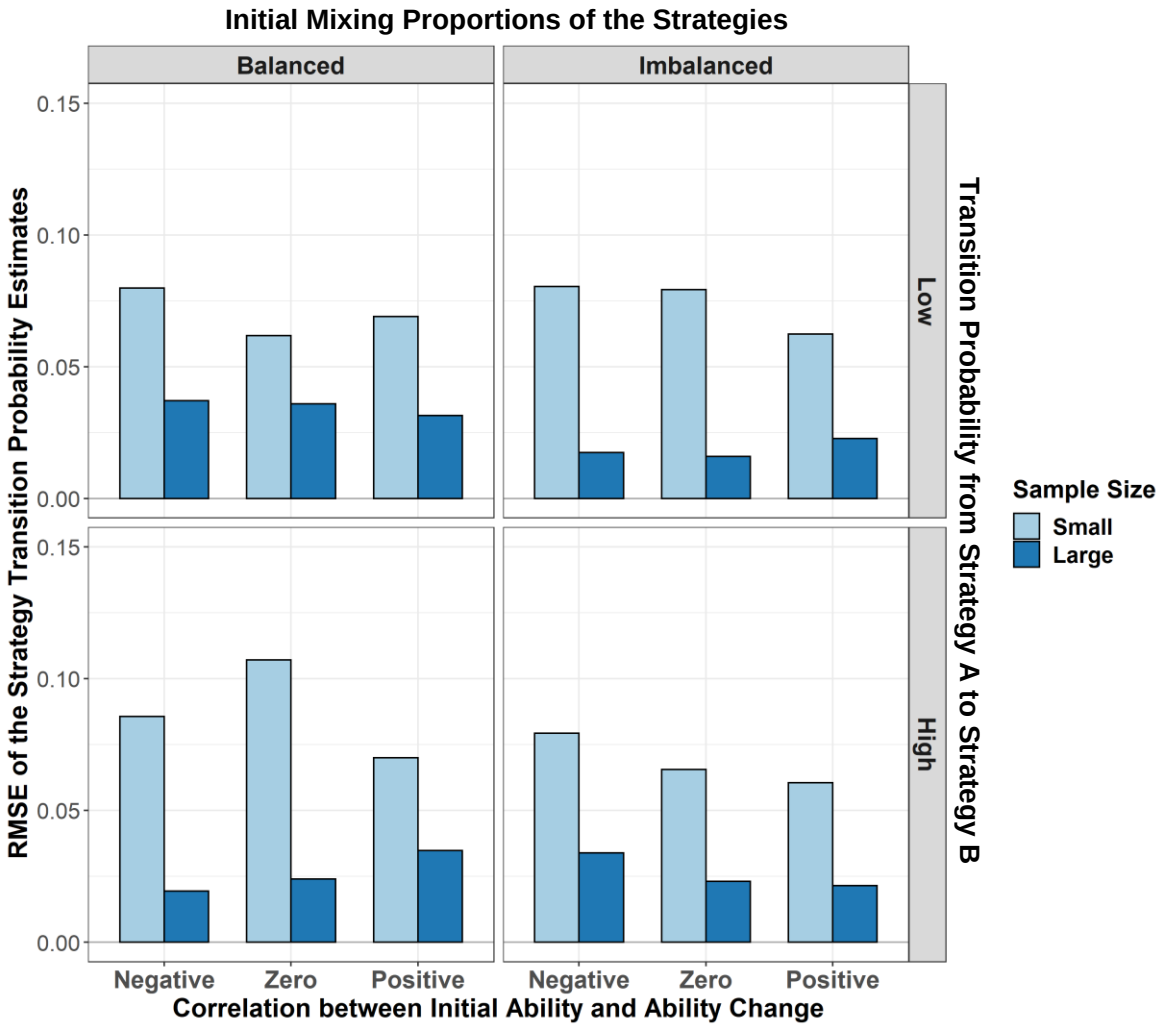


Figure 32. RMSE of the latent transition probability estimate from Strategy A to Strategy B, $\hat{\tau}^{(T_1)}_{M_B|M_A}$, based on the LTA-longitudinal-MCDM at each simulated condition.

4.3 Recovery of the Item Parameters

Item parameters refer to the parameters that delineate the characteristics of items and, thus, are item-specific. The item parameters considered in this study include the item intercept parameters ($\lambda_{i,0}$) and the attribute main effect parameters ($\lambda_{i,1,(k)}$), as shown in Table 12. These item parameters are estimated in all the three data-fitting models. Thus, in order to examine the effects of ignoring the multiple-

strategy scenarios on the item parameter recovery, the recovery outcome measures of the item parameters were compared across the three models with the mixed-effect ANOVAs. As elaborated in Section 3.3.5, in the mixed-effect ANOVAs, each item parameter was treated as a subject, the bias/SE/RMSE of the parameter was treated as the measurement (i.e., dependent variable) taken on each subject, the data-fitting model type was treated as the repeated-measure factor (i.e., within-subject factor) and the four manipulated factors were treated as the between-subject factors.

Table 27 summarizes the highest-order significant effects of the data-fitting model type and the manipulated factors on the item parameter estimates found in the mixed-effect ANOVAs. Elaborations and visualizations of these effects are provided in the following sections. Note that the number of attribute main effect parameters varies across the data-fitting models, i.e., the longitudinal LLM only has 22 attribute main effect parameters whereas the Longitudinal MCDM and LTA-longitudinal-MCDM have 35 that subsume the 22 in the Longitudinal LLM. The mixed-effect ANOVAs were only performed on the recovery outcome measures of the 22 attribute main effect parameters that are shared by the three data-fitting models; for the remaining 13 attribute main effect parameters that are only contained by the Longitudinal MCDM and LTA-longitudinal-MCDM, only marginal mean plots are inspected to examine the effects of the data-fitting model and manipulated factors on the parameter recovery, due to the insufficient group size for ANOVAs.

While the results from the mixed-effect ANOVAs inform the impact of ignoring the multiple-strategy scenarios on the item parameter recovery of the longitudinal CDMs, the four-way ANOVAs were also performed on the recovery

outcome measures of the item parameters to investigate the effects of the manipulated factors on the item parameter recovery of the proposed model (i.e., the LTA-longitudinal-MCDM). The four-way ANOVA results regarding the item intercept and attribute main effect parameters are presented at the end of Sections 4.3.1 and 4.3.2, respectively.

Table 27

Summary of Effect Sizes of the Highest-Order Significant Effects from the Mixed-Effect ANOVA on the Item Parameter Recovery

	Item Intercept Parameter ($\lambda_{i,0}$)			Attribute Main Effect Parameter[1] ($\lambda_{i,1,(k)}$)		
Effect	Bias	SE	RMSE	Bias	SE	RMSE
MODEL	0.262	0.021	0.257	0.272		
SIZE	0.083	0.863	0.288	0.038		
MIXING		0.015				
TR_Prob	0.100			0.019		
SIZE*MODEL					0.032	
MIXING*MODEL						0.011

Effect Size	Small (0.01≤partial η^2 <0.06)	Medium (0.06≤partial η^2 <0.14)	Large (partial η^2 ≥0.14)

Note. [1]Only the recovery of the attribute main effect parameters that are shared by all the three models were compared with ANOVA.
SIZE=Sample size (*J*); MIXING=Initial mixing proportions of Strategy A and Strategy B ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); MODEL=Data-fitting model type. The values in the cells are partial η^2. The table does not include simple effects.

4.3.1 Item intercept

As shown in the columns under "Item Intercept Parameter" in Table 27, the data-fitting model type (MODEL) as well as three manipulated factors, including sample size (SIZE), initial mixing proportions of the strategies (MIXING) and latent transition probability of strategy (TR_Prob), have significant main effects on one or more recovery outcome measures of $\lambda_{i,0}$. To be more specific, the data-fitting model type (MODEL) and sample size (SIZE) have significant main effects on the systematic errors of $\hat{\lambda}_{i,0}$ that are quantified by bias, random errors of $\hat{\lambda}_{i,0}$ that are quantified by SE, and the systematic and random errors of $\hat{\lambda}_{i,0}$ as a whole that are quantified by RMSE. In particular, the effects of MODEL on the bias (F=162.06, p<0.001, partial η^2=0.262) and RMSE (F=157.33, p<0.001, partial η^2=0.257) of $\hat{\lambda}_{i,0}$ are of large effect sizes, according to Table 28. Figures 33 and 34 show that the LTA-longitudinal-MCDM has the lowest marginal mean bias and RMSE of $\hat{\lambda}_{i,0}$ among the three competing models, while the Longitudinal LLM that ignores both between-person multiple strategies and within-person strategy shift has the highest mean bias and RMSE of $\hat{\lambda}_{i,0}$. In addition, SIZE is found to have large main effects on the SE (F=2862.75, p<0.001, partial η^2=0.863) and RMSE (F=184.26, p<0.001, partial η^2=0.288) of $\hat{\lambda}_{i,0}$, according to Tables 28 and 29. As shown in Figures 34 and 35, the marginal mean SE and RMSE of $\hat{\lambda}_{i,0}$ are lower in the large sample size conditions (J=800) than those in the small sample size conditions (J=100).

Moreover, as seen in Tables 28 and 29, TR_Prob and MIXING have small or medium significant main effects on the bias and SE of $\hat{\lambda}_{i,0}$, respectively. To visualize these significant main effects, the marginal means of the bias, SE and RMSE of $\hat{\lambda}_{i,0}$ by the levels of each factor are plotted in Figures 33, 34 and 35.

Table 28
Significant Effects in the Mixed-Effect ANOVA Results of the Bias and RMSE of the Item Intercept Estimates

Source	Bias of $\hat{\lambda}_{i,0}$			RMSE of $\hat{\lambda}_{i,0}$		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)						
MODEL	162.06	<0.001	0.262	157.33	<0.001	0.257
Between-Subject Effects						
SIZE	41.15	<0.001	0.083	184.26	<0.001	0.288
TR_Prob	4.58	0.033	0.010			

Note. TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); SIZE=Sample size (*J*); MODEL=Data-fitting model type

Table 29
Significant Effects in the Mixed-Effect ANOVA Results of the SE of the Item Intercept Estimates

Source	SE of $\hat{\lambda}_{i,0}$		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)			
MODEL	9.81	0.001	0.021
Between-Subject Effects			
SIZE	2862.75	<0.001	0.863
MIXING	7.10	0.008	0.015

Note. MIXING=Initial mixing proportions of Strategy A and Strategy B ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B}$); SIZE=Sample size (*J*); MODEL=Data-fitting model type.

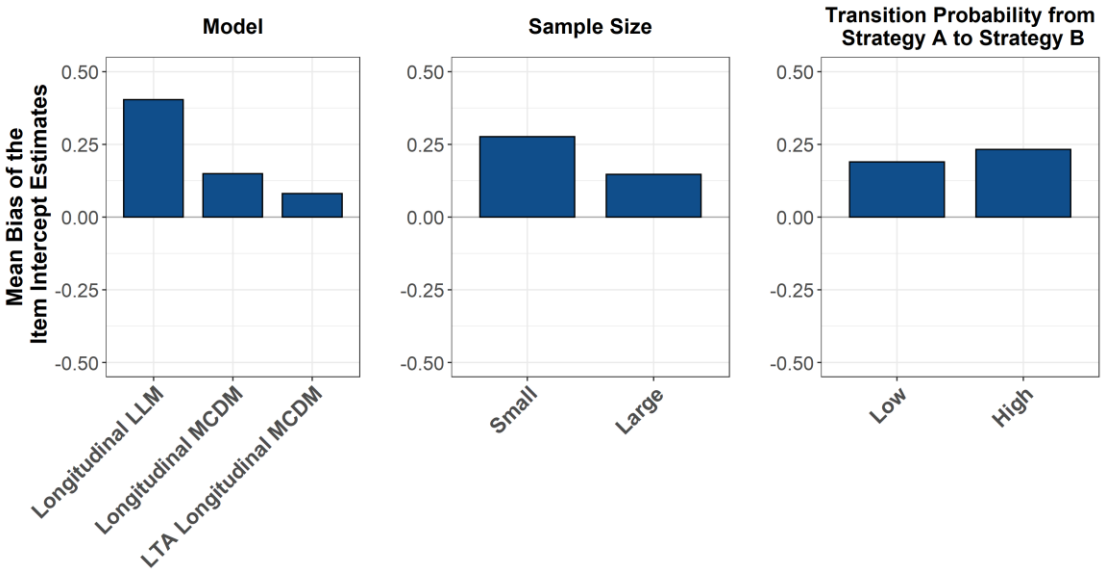


Figure 33. Significant main effects of MODEL, SIZE and TR_Prob on the bias of the item intercept parameter estimates, $\hat{\lambda}_{i,0}$. [Note. MODEL=Data-fitting model type; SIZE=Sample size (*J*); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$).]

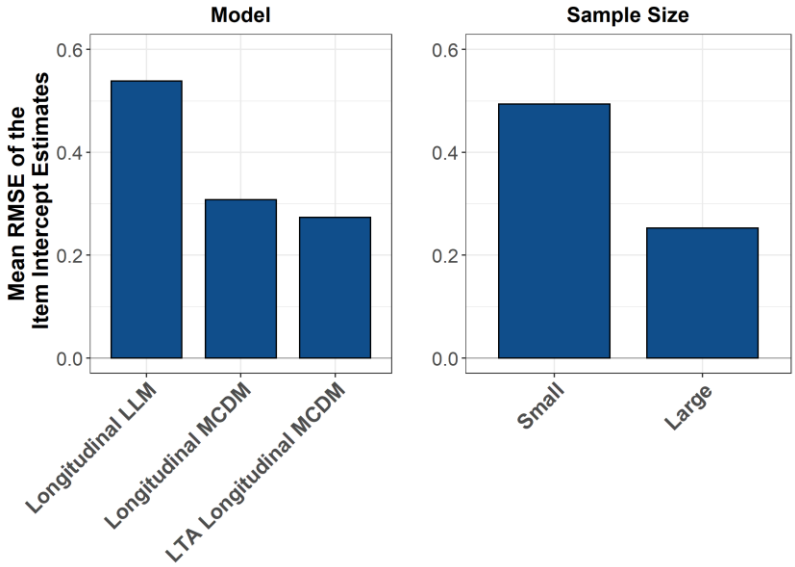


Figure 34. Significant main effects of MODEL and SIZE on the RMSE of the item intercept parameter estimates, $\hat{\lambda}_{i,0}$. [Note. MODEL=Data-fitting model type; SIZE=Sample size (*J*).]

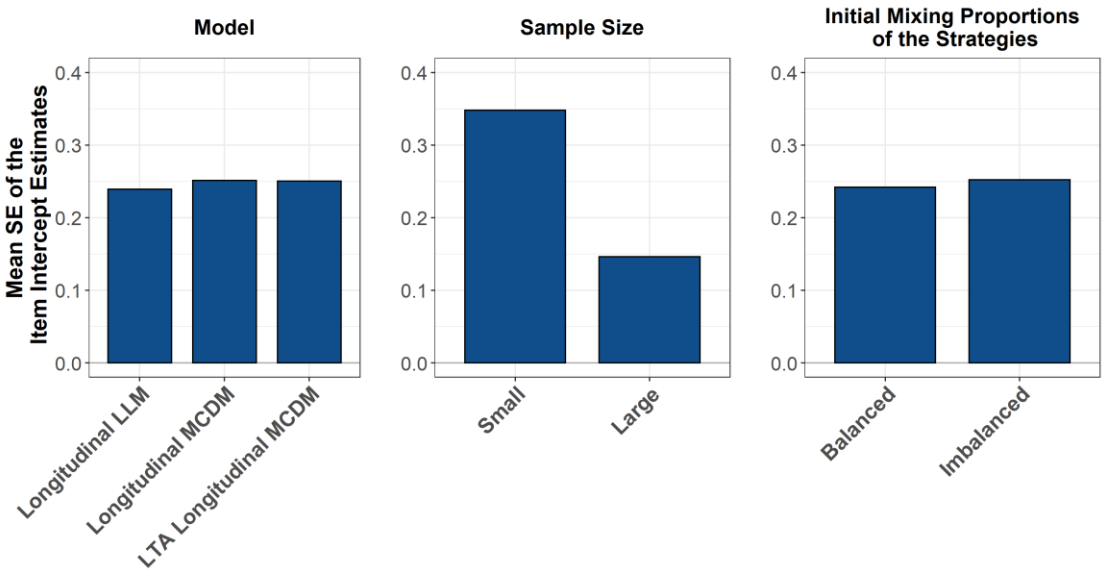


Figure 35. Significant main effects of MODEL, SIZE and MIXING on the SE of the item intercept parameter estimates, $\hat{\lambda}_{i,0}$. [Note. MODEL=Data-fitting model type; SIZE=Sample size (*J*); MIXING=Initial mixing proportions of Strategy A and Strategy B ($\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$).]

While the mixed-effect ANOVA results above indicated some of the manipulated factors have significant main effects on one or more recovery outcome measures of the item intercept parameters, four-way ANOVAs were conducted to further examine the effects of the manipulated factors on the item intercept parameter recovery of the proposed model. According to the results shown in Table 30, sample size (SIZE) has large effects on the bias (F=406.06, p<0.001, partial η^2=0.471), SE (F=1680.52, p<0.001, partial η^2=0.787) and RMSE (F=2196.89, p<0.001, partial η^2=0.828) of $\hat{\lambda}_{i,0}$ from the proposed model. An inspection in the marginal means indicated that the absolute values of the mean bias, SE and RMSE of $\hat{\lambda}_{i,0}$ from the proposed model are lower in the large sample size (J=800) conditions than those in the small sample size (J=100) conditions.

Table 30

Significant Effects in the Four-Way ANOVA Results of the Recovery of the Item Intercept Parameter from the LTA-longitudinal-MCDM

Source	Bias of $\hat{\lambda}_{i,0}$		SE of $\hat{\lambda}_{i,0}$		RMSE of $\hat{\lambda}_{i,0}$	
	p-value	Partial η^2	*p*-value	Partial η^2	*p*-value	Partial η^2
SIZE	<0.001	0.471	<0.001	0.787	<0.001	0.828
CORR	0.019	0.017				

Effect Size	Small	Medium	Large
	(0.01≤partial η^2 <0.06)	(0.06≤partial η^2 <0.14)	(partial η^2 ≥0.14)

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(\tau_1)}\Delta\theta}$); SIZE=Sample size (*J*); MODEL=Data-fitting model type

4.3.2 Attribute main effect

4.3.2.1 Attribute main effect parameters shared by the three models

As mentioned at the beginning of Section 4.3, the effects of the data-fitting model type and the manipulated factors on the recovery of the 22 attribute main effect parameters ($\lambda_{i,1,(k)}$) that are present in all the three data-fitting models were investigated with the mixed-effect ANOVAs. As indicated by the columns under "Attribute Main Effect Parameter" in Table 27, the data-fitting model type (MODEL) as well as sample size (SIZE) and initial mixing proportions of the strategies (MIXING) have significant main effects on the bias of $\hat{\lambda}_{i,1,(k)}$. SIZE and MODEL interact to affect the SE of $\hat{\lambda}_{i,1,(k)}$; MIXING and MODEL interact to affect the RMSE of $\hat{\lambda}_{i,1,(k)}$. These significant effects are elaborated in the following paragraphs.

According to Table 31, MODEL has a significant main effect with a large effect size on the systematic errors of the attribute main effect estimates, $\hat{\lambda}_{i,1,(k)}$, that

are quantified by bias (F=188.25, p<0.001, partial η^2=0.272). A visualization of the MODEL main effect in Figure 36 indicates that the LTA-longitudinal-MCDM has the lowest marginal mean bias of $\hat{\lambda}_{i,1,(k)}$ among the three models, while the Longitudinal LLM has the highest mean bias of $\hat{\lambda}_{i,1,(k)}$. Additionally, SIZE (F=19.74, p<0.001, partial η^2=0.038) and TR_Prob (F=9.89, p=0.002, partial η^2=0.019) are found to have significant main effects with small effect sizes on the bias of $\hat{\lambda}_{i,1,(k)}$.

As for the random errors, a significant two-way interaction of SIZE*MODEL (F=16.71, p<0.05, partial η^2=0.032) with a small effect size is found on the SE of $\hat{\lambda}_{i,1,(k)}$, as shown in Table 32. As visualized in Figure 37, the Longitudinal LLM has the lowest marginal mean SEs of $\hat{\lambda}_{i,1,(k)}$ among the three models in the either the large or small sample size (i.e., either J=800 or J=100) conditions. Moreover, a significant two-way interaction of MIXING*MODEL (F=5.66, p=0.016, partial η^2=0.011) with a small effect size is found on the RMSE of $\hat{\lambda}_{i,1,(k)}$. According to Figure 38, the LTA-longitudinal-MCDM has the lowest marginal mean RMSEs of $\hat{\lambda}_{i,1,(k)}$ among the three models in the conditions with either balanced or imbalanced initial mixing proportions of the strategies (i.e., either $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.6 : 0.4$ or $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.8 : 0.2$).

Table 31

Significant Effects in the Mixed-Effect ANOVA Results of the Bias of the Attribute Main Effect Estimates

Source	Bias of $\hat{\lambda}_{i,1,(k)}$		
	F Statistics	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)			
MODEL	188.25	<0.001	0.272
Between-Subject Effects			
SIZE	19.74	<0.001	0.038
TR_Prob	9.89	0.002	0.019

Note. TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); SIZE=Sample size (*J*); MODEL=Data-fitting model type. Only the recovery of the 22 attribute main effect parameters that are shared by all the three data-fitting models were compared with mixed-effect ANOVA.

Table 32

Significant Effects in the Mixed-Effect ANOVA Results of the SE and RMSE of the Attribute Main Effect Estimates

Source	SE of $\hat{\lambda}_{i,1,(k)}$			RMSE of $\hat{\lambda}_{i,1,(k)}$		
	F	*p*-value	Partial η^2	*F*	*p*-value	Partial η^2
Within-Subject Effects (with Greenhouse-Geisser Adjustment)						
MODEL	111.04	<0.001	0.181	202.27	<0.001	0.286
SIZE*MODEL	16.71	<0.001	0.032			
MIXING*MODEL				5.66	0.016	0.011
Between-Subject Effects						
SIZE	5270.63	<0.001	0.913	155.40	<0.001	0.236
TR_Prob				11.67	0.001	0.023

Note. MIXING=Initial mixing proportions of Strategy A and Strategy B ($\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); SIZE=Sample size (*J*); MODEL=Data-fitting model type. Only the recovery of the 22 attribute main effect parameters that are shared by all the three data-fitting models were compared with mixed-effect ANOVA.

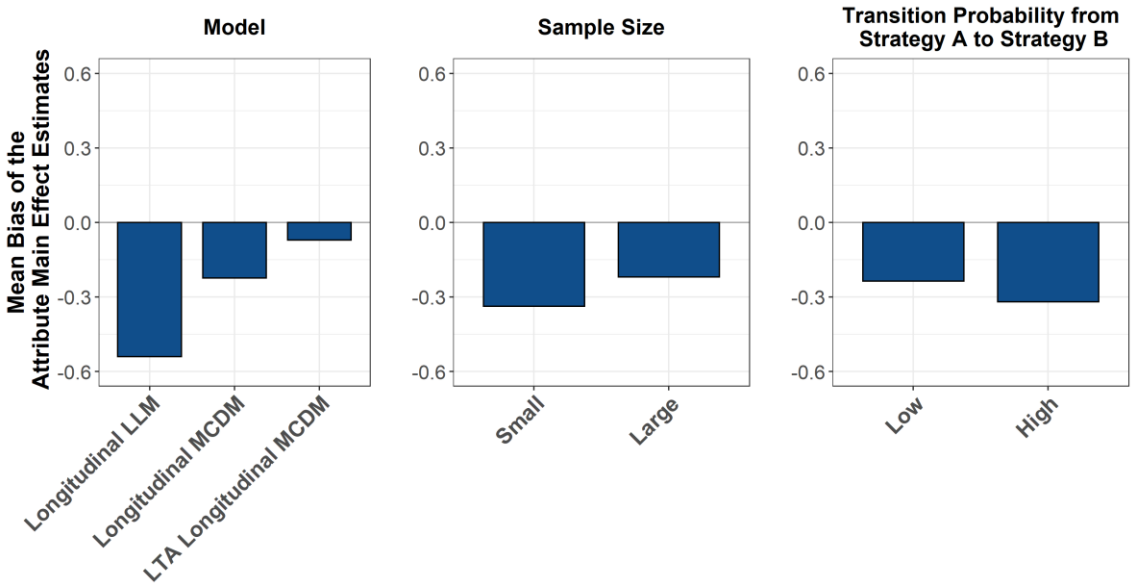


Figure 36. Significant main effects of MODEL, SIZE and TR_Prob on the bias of the attribute main effect parameter estimates, $\hat{\lambda}_{i,1,(k)}$, based on the 22 attribute main effect parameters that are shared by the Longitudinal LLM, Longitudinal MCDM and LTA-longitudinal-MCDM. [Note. MODEL=Data-fitting model type; SIZE=Sample size (J); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$).]

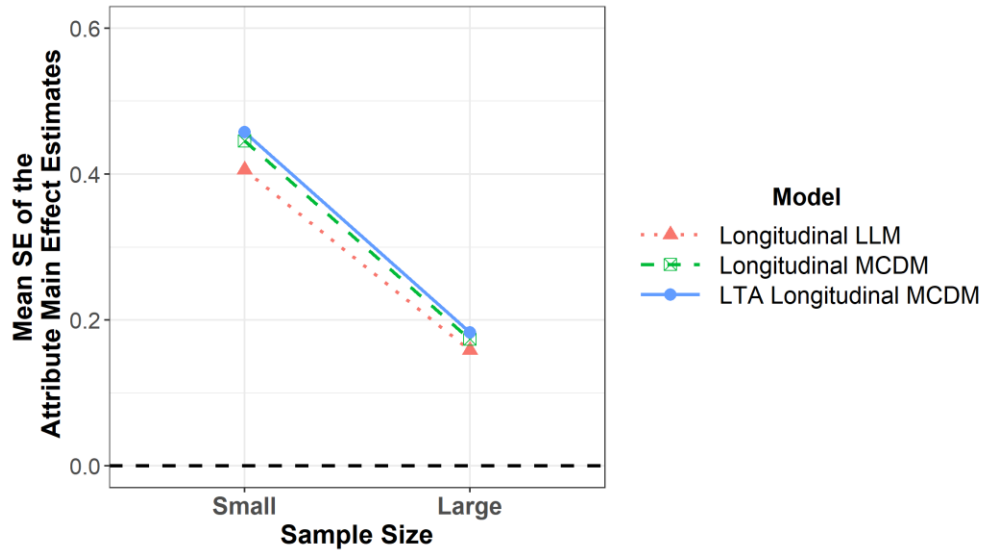


Figure 37. Significant two-way interaction of SIZE*MODEL on the SE of the attribute main effect parameter estimates, $\hat{\lambda}_{i,1,(k)}$, based on the 22 attribute main effect parameters that are shared by the Longitudinal LLM, Longitudinal MCDM and LTA-longitudinal-MCDM. [Note. MODEL=Data-fitting model type; SIZE=Sample size (J).]

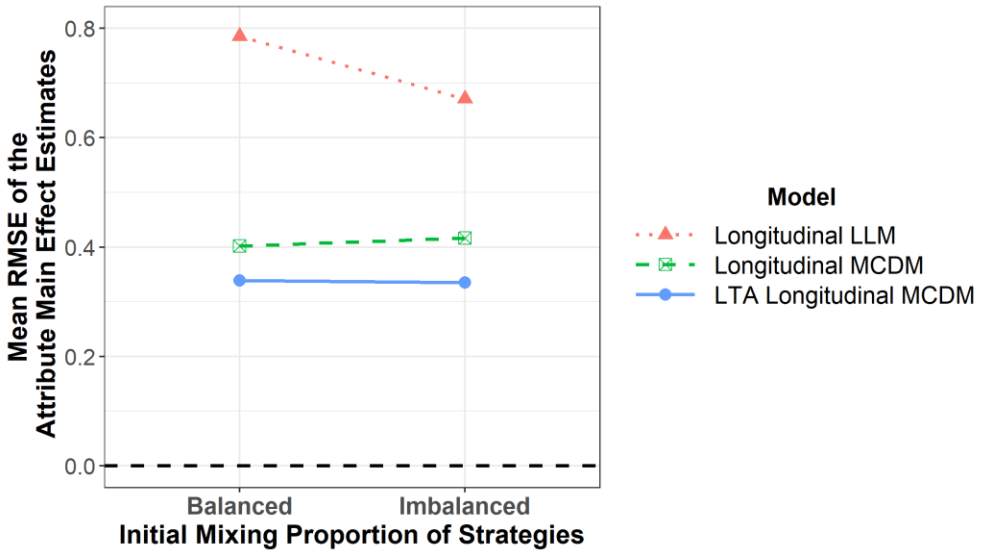


Figure 38. Significant two-way interaction of MIXING*MODEL on the RMSE of the attribute main effect parameter estimates, $\hat{\lambda}_{i,1,(k)}$, based on the 22 attribute main effect parameters that are shared by the Longitudinal LLM, Longitudinal MCDM and LTA-longitudinal-MCDM. [Note. MODEL=Data-fitting model type; MIXING=Initial mixing proportions of Strategy A and Strategy B ($\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$).]

4.3.2.2 Attribute main effect parameters only contained by the Longitudinal MCDM and LTA-longitudinal-MCDM

For the 13 attribute main effect parameters ($\lambda_{i,1,(k)}$) that are only contained by the Longitudinal MCDM and LTA-longitudinal-MCDM, the marginal mean bias, SE and RMSE of $\hat{\lambda}_{i,1,(k)}$ by the levels of each manipulated factor are plotted separately for the LTA-longitudinal-MCDM and Longitudinal MCDM (See Figures 39, 40 and 41), in order to investigate the effects of the neglect of the within-person strategy shift as well as the manipulated factors on the recovery of $\lambda_{i,1,(k)}$. In general, the LTA-longitudinal-MCDM yields lower marginal mean biases and RMSEs of $\hat{\lambda}_{i,1,(k)}$ than the Longitudinal MCDM does, but the two models produce similar marginal mean

SEs of $\hat{\lambda}_{i,1,(k)}$. Such results imply that, compared to the random errors of $\hat{\lambda}_{i,1,(k)}$, the systematic errors of $\hat{\lambda}_{i,1,(k)}$ may be more sensitive to the neglect of the within-person strategy shift in the model specification.

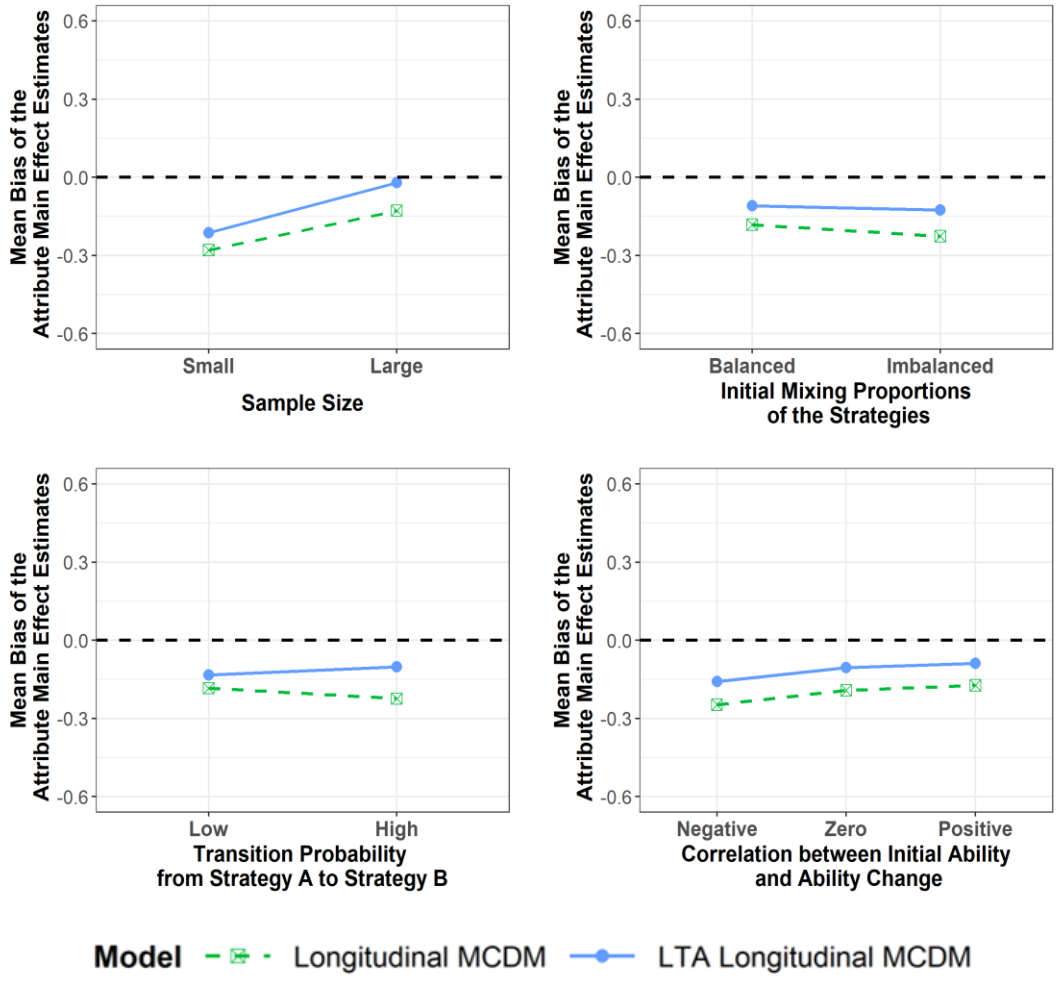


Figure 39. Marginal mean bias of the attribute main effect estimates, $\hat{\lambda}_{i,1,(k)}$, at each level of the manipulated factors, based on the 13 attribute main effect parameters that are only contained by the Longitudinal MCDM and LTA-longitudinal-MCDM.

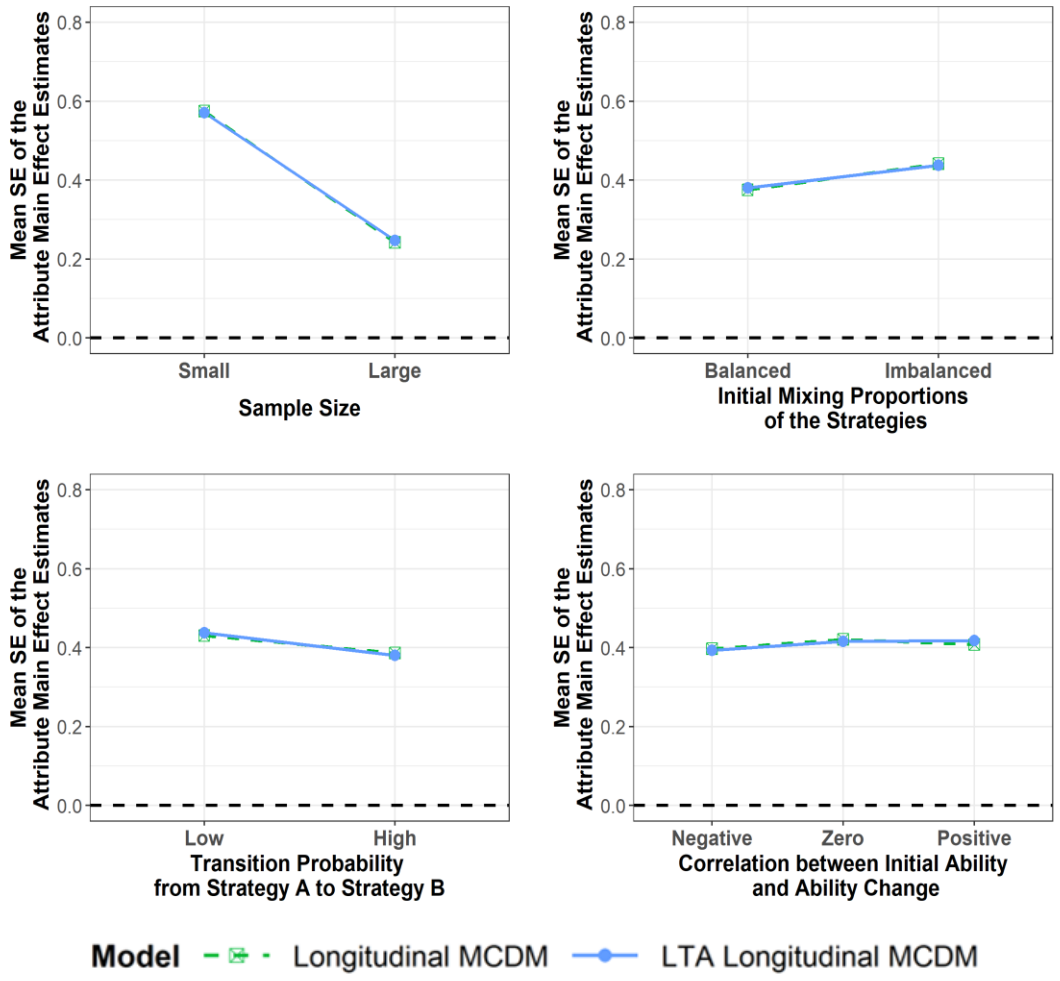


Figure 40. Marginal mean SE of the attribute main effect estimates, $\hat{\lambda}_{i,1,(k)}$, at each level of the manipulated factors, based on the 13 attribute main effect parameters that are only contained by the Longitudinal MCDM and LTA-longitudinal-MCDM.

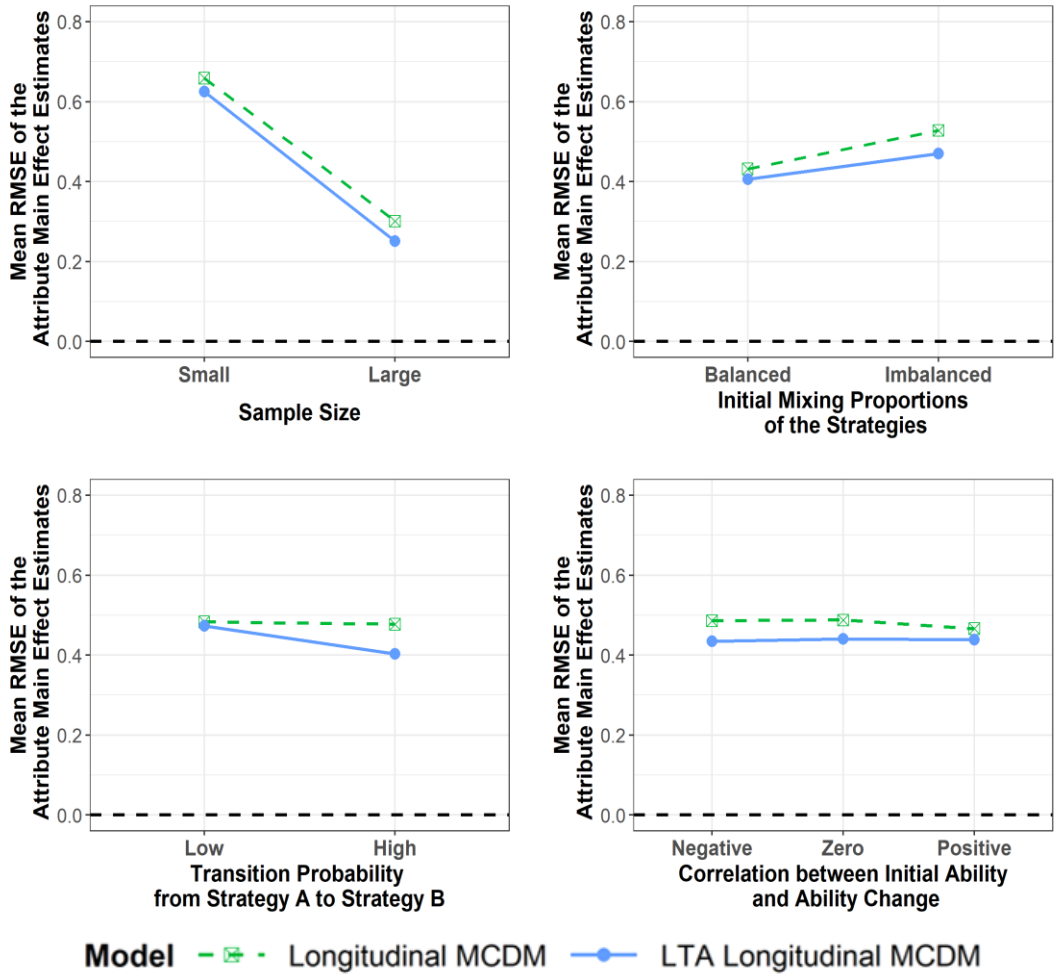


Figure 41. Marginal mean RMSE of the attribute main effect estimates, $\hat{\lambda}_{i,1,(k)}$, at each level of the manipulated factors, based on the 13 attribute main effect parameters that are only contained by the Longitudinal MCDM and LTA-longitudinal-MCDM.

4.3.2.3 The effects of the manipulated factors on the attribute main effect parameter recovery of the proposed model

To examine the effects of the manipulated factors on the attribute main effect parameter recovery of the proposed model, four-way ANOVAs were performed on the recovery outcome measures of the 35 attribute main effect parameters of the LTA-longitudinal-MCDM. As shown in Table 33, sample size (SIZE) has large effects on the bias (F=602.99, p<0.001, partial η^2=0.425), SE (F=3055.88, p<0.001,

partial η^2=0.789) and RMSE (F=3246.65, p<0.001, partial η^2=0.799) of $\hat{\lambda}_{i,1,(k)}$ from the proposed model. An inspection of the marginal means indicated that the smaller sample size conditions are associated with higher mean SE and RMSE of $\hat{\lambda}_{i,1,(k)}$. While a significant interaction between SIZE and CORR is found on the bias of $\hat{\lambda}_{i,1,(k)}$, the small sample size conditions (J=100) tend to have higher mean bias of $\hat{\lambda}_{i,1,(k)}$ than the large sample size conditions (J=800) at all the levels of correlation between the initial ability and ability change, despite the different magnitudes of differences.

Table 33

Significant Effects in the Four-Way ANOVA Results of the Recovery of the Attribute Main Effect Parameter from the LTA-longitudinal-MCDM

Source	Bias of $\hat{\lambda}_{i,1,(k)}$		SE of $\hat{\lambda}_{i,1,(k)}$		RMSE of $\hat{\lambda}_{i,1,(k)}$	
	p-value	Partial η^2	p-value	Partial η^2	p-value	Partial η^2
SIZE	<0.001	0.425	<0.001	0.789	<0.001	0.799
CORR	<0.001	0.023				
MIXING			<0.001	0.016	<0.001	0.016
TR_Prob			0.003	0.011	0.001	0.014
SIZE*CORR	0.001	0.017				

Effect Size	Small	Medium	Large
	(0.01≤partial η^2<0.06)	(0.06≤partial η^2<0.14)	(partial η^2≥0.14)

Note. CORR=Correlation between the initial ability and ability change ($\rho_{\theta^{(T_1)}\Delta\theta}$); MIXING=Initial mixing proportions of Strategy A and Strategy B ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B}$); TR_Prob=Transition probability from Strategy A to Strategy B ($p_{M_B|M_A}$); SIZE=Sample size (J); MODEL=Data-fitting model type.

4.4 Recovery of the Higher-Order Structural Parameters

The higher-order structural parameters, including the attribute easiness parameters (β_k) and the attribute discrimination parameters (ξ_k), characterize the relationship between the continuous skill implementation abilities ($\boldsymbol{\theta}_j$) and the discrete attribute mastery statuses ($\boldsymbol{\alpha}_j$) in the higher-order structure. In each model, each higher-order structural parameter corresponds to a specific attribute. Recall that the four attributes were simulated to have different levels of attribute easiness (i.e., true values of β_k were set at 1, 0.5, -0.5 and -1, for k = 1, 2, 3, 4), the implication of which is that the four attributes could be different in nature, thus the recoveries of β_k and ξ_k are summarized separately for each attribute.

4.4.1 Attribute easiness

In general, the effects of model specification on the marginal mean biases and RMSEs of $\hat{\beta}_k$ are inconsistent across different attributes and different levels of the manipulated factors, according to Figures 42 and 44. Specifically, for Attributes 1 and 3, the absolute values of the marginal mean biases and RMSEs of $\hat{\beta}_k$ (k=1, 3) estimated from the Longitudinal LLM are greater than those estimated from the LTA-longitudinal-MCDM or Longitudinal MCDM. For Attribute 4, the marginal mean biases and RMSEs of $\hat{\beta}_k$ (k=4) yielded by the three models are close to each other. However, when it comes to attribute 2, the Longitudinal LLM the has the lowest marginal biases and RMSEs of $\hat{\beta}_k$ (k=2) among the three models. For each attribute,

the magnitudes of discrepancies in the mean biases or RMSEs of $\hat{\beta}_k$ among the models vary across different levels of the manipulated factors.

In contrast to the inconsistent patterns observed for the biases and RMSEs of $\hat{\beta}_k$, the discrepancies in the marginal mean SEs of β_k among the models are relatively consistent and small, as seen in Figure 43. At all the levels of all the four manipulated factors, the Longitudinal LLM yields higher marginal mean SEs of $\hat{\beta}_k$ (k=3) than the other two models do for Attribute 3, while, for the other three attributes, the discrepancies in the marginal mean SEs of $\hat{\beta}_k$ (k=1,2,4) among the models are unobvious.

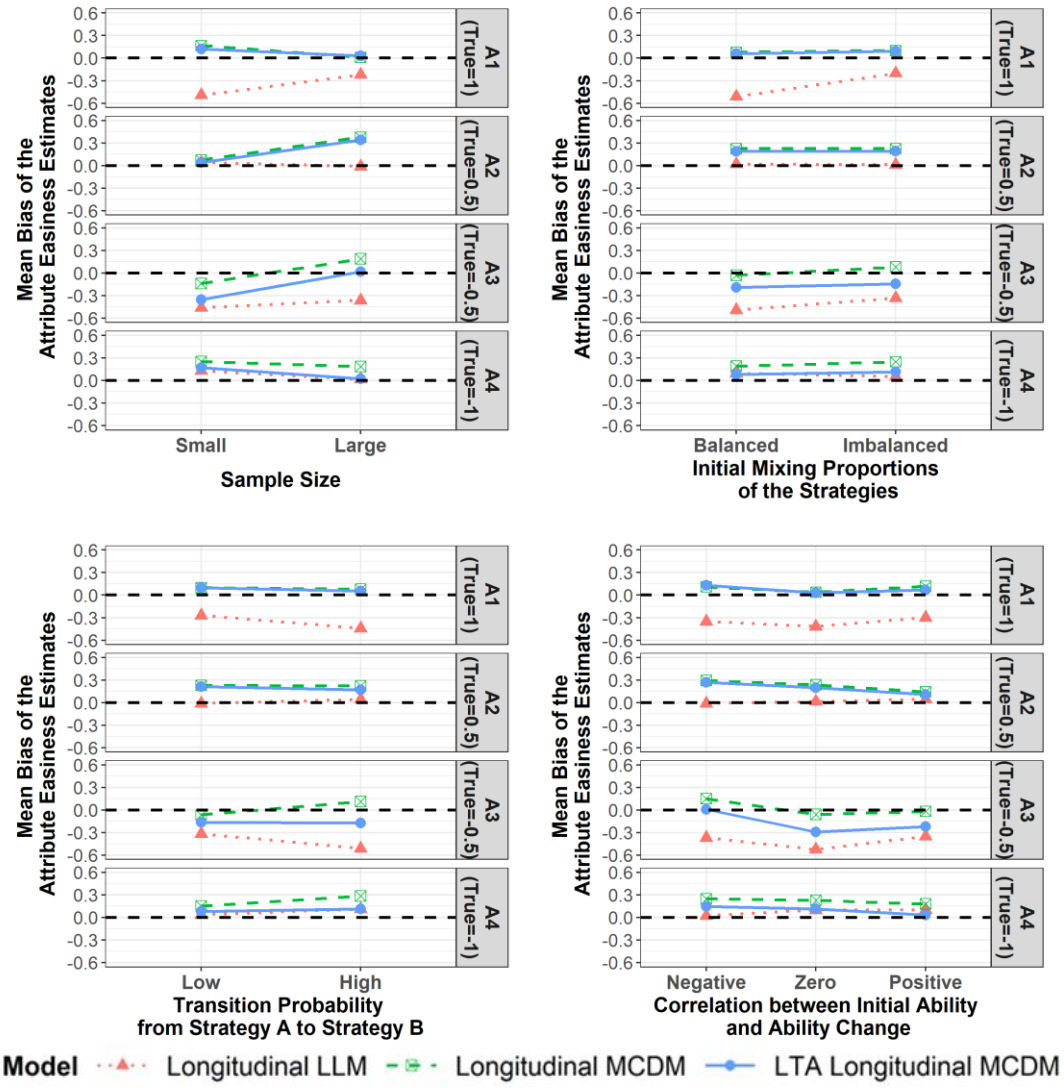


Figure 42. Marginal mean bias of the attribute easiness parameter estimates, $\hat{\beta}_k$, at each level of the manipulated factors. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute easiness parameters corresponding to different attributes. The easiness of an attribute being mastered increases as the attribute easiness parameter increases.

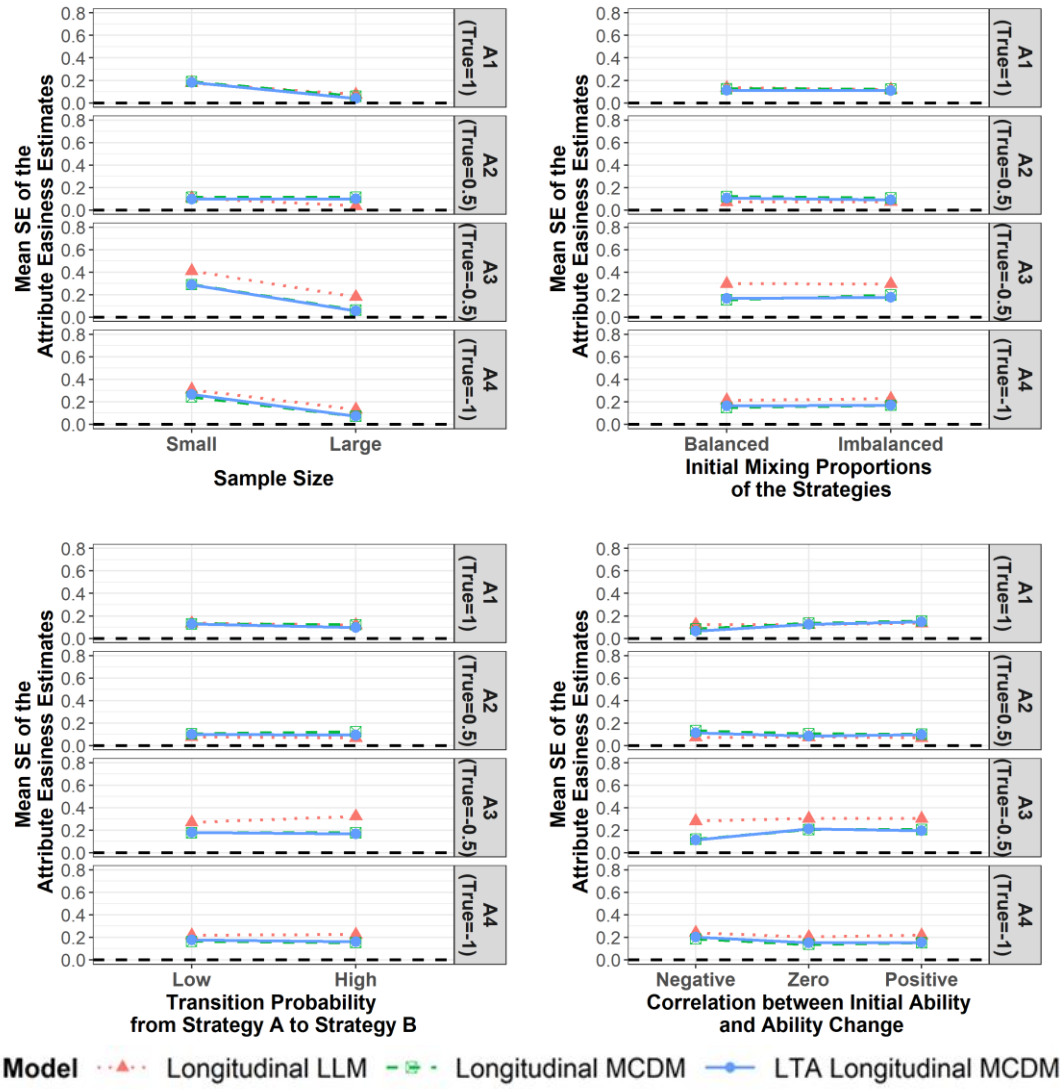


Figure 43. Marginal mean SE of the attribute easiness parameter estimates, $\hat{\beta}_k$, at each level of the manipulated factors. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute easiness parameters corresponding to different attributes. The easiness of an attribute being mastered increases as the attribute easiness parameter increases.

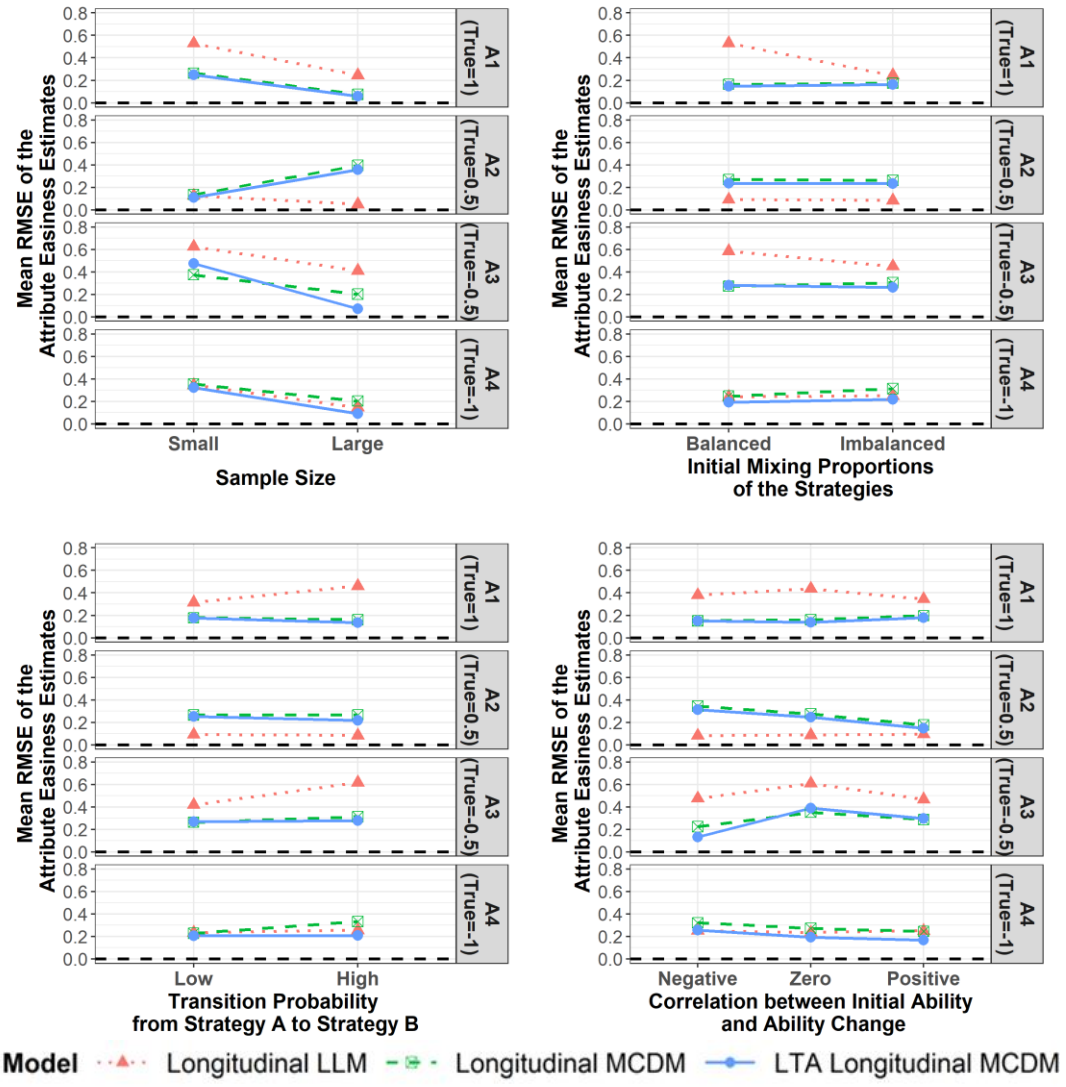


Figure 44. Marginal mean RMSE of the attribute easiness parameter estimates, $\hat{\beta}_k$, at each level of the manipulated factors. A1-A4 represent Attribute 1-Attribute 4. The values in the parentheses are the true values of the attribute easiness parameters corresponding to different attributes. The easiness of an attribute being mastered increases as the attribute easiness parameter increases.

4.4.2 Attribute discrimination

The patterns of the recovery of the attribute discrimination parameter (ξ_k) displayed in Figures 45, 46 and 47, are similar to those of the recovery of the attribute easiness parameters (β_k) presented above. As shown in Figures 45 and 47, for Attributes 1 and 3, the Longitudinal LLM has greater absolute values of marginal

mean biases and RMSEs for $\hat{\xi}_k$ (k=1, 3) than the LTA-longitudinal-MCDM or Longitudinal MCDM does; for Attribute 4, the marginal mean biases and RMSEs of $\hat{\xi}_k$ (k=4) yielded by the three models are close to each other; for Attribute 2, the Longitudinal LLM the has the lowest marginal biases and RMSEs of $\hat{\xi}_k$ (k=2) among the three models. The three models are close in terms of the marginal mean SEs of $\hat{\xi}_k$, according to Figure 46.

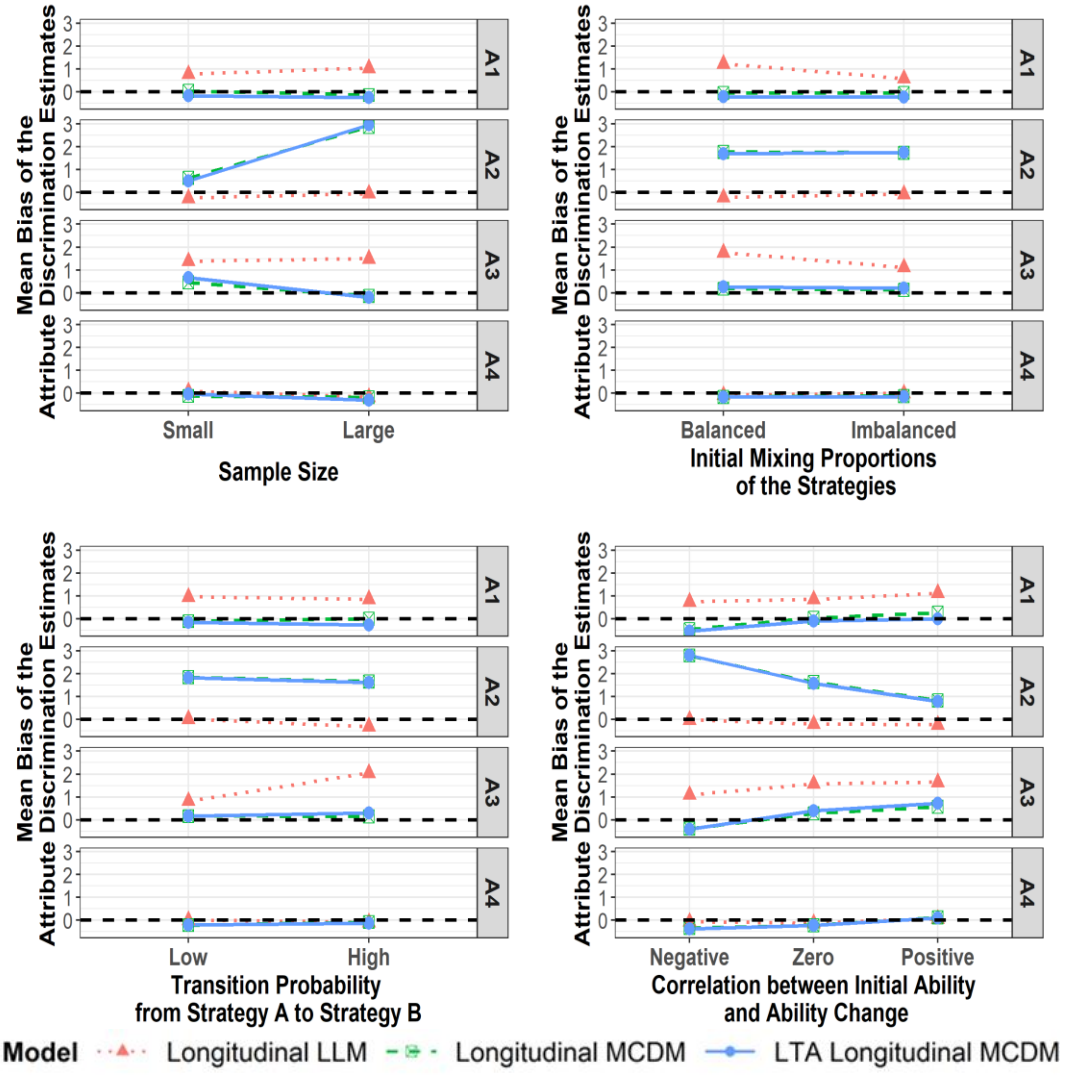


Figure 45. Marginal mean bias of the attribute discrimination parameter estimates, $\hat{\xi}_k$, at each level of the manipulated factors. A1-A4 represent Attribute 1-Attribute 4.

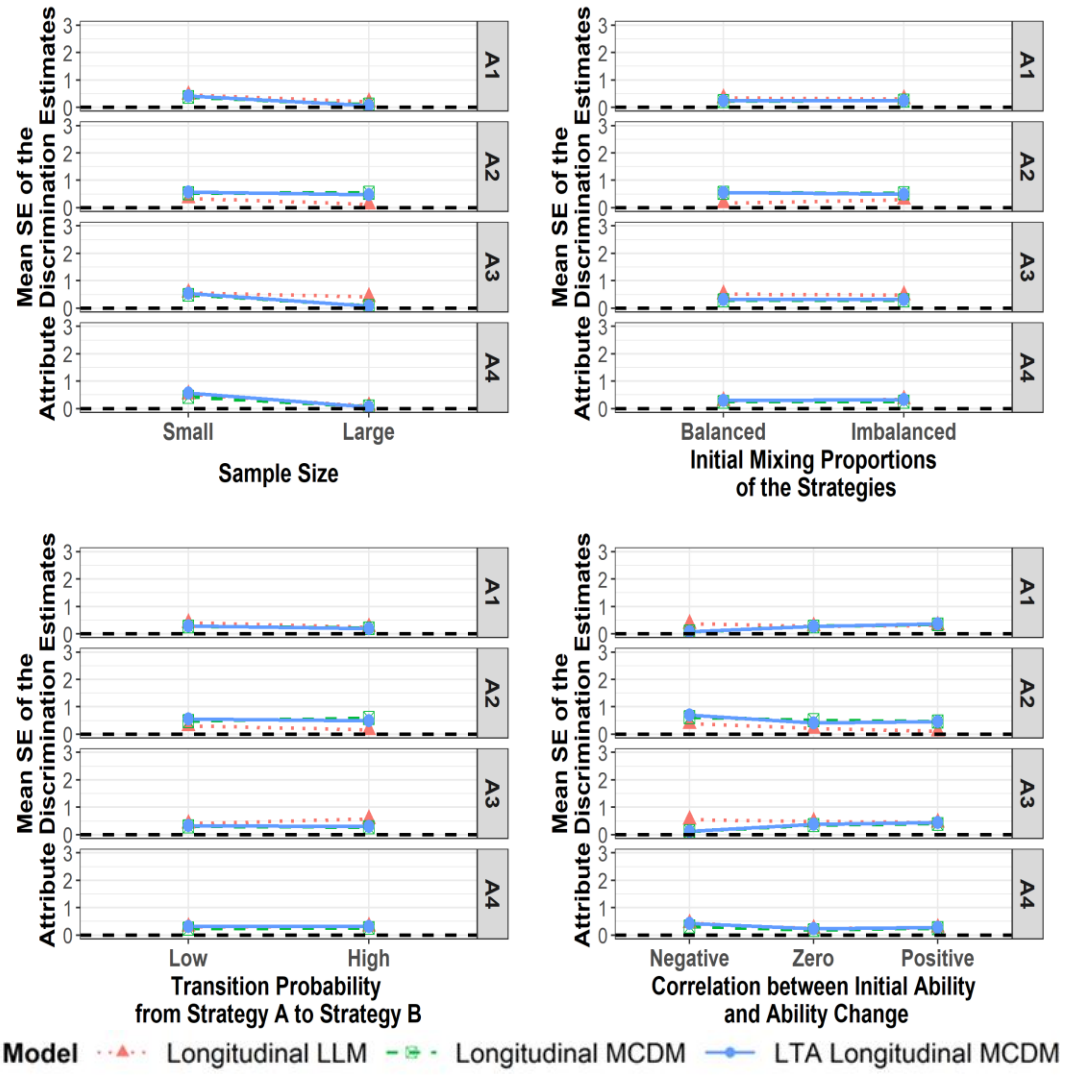


Figure 46. Marginal mean SE of the attribute discrimination parameter estimates, $\hat{\xi}_k$, at each level of the manipulated factors. A1-A4 represent Attribute 1-Attribute 4.

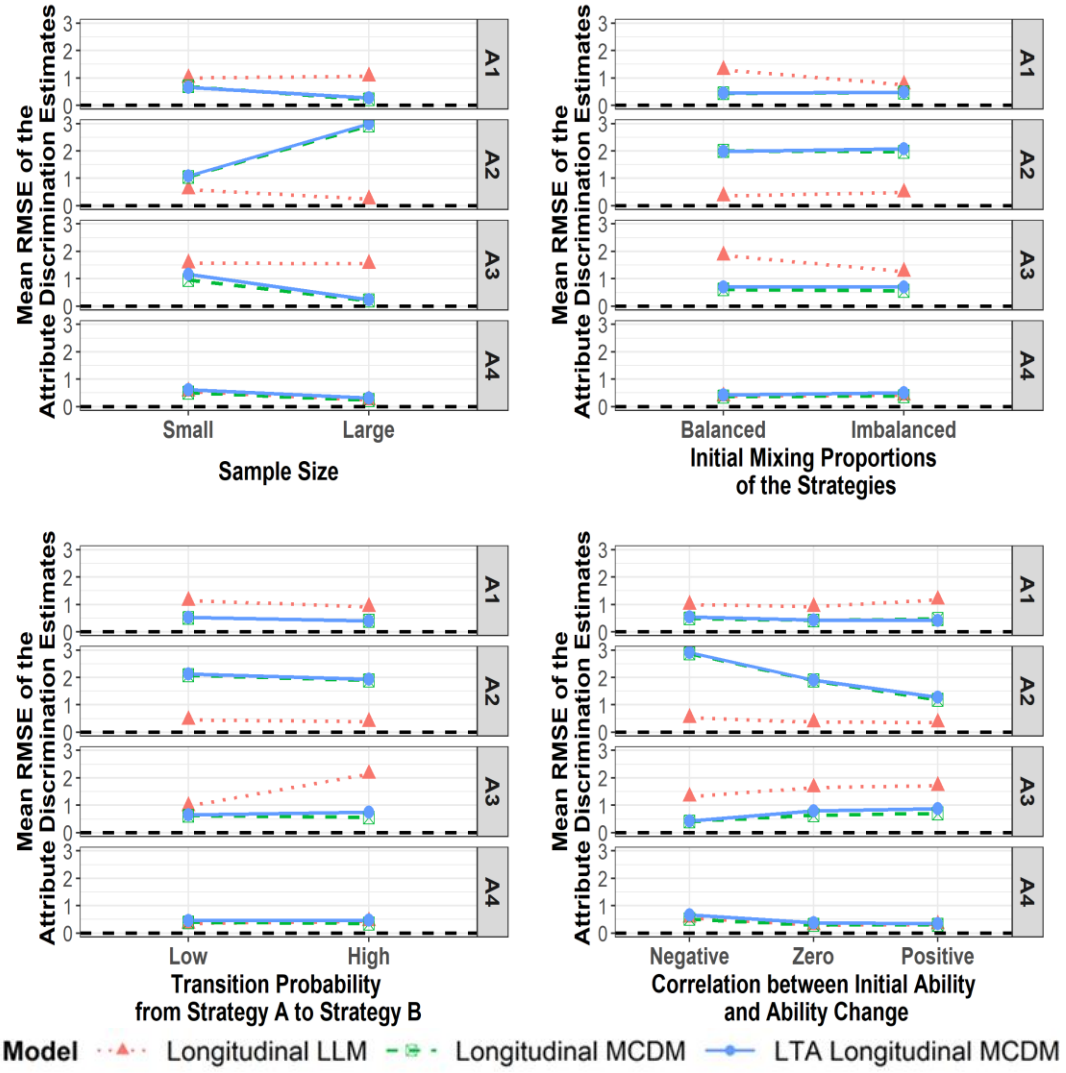


Figure 47. Marginal mean RMSE of the attribute discrimination parameter estimates, $\hat{\xi}_k$, at each level of the manipulated factors. A1-A4 represent Attribute 1-Attribute 4.

4.5 Summary of the Simulation Study Results

This section briefly recaps the key results presented in this chapter. The findings from the simulation study in response to each research question are summarized and discussed more extensively in Section 6.1.

The performance of the relative model fit indices in the presence of between-person multiple strategies and within-person strategy shift was examined by investigating the number of replications where each model fit index correctly

identified the LTA-longitudinal-MCDM as the best-fitting model (Table 13) and the number of replications where the evidence ratio of the LTA-longitudinal-MCDM as the best-fitting model to the alternative models being larger than 55 (Table 14). It was found that both AIC and BIC were able to correctly identify the LTA-longitudinal-MCDM as the best-fitting model in all the simulated conditions and nearly all the replications. The performance of DIC was more sensitive to the true initial mixing proportions of strategies and the strategy latent transition probability, i.e., the Longitudinal MCDM which ignores the within-person strategy shift tended to have lower DIC than the LTA-longitudinal-MCDM when the initial mixing proportions of strategies was balanced and the latent transition probability from Strategy A to Strategy B was low. However, the evidence ratio results indicated that the discrepancies in DIC between the Longitudinal MCDM as the best-fitting model and the proposed model as the second-best-fitting model were not significantly large.

The impact of ignoring the between-person multiple strategies and within-person strategy shift in the model on the parameter recovery of the longitudinal CDMs was examined by inspecting the marginal mean plots of the parameter recovery outcome measures of each model parameter against the data-fitting model types and the levels of the manipulated factors (See Figure 10 as an example). The effects of data-fitting model type on the parameter recovery shed light on the impact of ignoring the multiple-strategy scenarios on the parameter recovery of the longitudinal CDMs. Results from Figures 9 and 10 indicated that the attribute (profile) classification accuracy is reduced when between-person multiple strategies and/or within-person strategy shift is ignored in the model. For parameters with

sufficient sample size (i.e., the number of parameters of the same type being greater than 20), the mixed-effect ANOVAs were conducted to examine the statistical and practical significance of the effects of data-fitting model type on the parameter recovery. The results of the mixed-effect ANOVAs indicated that the data-fitting model type interacted with the correlation between the initial ability and ability change (CORR) and the true transition probability from Strategy A to Strategy B (TR_Prob) to affect the recovery of the first-level skill implementation ability parameters (See Table 16); and the data-fitting model type interacted with the sample size (SIZE) and initial mixing proportions of the strategies (MIXING) to affect the recovery of the item parameters (See Table 27).

The effects of the manipulated factors on the parameter recovery of the proposed model was examined by inspecting the marginal mean plots of the parameter recovery outcome measures of each model parameter of the proposed model against the levels of the manipulated factors (See the solid lines in Figure 10 as an example).The three-way or four-way ANOVAs were performed to investigate the significance of the manipulated factor effects on the recovery of the first-level skill implementation ability parameters and the item parameters of the proposed model. The ANOVA results indicated that both the correlation between the initial ability and ability change (CORR) and the transition probability from Strategy A to Strategy B (TR_Prob) have significant effects on at least one recovery outcome measure of the skill implementation ability parameters of the proposed model (See Tables 19 and 24); all the four manipulated factors have significant effects on at least one recovery

outcome measure of the item parameters of the proposed model (See Tables 30 and 33).

Chapter 5: Empirical Data Analysis Results

As an empirical data demonstration, the proposed model was applied to a dataset from a study (Bottge et al., 2015) that was designed to assess the effectiveness of the Enhanced Anchored Instruction (EAI; Bottge, 2001) and compare the effects of EAI to those of the business as usual (BAU) on students' problem-solving performance. The effectiveness study had a repeated-measure pretest-posttest design. Both the pretest and posttest consisted of 21 items measuring four attributes, including 1) ratios and proportional relationships (RPR), 2) measurement and data (MD), 3) number system – fractions (NSF) and 4) geometry – graphing (GG). Two empirical Q-matrices (see Table 7) that were learned from a data-driven nonparametric Q-matrix refinement method (Chiu, 2013) were used as inputs to the proposed model. Strategies corresponding to the two empirical Q-matrices were labeled as the empirical complex strategy ("the complex strategy") and the empirical simple strategy ("the simple strategy"), as items tended to load on more attributes in the former than the latter. The detailed dataset information and data analysis procedure can be found in Section 3.4.

This chapter presents the results of the empirical data analysis and consists of two sections. Section 5.1 documents the model fit indices of the data-fitting models and serves to justify the use of the empirical Q-matrices and the LTA-longitudinal MCDM to draw diagnostic inferences. Section 5.2 demonstrates the diagnostic information on the strategy choice, skill implementation ability and attribute mastery status drawn from the LTA-longitudinal-MCDM parameter estimates, which aims at addressing the two research questions, i.e.,

1) How do students' strategy choice, overall skill implementation ability and attribute mastery status change from the pretest to the posttest?
2) Do EAI and BAU differ in terms of their effects on students' learning outcomes regarding the strategy choice, overall skill implementation ability and attribute mastery status?

Results in this chapter are based on the testing dataset containing 749 students, 367 and 382 of whom have been assigned to the EAI and BAU instructional conditions, respectively.

5.1 Empirical Q-matrix Validation and Model Fit

According to the single-time-point analysis results (shown in Table 34), the S-MCDM-EE that utilizes the two empirical Q-matrices is identified by AIC, BIC and DIC as the best-fitting model (i.e., has the lowest relative model fit index) among the four competing single-time-point models in the posttest (See Table 8 for the detailed model specifications). Further, all the evidence ratios of the S-MCDM-EE to the other three single-time-point models derived from AIC, BIC and DIC are greater than 55 in the posttest, indicating that the discrepancies between the S-MCDM-EE and the other three single-time-point models, in terms of the relative model fit, are significant in the posttest. As for the pretest, the S-MCDM-EE has the lowest AIC and BIC among the four competing single-time-point models. All the evidence ratios of the S-MCDM-EE to the other three single-time-point models derived from AIC and BIC are greater than 55, except BIC evidence ratio of the S-MCDM-EE to the S-MCDM-TE. DIC favors the S-LLM-T that only utilizes the theoretical Q-matrix in the pretest.

Table 34
Model Fit Indices of the Single-Time-Point Models

Data	Model	Model fit index			
		AIC	BIC	DIC	PPP
Pretest	S-LLM-T	14143.18	14374.11	**16225.33**	0.560
	S-LLM-E	14154.19	14440.55	16314.70	0.564
	S-MCDM-TE	13618.31	13918.53	16757.78	0.638
	S-MCDM-EE	**13603.62**	**13917.69**	16598.59	0.592
Posttest	S-LLM-T	13950.2	14181.14	15370.32	0.685
	S-LLM-E	13869.22	14118.63	15283.10	0.616
	S-MCDM-TE	13670.86	13924.90	16137.86	0.686
	S-MCDM-EE	**13141.17**	**13455.24**	**15258.21**	0.637

Note. AIC=Akaike's information criterion; BIC=Bayesian information criterion; DIC=deviance information criterion; PPP=posterior predictive p-value. The lowest AIC, BIC and DIC values among the competing models are bolded.

As for the longitudinal analyses, the model fit indices are compared across the first four models listed in Table 35 (i.e., L-LLM-T, L-LLM-E-pre, L-LLM-E-post and L-MCDM-EE; see Table 9 for the detailed model specifications) to validate the empirical Q-matrices. The L-MCDM-EE that utilizes the two empirical Q-matrices is identified as the best-fitting model among the four competing longitudinal models by AIC and BIC. In addition, all the evidence ratios of the L-MCDM-EE to the other three models derived from AIC and BIC are greater than 55, supporting that the discrepancies in AIC and BIC between the L-MCDM-EE and the other three longitudinal models are significant. DIC favors the L-LLM-E-pre that only utilizes the empirical Q-matrix developed from the pretest. In sum, the model comparison results based on AIC and BIC support the use of the mixture of the two empirical Q-matrices in both the single-time-point and longitudinal analyses. DIC suggests the theoretical Q-matrix in the pretest in the single-time-point analyses and the empirical Q-matrix developed from the pretest in the longitudinal analyses. The mixture of the

two empirical Q-matrices is used for the subsequent LTA-longitudinal-MCDM as it is supported by the majority of the model fit indices assessed in this study.

Table 35
Model Fit Indices of the Longitudinal Models

Model	Model fit index			
	AIC	BIC	DIC	PPP
L-LLM-T	28301.48	28546.27	32082.86	0.559
L-LLM-E-pre	27869.05	28169.27	**31438.33**	0.582
L-LLM-E-post	28141.25	28404.52	31638.08	0.498
L-MCDM-EE	27259.21	27587.14	31728.62	0.549
LTA-L-MCDM-EE	**26841.77**	**27178.94**	32080.38	0.543

Note. AIC=Akaike's information criterion; BIC=Bayesian information criterion; DIC=deviance information criterion; PPP=posterior predictive p-value. The lowest AIC, BIC and DIC values among the competing models are bolded.

The LTA-longitudinal-MCDM is identified as the best-fitting model among the five longitudinal models by AIC and BIC, according to Table 35. The L-LLM-E-pre is identified as the best-fitting model by DIC. However, it should be noted that the simulation study results in Section 4.1 indicated that DIC may not be able to identify the LTA-longitudinal-MCDM as the best-fitting model even when multiple strategies exist in certain simulated conditions. As for the absolute model fit, the PPP value of the LTA-longitudinal-MCDM is 0.543, meaning that proportion of the replicated data generated from the proposed model having a sum of squares of standardized residuals that are greater than that of the observed data is 0.543. Such PPP value is not extremely close to 0, supporting that the observed data are likely to be seen in the replicated data if the LTA-longitudinal-MCDM is the true model. Thus, the PPP result provides a piece of evidence that the LTA-longitudinal-MCDM fits the empirical dataset adequately.

5.2 Diagnostic Inferences

This section demonstrates the diagnostic inferences drawn from the person parameter estimates of the LTA-longitudinal-MCDM. This study classifies the diagnostic information into three categories, i.e., strategy choice, skill implementation ability and attribute mastery. Therefore, the person parameters relevant to different categories are reported and interpreted separately. As an overview, the second-level person parameter estimates are listed Table 36. Since the item parameters and higher-order structural parameters are not the focus of this empirical data demonstration, the estimates of these parameters are supplied in Appendix B. Furthermore, findings from the statistical tests that compare the effects of EAI and BAU on students' learning outcomes in terms of strategy choice, skill implementation ability and attribute mastery are reported.

Table 36
Second-Level Person Parameter Estimates of the LTA-longitudinal MCDM

Parameter		Description	Estimate (SE)
Estimated parameters	$\mu_{\Delta\theta}$	Mean of the skill implementation ability change	0.51 (0.08)
	$\sigma^2_{\Delta\theta}$	Variance of the skill implementation ability change	1.25 (0.28)
	$\sigma_{\theta^{(T_1)}\Delta\theta}$	Covariance between the initial skill implementation ability and ability change	0.01 (0.12)
	$\pi^{(T_1)}_{M_{E,Complex}}$	Initial mixing proportion of the empirical complex strategy	0.43 (0.05)
	$\tau^{(T_1)}_{M_{E,simple}\|M_{E,complex}}$	Latent transition probability from the complex strategy to the simple strategy	0.37 (0.24)
	$\tau^{(T_1)}_{M_{E,complex}\|M_{E,simple}}$	Latent transition probability from the simple strategy to the complex strategy	0.63 (0.17)
Derived parameters	$\rho_{\theta^{(T_1)}\Delta\theta}$	Correlation between the initial skill implementation ability and ability change	0.01

	$\rho_{\theta^{(T_1)}\theta^{(T_2)}}$	Correlation between the initial skill implementation ability and the ability at the second timepoint	0.67

5.2.1 Strategy choice

The estimated strategy mixing proportions and latent transition probabilities are displayed in Table 37. The estimated initial strategy mixing proportions of the empirical complex and simple strategies ($\hat{\pi}_m^{(T_1)}$) are 0.429 and 0.571, respectively, meaning that, at the pretest, the expected percentages of students being classified into the complex and simple strategy latent classes are 42.9% and 57.1%, respectively. The estimated latent transition probability from the simple strategy to the complex strategy ($\hat{\tau}^{(T_1)}_{M_{E,simple}|M_{E,complex}}$) is 0.631, meaning that, the probability of transitioning to the complex strategy at the posttest conditional on the membership in the simple strategy latent class at the pretest is 0.631. In other words, among the students who are in the simple strategy latent class at the pretest, 63.1% are expected to be classified into the complex strategy latent class at the posttest. The estimated latent transition probability from the complex strategy to the simple strategy ($\hat{\tau}^{(T_1)}_{M_{E,simple}|M_{E,complex}}$) is 0.371, denoting that, among the students who are in the complex strategy latent class at the pretest, 37.1% are expected to be classified into the simple strategy latent class at the posttest. The strategy mixing proportions at the second timepoint ($\hat{\pi}_m^{(T_2)}$) are derived from the initial strategy mixing proportion and the latent transition probability estimates. The expected percentages of students being classified into the complex and simple strategy latent classes at the posttest are 63.2% and 36.8%, respectively.

Table 37

Strategy Mixing Proportion and Latent Transition Probability Estimates

Initial Strategy ($m^{(T_1)}$)	Initial strategy mixing proportions ($\pi_m^{(T_1)}$)	Strategy latent transition probability ($\tau_{m^{(T_2)}|m^{(T_2)}}^{(T_1)}$)	
		Complex ($M_{E,complex}^{(T_2)}$)	Simple ($M_{E,simple}^{(T_2)}$)
Complex ($M_{E,complex}^{(T_1)}$)	0.429	0.629	0.371
Simple ($M_{E,simple}^{(T_1)}$)	0.571	0.631	0.369
	Strategy mixing proportions at the second timepoint ($\pi_m^{(T_2)}$)	0.632	0.368

Four possible strategy choice trajectories, resulting from the four combinations of strategies at the pretest and posttest, are considered in this study. The four strategy choice trajectories are labeled as “complex to complex”, “complex to simple”, “simple to complex” and “simple to simple”. Each individual in the testing sample was classified into one of the four strategy trajectories. The distributions of the strategy choice trajectory classifications are summarized in Figure 48. Students who were classified into the “simple to complex” trajectory have taken up 51% of the testing sample. Nevertheless, less than 1% of the students were classified into the “complex to simple” trajectory. To examine whether the distributions of strategy choice trajectories differ across the BAU and EAI groups, a chi-square test for association was conducted. No significant association was found between the instructional condition and the distribution of the strategy choice trajectories.

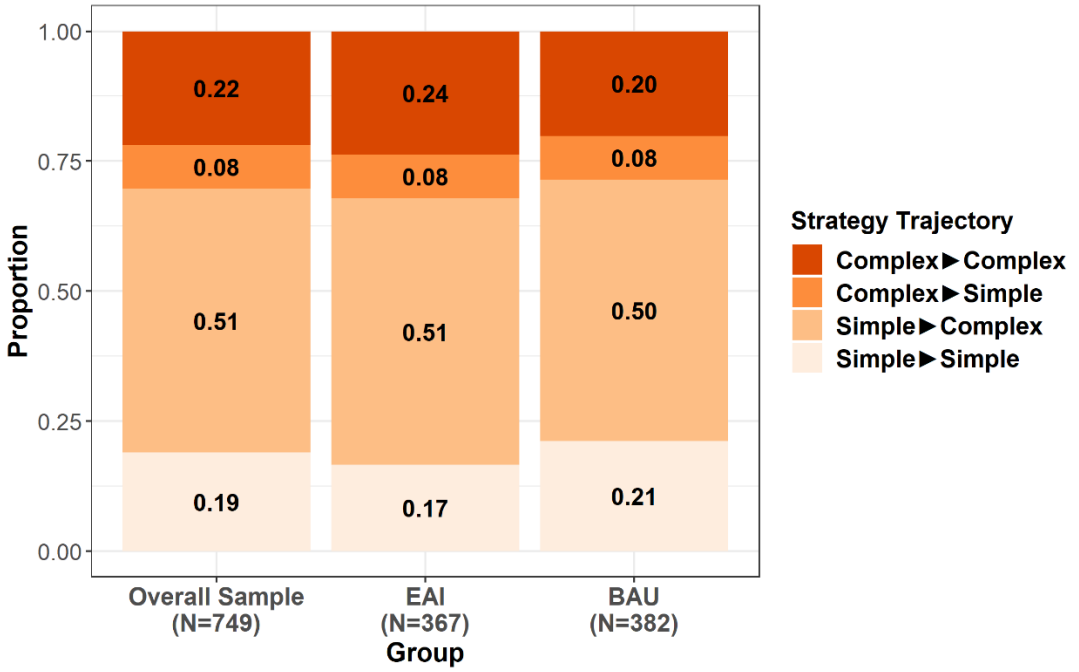


Figure 48. Distribution of the strategy choice trajectory classifications in the overall testing dataset and by instructional condition groups (EAI and BAU). EAI=Enhanced Anchored Instruction; BAU=Business as usual.

5.2.2 Skill implementation ability change

The estimated mean of the skill implementation ability change ($\hat{\mu}_{\Delta\theta}$) is 0.51, the 95% Bayesian credible interval of which is [0.36, 0.69]. As the 95% credible interval of $\hat{\mu}_{\Delta\theta}$ does not contain 0, the mean skill implementation ability change over time is statistically significant. Further, the means of the individual ability change estimates ($\Delta\hat{\theta}$) were compared across the EAI and BAU groups with an independent-samples *t*-test. Having confirmed that there is no severe assumption violation of the independent-samples *t*-test, i.e., no outliers, approximately normally-distributed residuals and the homogeneity of residual variances, a statistically significant difference was found between the EAI (M=0.63, SD=0.73) and the BAU (M=0.40, SD=0.73) groups in the mean ability change estimates (t=4.36, df=747,

p<0.001, Cohen's d=0.32). The distributions of the ability change estimates in the two instruction groups are plotted in Figure 49. It can be inferred that the average skill implementation ability growth for students in the EAI group is larger than that in the BAU group.

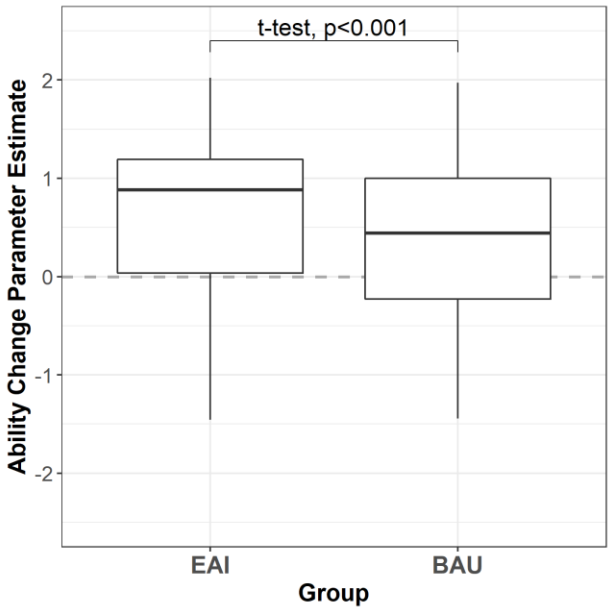


Figure 49. Distribution of the ability change parameter estimates in the EAI and BAU groups in the testing dataset. EAI=Enhanced Anchored Instruction; BAU=Business as usual.

The estimated variance of the ability change ($\hat{\sigma}^2_{\Delta\theta}$) is 1.25, which is larger than the variance of the initial ability ($\sigma^2_{\theta^{(T_1)}}$) that has been constrained at 1 for scale identification. The estimated covariance between the initial ability and ability change ($\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$) is 0.01. Thus, the derived correlation between the initial ability and ability change approximates zero. The derived correlation between the abilities at the first and second timepoints is 0.67, which is a moderate correlation. However, cautions should be taken to draw inferences from the covariance estimate between the initial

ability and ability change, as the simulation study results in Section 4.2 suggested that $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$ tends to be biased.

5.2.3 Attribute mastery status

Regarding each attribute, each individual has four possible mastery trajectories in the pretest-posttest scenario, i.e., non-mastery to non-mastery (0→0), non-mastery to mastery (0→1), mastery to non-mastery (1→0) and mastery to mastery (1→1). The distributions of the classified attribute mastery trajectories are summarized in Figure 50. The numbers labelled on the bars represent the proportions of students in the testing sample that are classified in particular attribute mastery trajectories. For instance, 30% of the students in the testing sample were classified as not mastering the ratios & proportional relationships (RPR) at the pretest but mastering the RPR at the posttest. Among the four attributes, RPR has the highest proportion of the "non-mastery to mastery" trajectory, followed by the geometry – graphing (GG); the measurement & data (MD) attribute has the lowest proportion of "non-mastery to mastery" trajectory. For each attribute, a small proportion (up to 0.12) of students have a "mastery to non-mastery" trajectory.

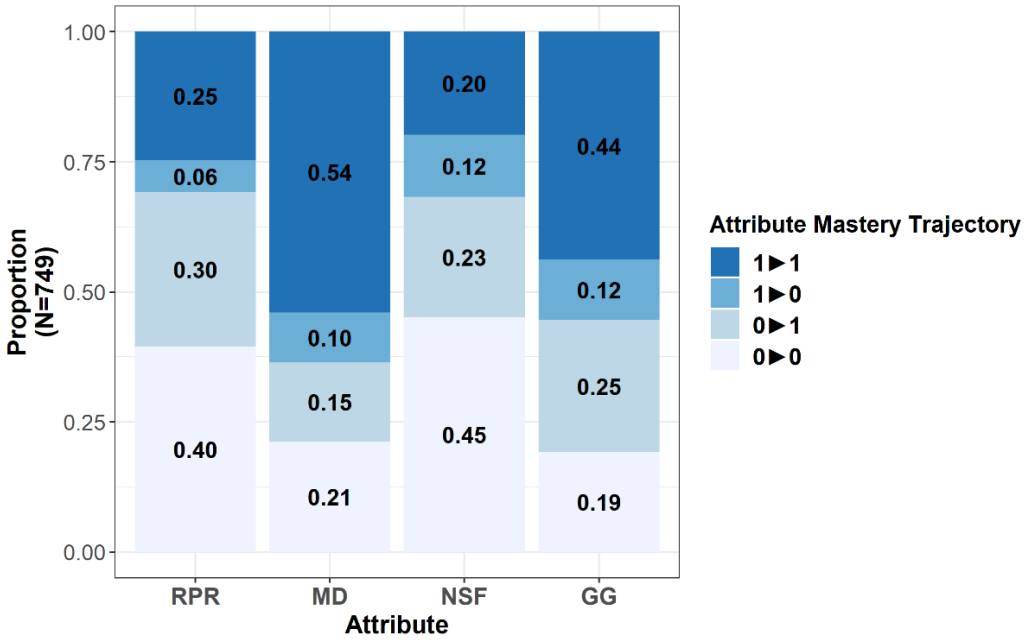


Figure 50. Distribution of the attribute mastery trajectory classifications in the testing dataset. RPR=ratios and proportional relationships; MD=measurement and data; NSF=number system – fractions; GG=geometry – graphing.

Figure 51 contrasts the proportions of the students with attribute non-mastery at the pretest being classified as attribute mastery at the posttest between the EAI and BAU groups. Note that the proportions in Figure 51 were calculated differently from those in Figure 50: the proportions in Figure 50 used the whole sample as the denominator, while the proportions in Figure 51 used those who were classified as attribute non-mastery at the pretest as the denominator. Specifically, the numbers labelled above each bar in Figure 51 clarify how the proportions were calculated, i.e., the number of students with attribute non-mastery at the pretest who were classified as attribute mastery at the posttest divided by the number of students who were classified as attribute non-mastery at the pretest. The associations between the proportions of students with attribute non-mastery-to-mastery transition and instructional condition (i.e., EAI or BAU) were examined with chi-square tests for association. As multiple chi-square tests were performed, one for each attribute, the

Dunn-Šidák correction (Šidák, 1967) was used to control the familywise error rate. With the Dunn-Šidák correction, the alpha level used for each chi-square test is 0.012, which corresponds to a familywise Type I error rate of 0.05. Statistically significant associations have been found between the instructional condition and the proportion of students with non-mastery-to-mastery transition on the ratios & proportional relationships (RPR; χ^2 =23.75, p<0.001, $\varphi = 0.21$) and geometry – graphing (GG; χ^2 =6.69, p=0.010, $\varphi = 0.14$) attributes. It can be seen from Figure 51 that, compared to the BAU group, the EAI group has higher proportions of students with non-mastery-to-mastery transition on RPR and GG.

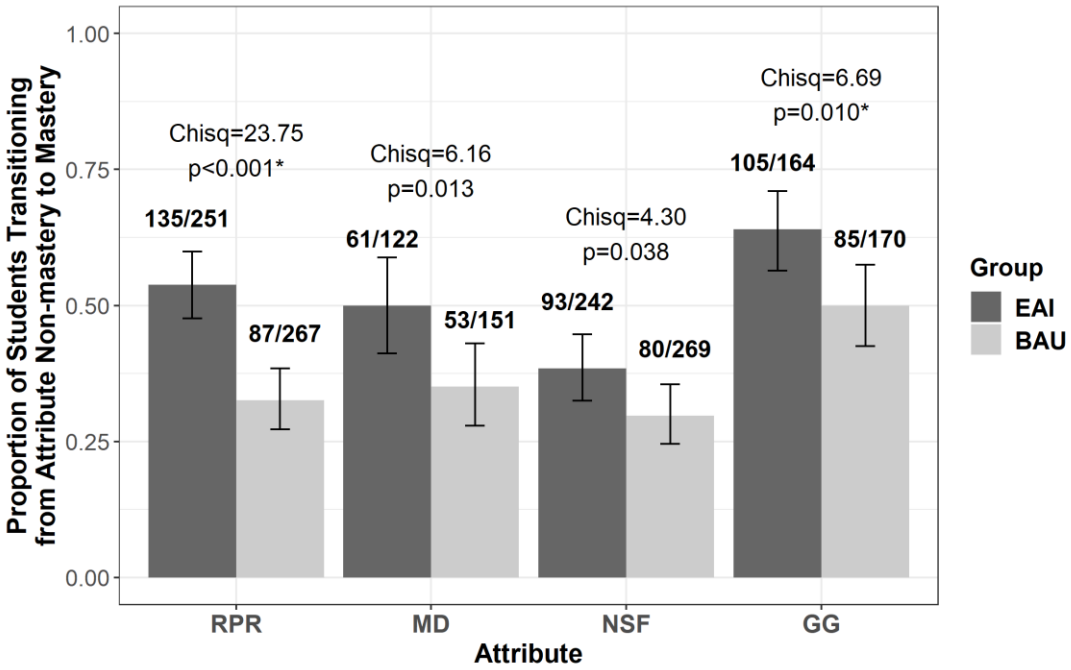


Figure 51. Proportion of students transitioning from attribute non-mastery to mastery (conditional on the non-mastery at the pretest) in the EAI and BAU groups. The numbers above each bar represent the number of students with attribute non-mastery at the pretest who were classified as attribute mastery at the posttest/the number of students who were classified as attribute non-mastery at the pretest. RPR=ratios and proportional relationships; MD=measurement and data; NSF=number system – fractions; GG=geometry – graphing; EAI=Enhanced Anchored Instruction; BAU=Business as usual. *p smaller than the Dunn-Šidák-corrected alpha level, 0.012.

Chapter 6: Discussion

As an increasing number of instructional programs are designed to improve students' problem solving (e.g., Bottge et al., 2003; Jitendra et al., 2002), there is an increasing need to evaluate the effectiveness of these programs from the perspective of students' problem-solving strategy shift. To this end, this study proposed the LTA-longitudinal-MCDM, which is a longitudinal CDM that can model both between-person multiple strategies and within-person strategy shift overtime. Compared to diagnostic inferences provided by the traditional longitudinal CDMs, what the proposed model provides is more informative and more relevant to problem solving: traditional longitudinal CDMs could only inform the change in students' attribute mastery status, while the proposed model can inform the change in students' strategy choice, skill implementation ability in addition to their attribute mastery status.

A simulation study was conducted to investigate the consequence of ignoring the multiple-strategy scenarios in the longitudinal CDMs and to examine the parameter recovery of the proposed model under various simulated conditions. Four factors were manipulated in the simulation study, including the sample size, the initial mixing proportions of strategies, the strategy latent transition probability and the correlation between the initial ability and ability change. The application of the proposed model to provide diagnostic inferences on students' strategy choice as well as skill implementation ability and attribute mastery status was demonstrated with an empirical data analysis. Sections 6.1 and 6.2 are arranged by the five research questions posed at the end of Section 1.2, summarizing the key findings from the

simulation study and the empirical data analysis in response to each research question. Limitations, implications and future directions are discussed in Section 6.3.

6.1 Findings from the Simulation Study

The simulation study was intended to examine the following three aspects, each of which serves to address a research question: 1) the performance of AIC, BIC and DIC in correctly selecting the LTA-longitudinal-MCDM as the best-fitting model in the presence of between-person multiple strategies and within-person strategy shift; 2) the impact of ignoring the multiple-strategy scenarios in the model on the parameter recovery of the longitudinal CDMs; and 3) the effect of the manipulated factors on the parameter recovery of the proposed model, i.e., the LTA-longitudinal-MCDM.

How do the relative model fit indices perform in the presence of between-person multiple strategies and within-person strategy shift? The performance of three commonly used model fit indices, i.e., AIC, BIC and DIC, in correctly selecting the LTA-longitudinal-MCDM as the best-fitting model in the presence of between-person multiple strategies and within-person strategy shift were evaluated. Both AIC and BIC were able to correctly identify the LTA-longitudinal-MCDM as the best-fitting model in all the simulated conditions and nearly all (at least 29 out of 30) the replications, while the performance of DIC was more sensitive to the manipulated factors that affect the strategy trajectory distribution in the population (i.e., the true initial mixing proportions of strategies and the strategy latent transition probability). Specifically, the Longitudinal MCDM that ignores the within-person strategy shift had the lowest DIC among the three data-fitting models in most replications when the

initial mixing proportions of strategies were balanced and the latent transition probability from Strategy A to Strategy B was low. Nevertheless, the discrepancies in DIC between the Longitudinal MCDM as the best-fitting model and the LTA-longitudinal-MCDM model as the second-best-fitting model were not significant according to the evidence ratio.

While the reasons why the performance of DIC in identifying the true model is undermined in certain conditions still need further explorations, there are some controversies in applying DIC to the mixture models. For example, according to the seminal paper by Spiegelhalter et al. (2002), DIC was not originally designed for the mixture models even though the possibility of extending DIC to the mixture models was mentioned in the paper. DeIorio and Robert (2002) indicated the lack of consistent definitions of DIC in the settings of the mixture models. Celeux et al. (2006) have explored several variations of DIC for mixture models and found that their performances in identifying the correct number of latent classes varied. McGrory and Titterington (2007) derived DIC based on a variational Bayes approach and found that this variation of DIC performed satisfactorily in choosing the correct number of latent classes. Thus, future studies could investigate the performance of the variations of DIC that are designed for latent class models in correctly identifying the proposed model under the simulated conditions.

What is the impact of ignoring between-person multiple strategies and/or within-person strategy shift on the parameter recovery of the longitudinal CDMs? Effects of ignoring between-person multiple strategies and within-person strategy shift on the model parameter recovery were examined by comparing the

parameter recovery outcome measures across different data-fitting models, including the LTA-longitudinal-MCDM that models both between-person multiple strategies and within-person strategy shift, the Longitudinal MCDM that ignores within-person strategy shift and the Longitudinal LLM that ignores both between-person multiple strategies and within-person strategy shift. On average, the LTA-longitudinal-MCDM had the highest classification accuracy of the attribute mastery profile, the lowest bias and RMSE of the item intercept parameter estimates and the lowest bias of the attribute main effect estimates among the three models. Such results implied that ignoring between-person multiple strategies and/or within-person strategy shift could lower the classification accuracy of the attribute mastery status profile and introduce errors to the item parameter estimates of the longitudinal CDMs.

The data-fitting model type also interacted with some manipulated factors to affect the recovery of certain parameters of the longitudinal CDMs. Notably, under the large sample size (J=800) conditions, there is an interaction between the data-fitting model type and the strategy latent transition probability ($p_{M_B|M_A}$) on the SE of the ability change parameter estimates ($\Delta\hat{\theta}_j$). Specifically, at both levels of the strategy transition probability (i.e., either $p_{M_B|M_A} = 0.3$ or $p_{M_B|M_A} = 0.7$), the mean SEs of $\Delta\hat{\theta}_j$ from the Longitudinal LLM were higher than those from the Longitudinal MCDM or LTA-longitudinal-MCDM, implying that ignoring the within-person strategy shift may result in an increase in the random errors of $\Delta\hat{\theta}_j$. Further, the magnitude of the difference in the mean SE of $\Delta\hat{\theta}_j$ between the Longitudinal LLM

and the other two models were larger in the higher strategy transition probability conditions ($p_{M_B|M_A} = 0.7$).

Moreover, the data-fitting model type interacted with both initial mixing proportions of strategies ($\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B}$) and strategy latent transition probability ($p_{M_B|M_A}$) to affect the strategy (trajectory) classification accuracy. The LTA-longitudinal-MCDM had a higher strategy choice trajectory classification accuracy than the Longitudinal MCDM, except under some conditions with balanced initial mixing proportions of the strategies and a low strategy transition probability (i.e., $\pi^{(1)}_{M_A} : \pi^{(1)}_{M_B} = 0.6{:}0.4$ and $p_{M_B|M_A} = 0.3$) where the Longitudinal-MCDM was slightly higher in strategy choice trajectory classification accuracy than the LTA-longitudinal-MCDM. One thing worth noticing is that the conditions under which the Longitudinal-MCDM slightly outperformed the LTA-longitudinal-MCDM in the strategy choice trajectory classification accuracy were almost the same as those under which DIC incorrectly identified the Longitudinal MCDM as the best-fitting model. Further explorations are needed to determine whether this is only a coincidence or there is some connection between the performance of DIC and the accuracy of strategy latent class classification.

How is the recovery of the parameters in the proposed model affected by the manipulated factors? Overall, each of the four manipulated factors were found to have significant effects on the recovery of at least one parameter of the proposed model, and the different factors affected different aspects of the parameter recovery. The effects of sample size on the item parameter recovery were of large effect sizes,

while the other significant effects were of small effect sizes. Significant effects found of each manipulated factor are elaborated below.

Sizes of 100 and 800 were considered as small and large sample sizes, respectively, in this study. One interested question is whether the parameter recovery of the proposed model would be problematic when the sample size is as small as 100. While a previous study by Cho et al. (2010) has supported that stable estimates can be obtained for the LTA-mixture Rasch model when the sample size was 100, it remained to be explored whether such a small sample size could yield stable estimates in the CDM counterpart. In general, results from this simulation study have shown that a small sample size (i.e., 100) is associated with a diminished recovery of the item parameters and some second-level person parameters in the proposed model. However, whether the parameter recoveries are deemed problematic may vary on a case-by-case basis, depending on the parameters of interest and the acceptable level of errors of the particular study. Specifically, the smaller sample size conditions were higher in the mean bias, SE and RMSE of the item intercept ($\hat{\lambda}_{i,0}$) and attribute main effect ($\hat{\lambda}_{i,1,(k)}$) estimates. Such findings are consistent with those from previous literature on longitudinal CDMs: Madison and Bradshaw (2018b), comparing sample sizes of 500 versus 2,000, found that the smaller sample size conditions had less accurate item parameter recovery as quantified by the median absolute deviation; Zhan, Jiao, Liao, et al. (2019), comparing sample sizes of 200 versus 500, found that the smaller sample size conditions were associated with higher mean bias and RMSE of the item parameters. As for the effects of sample size on the person parameters, the smaller sample size conditions tended to have higher mean SE and RMSE of the

mean estimates of ability change ($\hat{\mu}_{\Delta\theta}$) of the proposed model. While the proposed model has a different parameterization of ability from those longitudinal models proposed by Cho et al. (2010) and Zhan, Jiao, Liao, et al. (2019), i.e., the proposed model followed the Embretson-type parameterization (Embretson, 1991) while the other studies followed the Anderson-type parameterization (Andersen, 1985), findings about the effects of sample size on the recovery of mean ability parameter(s) are consistent across these studies. Furthermore, the small sample size conditions had higher SEs and RMSEs of the strategy latent transition probability estimates ($\hat{\tau}^{(T_1)}_{M_B|M_A}$) than the large sample size conditions. While no study in the CDM framework has been done to investigate the recovery of strategy transition probability, a similar effect of sample size on the latent class transition probability was found in the IRT framework. In particular, Cho et al. (2010), comparing sample sizes of 100, 1,000 and 3,000, found that the smaller sample size conditions tended to have higher RMSEs of the transition probability in the LTA-mixture Rasch model.

The initial mixing proportions of strategies and strategy latent transition probability were manipulated to simulate populations that vary on the distribution of strategy choice trajectories. Two levels of initial mixing proportions of strategies, i.e., 0.6:0.4 and 0.8:0.2, were chosen to mimic a balanced and an imbalanced initial population decomposition of strategy choice, respectively. Latent transition probabilities of 0.7 and 0.3 were selected to simulate a high and a low strategy transition probability, respectively. Both the initial mixing proportions of strategies and strategy latent transition probability were found to have small main effects on the SE and RMSE of the attribute main effect parameter estimates ($\hat{\lambda}_{i,1,(k)}$) of the

proposed model. In particular, the SE and RMSE of $\hat{\lambda}_{i,1,(k)}$ of the proposed model tended to be higher in the conditions with imbalanced initial mixing proportions of strategies than in those with balanced initial mixing proportions of strategies, and be higher in the low strategy transition probability conditions than in the high strategy transition probability conditions. Moreover, the strategy latent transition probability affected the recovery of the first-level person parameters of the proposed model under certain conditions: when the sample size was small, the mean SEs of $\hat{\theta}_j^{(T_1)}$ and $\Delta\hat{\theta}_j$ are higher in the lower strategy transition probability conditions.

In addition, this study has considered three levels of true correlation between the ability and ability change, i.e., negative ($\rho_{\theta^{(T_1)}\Delta\theta}$=-0.3), none ($\rho_{\theta^{(T_1)}\Delta\theta}$=0) and positive ($\rho_{\theta^{(T_1)}\Delta\theta}$=0.3), which corresponded to medium-to-high true correlations between the abilities at the two timepoints, ranging from 0.59 to 0.81. The true correlation between the ability and ability change were found to have small effects on the biases of the item parameters, including the item intercept ($\hat{\lambda}_{i,0}$) and attribute main effects ($\hat{\lambda}_{i,1,(k)}$). Furthermore, the correlation between the initial ability and ability change had small effects of the SE of $\hat{\theta}_j^{(T_1)}$ and $\Delta\hat{\theta}_j$; in the small sample size conditions, the correlation between the initial ability and ability change also had small effects of the bias and RMSE of $\Delta\hat{\theta}_j$.

6.2 Findings from the Empirical Data Analysis

To demonstrate that the LTA-longitudinal-MCDM is able to provide richer diagnostic information than the traditional longitudinal CDMs do, the LTA-

longitudinal-CDM was applied to an empirical dataset from an effectiveness study (Bottge et al., 2015), which has a repeated-measure pretest-posttest design, of the Enhanced Anchored Instruction (EAI; Bottge, 2001). Taking two empirical Q-matrices learned from a data-driven Q-matrix refinement method as input – one was labeled as the empirical complex strategy ("the complex strategy") and the other was labeled as the empirical simple strategy ("the simple strategy") – the LTA-longitudinal-MCDM was used to address two research questions: 1) How do students' strategy choice, overall skill implementation ability and attribute mastery status change from the pretest to the posttest? And 2) Do EAI and business as usual (BAU) instructional method differ in terms of their effects on students' learning outcomes regarding the strategy choice, overall skill implementation ability and attribute mastery status?

How do students' strategy choice, overall skill implementation ability and attribute mastery status change from the pretest to the posttest? Inferences about students' change in strategy choice were drawn from the estimated strategy mixing proportions at each timepoint and strategy latent transition probabilities. At the pretest, the expected percentages of students being classified into the complex and simple strategy latent classes were 42.9% and 57.1%, respectively; the corresponding percentages became 63.2% and 36.8% at the posttest. Further, according to the strategy latent transition probability estimates, $\hat{\tau}^{(T_1)}_{M_{E,simple}|M_{E,complex}} = 0.631$ and $\hat{\tau}^{(T_1)}_{M_{E,simple}|M_{E,complex}} = 0.371$, among the students who were in the simple strategy latent class at the pretest, 63.1% were expected to be classified into the complex strategy latent class at the posttest; among the students who were in the complex strategy

latent class at the pretest, 37.1% were expected to be classified into the simple latent class at the posttest. In sum, the majority of students chose the simple strategy at the pretest while the majority of students chose the complex strategy at the posttest. The probability of transitioning from the simple strategy to the complex strategy was higher than the other way around.

Inferences about the change in the overall skill implementation ability was drawn from the mean estimate of ability change ($\hat{\mu}_{\Delta\theta}$). A $\hat{\mu}_{\Delta\theta}$ of 0.51 with a 95% Bayesian credible interval not containing 0 implied that the increase in the mean skill implementation ability from the pretest to the posttest was statistically significant. Inferences about the change in attribute mastery status was obtained by summarizing the distributions of the classified attribute mastery trajectories for each attribute. The proportions of students having a "non-mastery to mastery" trajectory vary across attributes. The proportions of students having a "non-mastery to mastery" trajectory on the four attributes (from high to low) are: 0.30 for ratios & proportional relationships (RPR), 0.25 for geometry – graphing (GG), 0.23 for number system – fractions (NSF), and 0.15 for measurement and data (MD). In addition, a small proportion (up to 0.12) of students were found to have a "mastery to non-mastery" trajectory for each attribute, which may imply the existence of the forgetting effect.

Do EAI and BAU differ in terms of their effects of on students' learning outcomes regarding the strategy choice, overall skill implementation ability and attribute mastery status? In this study, the learning outcome of strategy choice is operationally defined as the distribution of strategy choice trajectory; the learning outcome of overall skill implementation ability is operationally defined as the ability

change estimates of the individuals; the learning outcome of the attribute mastery status is operationally defined as the proportion of attribute non-mastery students at the pretest who are classified as attribute mastery at the posttest. Results has shown that the EAI outperformed the BAU in terms of its effect on students' overall skill implementation ability and the mastery statuses of ratios & proportional relationships (RPR) and geometry – graphing (GG). Nevertheless, no significant difference between the EAI and BAU groups was observed on the learning outcomes regarding the mastery statuses of the measurement & data (MD) or number system – fractions (NSF) attribute, or in terms of the strategy choice. The results about attribute mastery status found in this study are consistent with those from studies that addressed a similar research question but with different methods (Bottge et al., 2014; Madison & Bradshaw, 2018a). Specifically, Bottge et al. (2014) and Madison and Bradshaw (2018a) found that, for RPR and GG attributes, the differences in the scores as well as the nonmastery-to-mastery transition probabilities between the EAI and BAU groups are statistically significant; no significant group difference was found for the MD or NSF attribute.

6.3 Limitations and Future Directions

As with any research study, this study has limitations, and there is room for future explorations. Nine aspects that worth further exploring are identified and elaborated below.

Single-time-point alternatives for strategy shift classification. While this study proposed a longitudinal model to model strategy shift, there could be some more time-efficient single-time-point alternatives if the skill implementation ability

change is not of interest. For example, in the empirical data analysis, an alternative way to figure out the strategy shift over time is to fit the single-time-point MCDM to the pretest and posttest data, separately, and then compare the strategy mixing proportions and strategy choice classifications over time. As an initial exploration, this single-time-point method was applied to the empirical dataset and the analysis results were compared with those from the LTA-longitudinal-MCDM. In the single-time-point analyses, the estimated mixing proportions of the complex and simple strategies are 0.40 and 0.60, respectively, at the pretest, and 0.64 and 0.36, respectively, at the posttest, the patterns of which were similar to those in the LTA-longitudinal-MCDM. Further, the strategy trajectory and attribute mastery profile trajectory classifications yielded from the single-time-point method and the LTA-longitudinal-MCDM were highly consistent (i.e., 88.8% of the students have the same strategy trajectory classifications using the two methods; 73.3% of the students have the same attribute mastery profile trajectory classifications using the two methods). Simulation studies could be conducted in the future to further explore the strategy classification accuracy as well as other aspects of the parameter recovery of the single-time-point methods for strategy shift classification.

The accuracy of the empirical Q-matrices. The empirical Q-matrices may have limited accuracy due to the limitations of the empirical Q-matrix development method. This study only employed one of many existing empirical Q-matrix development methods, and this method as well as the other existing empirical Q-matrix development methods assumed a correctly specified number of attributes and a single strategy. These assumptions remain unassessed and, if violated, the accuracy of

the resulting empirical Q-matrices could be diminished. Further, if the empirical Q-matrices could not accurately reflect the mapping relations between items and attributes, the accuracy of the diagnostic information drawn from the LTA-longitudinal-MCDM would be threatened, given that the Q-matrix misspecification could lead to a decrease in the attribute mastery classification accuracy (e.g., Rupp & Templin, 2008a).

Several measures could be considered in the future to enhance the accuracy of the empirical Q-matrices, which include: a) Complement the single empirical Q-matrix development method by trying out other Q-matrix development methods (e.g., de la Torre & Chiu, 2016; DeCarlo, 2012; Desmarais & Naceur, 2013), comparing the resulting empirical Q-matrices and using tree-based classification models to combine the results from different Q-matrix development methods (e.g., Desmarais et al., 2015; Xu & Desmarais, 2016). b) Validate the number of attributes in each Q-matrix and allow Q-matrices associated with different strategies to contain different sets of attributes. However, the successful implementation of such measures is contingent on the advances in the empirical Q-matrix development methods. Most existing empirical Q-matrix development methods, including the one employed by this study, requires the number of attributes to be pre-specified. Some matrix factorization methods have been used by Beheshti et al. (2012) to learn the number of attributes from the response data, which could potentially be combined with the empirical Q-matrix development methods in the future. c) Validate or explore the number of strategies. The empirical Q-matrix development method utilized in this study assumes that there is only one "correct" Q-matrix for an assessment, the

underlying assumption of which is that there is only a single strategy. Nevertheless, in the presence multiple strategies, the single empirical Q-matrix yielded from the existing methods may not be the "correct" Q-matrix; instead, it could be a result of a mixture of multiple "correct" Q-matrices. This study, having found different empirical Q-matrices from the repeated-measure pretest and posttest with the same empirical Q-matrix development method, adds to the possibility of the existence of multiple strategies. To the author's knowledge, no exploratory method is available currently to determine the number of "correct" Q-matrices that corresponds to the number of strategies. Therefore, there is a pressing need to develop such an empirical Q-matrix development method that can provide refinement suggestions on the number of "correct" Q-matrices.

The interpretability of the empirical Q-matrices. Interpreting an empirical Q-matrix has always been challenging, let alone making meaningful interpretation about multiple strategies from multiple empirical Q-matrices. Although it could be observed that one empirical Q-matrix was more complex (i.e., with items loading on more attributes) than the other, the meaning of the strategies corresponding to the two empirical Q-matrices remain unknown. The difficulty in identifying the meaningful strategies underlying the empirical Q-matrices makes it hard to operationally define the desired strategy choice learning outcome or evaluate the effectiveness of EAI in terms of students' strategy choice. For instance, while the empirical data analysis results showed that the transition probability from the empirical simple strategy to the empirical complex strategy is 0.63, one cannot tell whether such strategy transition is desirable or not with the current available information.

Expert opinions could be useful in improving the interpretability of the empirical Q-matrices. Experts on mathematical problem solving could be involved to inspect the empirical Q-matrices and determine whether these Q-matrices reflect any meaningful problem-solving strategies. In the long term, in order to gain more valid diagnostic inferences on strategy choice, it is recommended that test developers take multiple strategies into account in the early test development phase. For example, content experts could be asked to judge whether the items are expected to be solved with different strategies and whether students' strategy choices are expected to change after certain instructional interventions. If the answers to the questions above are "yes", multiple theoretical Q-matrices could be constructed, one for each strategy. Moreover, content experts could identify the impossible and/or desirable strategy trajectories, which can inform the setup of model constraints and provide criteria to evaluate the effectiveness of the intervention in terms of strategy choice.

Another issue relevant to the Q-matrix interpretation is the implication of modifying the Q-matrix for the item difficulty or complexity. In other words, what are the implications of an item loading on more (or fewer) attributes in the Q-matrix? Since the Q-matrix elements play different roles in the model equations of different CDMs, the implications of modifying the Q-matrix on the item properties vary across models. For instance, in the DINA model where, ideally, one could only correctly respond to an item when he or she masters all the required attributes of the item as specified in the Q-matrix, an item loading on more attributes in the Q-matrix implies that the item gets more complex since one needs to master more attributes in order to succeed. In contrast, in the DINO model where one could succeed as long as he or she

masters at least one of the required attributes of the item, an item loading on more attributes in the Q-matrix implies that the item gets easier. While the DINA and the DINO models represent the extreme cases where the required attributes are either conjunctive or disjunctive, the LLM utilized in this study falls somewhere in the middle of the conjunctive-disjunctive spectrum. Given that the LLM which is the measurement model utilized in this study assumes additive relations among the attribute main effects, the attributes are assumed to be the compensatory in the LLM (Templin & Hoffman, 2013). Since the attributes in the DINO model are also compensatory ("disjunctive" is a special case of "compensatory"), the implication of modifying the Q-matrix in the LLM is more aligned with the case of the DINO model: an item loading on more attributes in the Q-matrix implies that the item gets easier.

Measurement invariance assumption. The LTA-longitudinal-MCDM assume measurement invariance, meaning that the item response distributions conditional on the same strategy choice and attribute mastery pattern are identical, which further implies that the item parameters (i.e., item intercepts and attribute main effects) are assumed to be invariant over time. By fitting the LTA-longitudinal-MCDM, which constrained the item parameters to be equal across the two timepoints, to the empirical data, this study assumes the assessment used in the empirical data analysis to be measurement invariant over time. The measurement invariance assumption made in this empirical data analysis is largely attributed to a previous study by Madison and Bradshaw (2018a) who used the same empirical dataset to assess the measurement invariance of the same assessment and found that the item

parameter drift over time was not substantial. However, it is suggested that the measurement invariance assumption should be checked if the LTA-longitudinal-MCDM is to be applied to a different empirical dataset. The measurement invariance assumption could be assessed by comparing the model-data fit of an LTA-longitudinal-MCDM with all the item parameters constrained to be equal over time to an LTA-longitudinal-MCDM with time-specific item parameters. The latter having a significantly better model-data fit than the former could be a sign of violation to the measurement invariance assumption. Other methods for detecting differential item functioning, such as methods that are based on exploratory structural equation modeling (Marsh et al., 2009) and methods that utilize the regularization techniques with a penalty term (e.g., Bauer et al., 2020), could be adapt for the proposed model in the future to examine the measurement invariance assumption. In addition, simulation studies could be conducted to investigate the effects of violation to the measurement invariance assumption on the parameter recovery of the LTA-longitudinal-MCDM.

Over-specification of the number of strategies and the selection of the number of strategies. The simulation study focused on examining the effects of ignoring the multiple-strategy scenarios in the model (i.e., the under-specification of the number of strategies) on the performance of the model fit indices and the recovery of the CDM parameters (e.g., attribute mastery status classifications), given that most existing CDMs tended to under-specify the number of strategies. However, it would also be interesting to investigate the effects of over-specification of the number of strategies on the model selection and parameter recovery. One way of exploring the

effect of over-specification of the number of strategies is to fit the same data-fitting models as the simulation study, described in Section 3.3.4, to the simulated datasets in the absence of between-person multiple strategies and/or within-person strategy shift, and then compare the parameter recoveries across the data-fitting models.

Another perspective that worth further exploring is the performance of the information-based model fit indices in terms of correctly selecting the number of strategies. While this study has found that AIC and BIC outperformed DIC in selecting the true model when there are two strategies under most simulated conditions, it remains to be examined how these model fit indices perform when the true number of strategies is different. Given that selecting the number of strategies is analogous to selecting the number of latent classes in a finite mixture model, previous findings in the literature about using the information-based criteria in choosing the number of latent classes in the mixture models may shed light on future studies on choosing the number of strategies. For example, Steele and Raftery (2010) has compared the performance of AIC, BIC and DIC in selecting the number of latent classes in Gaussian mixture models with Bayesian estimation and found that BIC yielded the most accurate number of latent classes. In addition, several studies have found that AIC tended to overestimate the number of latent classes in the mixture models (e.g., Celeux & Soromenho, 1996; Koehler & Murphree, 1988; Steele & Raftery, 2010). The numbers of latent classes suggested by DIC were inaccurate under all the simulated conditions designed by Steele and Raftery (2010).

Model identification issue due to the relaxed correlation between the initial ability and ability change. The proposed model allows the covariance

between the initial ability and ability change to be freely estimated due to the interest in learning the relationship between the initial ability and ability change. Nevertheless, the ability structure shown in Figure 6 resembles an oblique version of the bifactor structure (Holzinger & Swineford, 1937) where the initial ability latent variable underlying all the attributes in both time points resembles the general factor and the ability change variable underlying only the attributes at the second time point resembles the specific factor. Mulaik and Quartetti (1997) have indicated that some model identification issues may arise when the general factor and the specific factor of a bifactor model are allowed to covary. While a common practice of facilitating the identification of the bifactor models is to constrain the general factor and the specific factors to be uncorrelated (e.g., Cai et al., 2011; Y. Li et al., 2006), the proposed model did not impose such a constraint considering the finding by Jeon et al. (2013) that ignoring the correlation between the general factor and the specific factors in a multigroup bifactor model could result in biased item parameter estimates and the lack of evidence supporting the ability change to be uncorrelated with the initial ability. In fact, different theories have suggested different directions of correlation between the initial ability and ability change, depending on factors such as the subject, content domain, analysis method and population. For instance, the existence of the ceiling effect yielded a negative correlation between the initial ability and ability change when analyzing the longitudinal data from cognitive aging study with regular growth curve analyses (L. Wang et al., 2008), while the Matthew effect observed in reading achievement where better readers gain greater improvement in their reading proficiency (Stanovich, 1986) suggested a positive correlation between

the initial ability and ability change. In the current study, the model identification issue due to the unconstrained correlation between the initial ability and ability change may manifest as the inaccurate estimates of the covariance between the initial ability and ability change. Thus, to mitigate the model identification issue due to the relaxed correlation between initial ability and ability change, future studies could consider constraining the correlation between the initial ability and ability change to a theoretical value or range after determining which theory best applies to the scenario under investigation.

Bayesian prior sensitivity analysis and posterior predictive model check. Due to the complexity of the proposed model and the alternative models, this study utilized relatively informative priors in the Bayesian MCMC estimation to facilitate the convergence of the model. Given that different choices of priors could affect the inferences drawn from the mixture models (e.g., Griffin, 2010; Miller & Harrison, 2018), future studies could experiment with other priors and investigate the sensitivity of the inferences drawn from the proposed model to the prior settings.

The posterior predictive model check was conducted as a measure of absolute model-data fit and the PPP values were calculated using the sum of squares residuals as the discrepancy measure. Given that different discrepancy measures delineate different aspects of the model and data, future studies could include other discrepancy measures to assess the adequacy of the model-data fit from other perspectives. Further, it should be noted that the power of the posterior predictive model check is highly dependent on the choice of the discrepancy measure, and the discrepancy measures of different power could be chosen for different study purposes, according

to Rubin (1996). Carlin and Louis (1996) indicated that the posterior predictive model check lacks power as the data were used twice in the model check process. The sample size and the PPP cut-off values could also affect the power of the posterior predictive model check. When the sample size is small, PPP could be sensitive to the priors and it remains unclear how it would affect the power of the posterior predictive model check (Berkhof et al., 2000). Given that no specific suggestion on the PPP cut-off value was found, this study rejects a model when the PPP value is lower than 0.05 which is a reasonable range from the frequentist perspective. However, it should be noted that a slight improvement in the model could bring a PPP into the acceptable range and that the PPP only measures the "statistical significance" of the difference between the data and the model (Gelman et al., 2003). To better decide whether a model should be rejected, future studies could further take into account the practical significance of the difference between the observed data and the model, which is, to a large extent, determined by the purpose and substantive interest of the study (Gelman et al., 2003).

Local item dependencies. In the LTA-longitudinal-MCDM, local item independence is assumed conditional on the strategy choice and attribute mastery status. However, several factors in the empirical dataset could potentially result in item dependencies and, thus, the violation of the local item independence assumption. On one hand, dependencies may exist among the repeated items across timepoints. As an initial exploration, an extension of the LTA-longitudinal-MCDM with latent variables accounting for residual dependence of the repeated items was fit to the empirical dataset used in Section 3.4. However, the model with cross-timepoint

dependencies suffered from slow convergence, rendering its comparison with the original LTA-longitudinal-MCDM infeasible. Causes of this convergence issue need to be further explored. On the other hand, local item dependencies could be present if multiple problem-solving items in the assessment formed a testlet and shared the same prompt. Future studies could extend the LTA-longitudinal-MCDM to account for item dependencies by incorporating testlet-specific latent variables.

Other extensions to the model. The LTA-longitudinal-MCDM explored in this study was limited to two timepoints, considered only the main effects of the attributes on the item response probabilities (i.e., used the LLM as the measurement model) and assumed the independence of skill implementation ability from strategy choice. Fortunately, the LTA-longitudinal-MCDM is flexible to be extended into more generalized forms and, in fact, the model equations have been written in more generalized forms – equation 4 has specified the measurement model as the LCDM that includes the interactions among attributes and equation 9 is applicable to scenarios with more than two timepoints. Further, when there are multiple time points, the choice of the reference point and scale could affect the parameter estimates of a growth model as demonstrated by Hancock and Choi (2006). To describe the growth trajectory in a more meaningful way, some scale-free statistics such as the relative aperture location proposed by Hancock and Choi (2006) could be calculated in the future. Specifically, in the case of this study, the aperture represents the time point where individuals are most similar in their true skill implementation ability, and locating the aperture and applying an intervention at the aperture could help maximize the effectiveness of the intervention (Hancock & Choi, 2006). The

assumption of independence between strategy choice and skill implementation ability could be relaxed by adding a higher-order latent variable that models the dependencies between the strategy choice parameter, *m*, and the skill implementation ability parameter, θ .

Despite the limitations, the contributions of this study to the CDM research literature, effectiveness evaluation, teaching and learning practices are significant. From the diagnostic modeling research perspective, the simulation study provided evidence that ignoring the multiple-strategy scenarios in the longitudinal CDMs, which is essentially a form of Q-matrix misspecification, could reduce the classification accuracy of the attribute mastery status. The proposed model, by considering multiple Q-matrices representing multiple strategies, can reduce the risk of Q-matrix misspecification due to the under-specification of multiple strategies. From the effectiveness evaluation perspective, for an instructional program to improve students' problem solving, it is not only important to train students' ability to implement the skills and help them achieve skill mastery, but it is also crucial to guide students to form effective strategies by choosing appropriate sets of skills to solve the problems (e.g., Afflerbach et al., 2008; Coughlin & Montague, 2011; Swanson, 2001). This study provides practitioners with a tool to evaluate the effectiveness of the instructional programs from both the strategy choice and skill implementation aspects. From the teachers and students' perspectives, the additional diagnostic information on strategy choice provided by the proposed model is useful to inform the teaching and learning practices.

Appendix A: Classification Accuracy, Bias, SE and RMSE Results by the Simulated Conditions

Table A. 1

Attribute Correct Classification Rate at Timepoint 1 (J=100)

			ACCR (Timepoint 1)											
			Attribute 1			Attribute 2			Attribute 3			Attribute 4		
$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
0.6:0.4	0.3	-0.3	0.74	0.96	0.97	0.96	0.98	0.98	0.88	0.89	0.90	0.89	0.90	0.91
		0	0.77	0.96	0.96	0.96	0.98	0.98	0.90	0.92	0.92	0.90	0.91	0.91
		0.3	0.84	0.97	0.96	0.95	0.97	0.97	0.91	0.92	0.92	0.89	0.90	0.92
	0.7	-0.3	0.72	0.95	0.96	0.96	0.97	0.98	0.89	0.86	0.90	0.89	0.88	0.92
		0	0.76	0.96	0.96	0.96	0.97	0.98	0.91	0.90	0.92	0.90	0.89	0.92
		0.3	0.82	0.96	0.96	0.96	0.97	0.97	0.92	0.89	0.93	0.89	0.88	0.92
0.8:0.2	0.3	-0.3	0.86	0.98	0.98	0.97	0.99	0.98	0.86	0.89	0.88	0.89	0.88	0.90
		0	0.89	0.98	0.98	0.98	0.98	0.98	0.89	0.89	0.90	0.89	0.90	0.90
		0.3	0.90	0.98	0.98	0.97	0.98	0.98	0.90	0.89	0.91	0.89	0.89	0.90
	0.7	-0.3	0.84	0.96	0.98	0.97	0.98	0.99	0.87	0.84	0.88	0.89	0.86	0.90
		0	0.87	0.97	0.98	0.98	0.98	0.98	0.90	0.86	0.91	0.89	0.88	0.90
		0.3	0.89	0.97	0.98	0.97	0.98	0.98	0.91	0.88	0.92	0.90	0.87	0.91

Note. ACCR=Attribute Correct Classification Rate; L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 2

Attribute Correct Classification Rate at Timepoint 1 (J=800)

			ACCR (Timepoint 1)											
			Attribute 1			Attribute 2			Attribute 3			Attribute 4		
$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)} \Delta \theta}$	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
0.6:0.4	0.3	-0.3	0.83	0.96	0.96	0.97	0.98	0.98	0.91	0.91	0.92	0.90	0.91	0.93
		0	0.85	0.96	0.96	0.97	0.98	0.98	0.91	0.91	0.92	0.90	0.91	0.92
		0.3	0.84	0.96	0.96	0.97	0.98	0.98	0.92	0.91	0.92	0.91	0.92	0.93
	0.7	-0.3	0.81	0.95	0.96	0.97	0.98	0.98	0.92	0.87	0.92	0.90	0.88	0.93
		0	0.84	0.95	0.96	0.97	0.98	0.98	0.92	0.87	0.92	0.89	0.88	0.93
		0.3	0.83	0.96	0.96	0.97	0.98	0.98	0.93	0.87	0.92	0.90	0.89	0.93
0.8:0.2	0.3	-0.3	0.92	0.97	0.97	0.98	0.99	0.99	0.90	0.88	0.90	0.90	0.89	0.91
		0	0.93	0.97	0.97	0.98	0.99	0.99	0.90	0.89	0.90	0.90	0.89	0.91
		0.3	0.92	0.97	0.97	0.98	0.99	0.99	0.91	0.89	0.91	0.91	0.89	0.91
	0.7	-0.3	0.91	0.96	0.97	0.98	0.98	0.99	0.91	0.85	0.90	0.90	0.85	0.91
		0	0.92	0.97	0.97	0.98	0.98	0.99	0.91	0.85	0.90	0.90	0.85	0.91
		0.3	0.91	0.97	0.97	0.98	0.98	0.99	0.92	0.85	0.91	0.90	0.86	0.92

Note. ACCR=Attribute Correct Classification Rate; L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 3

Attribute Correct Classification Rate at Timepoint 2 (J=100)

			ACCR (Timepoint 2)											
			Attribute 1			Attribute 2			Attribute 3			Attribute 4		
$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)} \Delta \theta}$	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
0.6:0.4	0.3	-0.3	0.77	0.94	0.95	0.95	0.97	0.97	0.88	0.92	0.91	0.88	0.92	0.92
		0	0.85	0.96	0.96	0.94	0.96	0.97	0.92	0.93	0.93	0.89	0.92	0.93
		0.3	0.81	0.95	0.96	0.94	0.96	0.97	0.90	0.92	0.92	0.90	0.94	0.94
	0.7	-0.3	0.66	0.91	0.93	0.94	0.97	0.97	0.90	0.94	0.94	0.88	0.95	0.95
		0	0.76	0.92	0.95	0.93	0.96	0.97	0.93	0.94	0.94	0.90	0.95	0.97
		0.3	0.69	0.90	0.94	0.94	0.96	0.97	0.93	0.95	0.95	0.91	0.96	0.97
0.8:0.2	0.3	-0.3	0.87	0.96	0.97	0.95	0.97	0.98	0.86	0.89	0.91	0.88	0.90	0.91
		0	0.90	0.96	0.96	0.94	0.96	0.97	0.91	0.92	0.93	0.88	0.90	0.92
		0.3	0.81	0.94	0.96	0.94	0.97	0.97	0.91	0.91	0.91	0.90	0.92	0.92
	0.7	-0.3	0.72	0.92	0.94	0.93	0.96	0.97	0.89	0.92	0.94	0.89	0.94	0.95
		0	0.78	0.92	0.94	0.92	0.95	0.96	0.93	0.94	0.95	0.89	0.94	0.94
		0.3	0.72	0.90	0.94	0.92	0.95	0.96	0.92	0.93	0.94	0.91	0.94	0.94

Note. ACCR=Attribute Correct Classification Rate; L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 4

Attribute Correct Classification Rate at Timepoint 2 (J=800)

			ACCR (Timepoint 2)											
			Attribute 1			Attribute 2			Attribute 3			Attribute 4		
$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)} \Delta \theta}$	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
0.6:0.4	0.3	-0.3	0.82	0.96	0.96	0.96	0.98	0.98	0.92	0.93	0.93	0.88	0.93	0.93
		0	0.82	0.95	0.96	0.96	0.98	0.98	0.92	0.93	0.93	0.90	0.93	0.93
		0.3	0.83	0.95	0.96	0.96	0.98	0.98	0.92	0.94	0.94	0.91	0.94	0.94
	0.7	-0.3	0.70	0.93	0.95	0.95	0.98	0.98	0.94	0.95	0.96	0.88	0.95	0.95
		0	0.71	0.93	0.95	0.95	0.98	0.98	0.94	0.95	0.96	0.90	0.95	0.95
		0.3	0.73	0.93	0.95	0.96	0.98	0.98	0.94	0.96	0.96	0.90	0.96	0.96
0.8:0.2	0.3	-0.3	0.87	0.95	0.96	0.96	0.98	0.99	0.91	0.92	0.92	0.89	0.91	0.92
		0	0.87	0.95	0.97	0.96	0.98	0.99	0.91	0.92	0.92	0.90	0.92	0.93
		0.3	0.87	0.95	0.97	0.97	0.98	0.99	0.91	0.93	0.93	0.91	0.93	0.93
	0.7	-0.3	0.77	0.93	0.95	0.94	0.97	0.98	0.93	0.94	0.95	0.88	0.94	0.95
		0	0.77	0.93	0.95	0.95	0.97	0.98	0.93	0.94	0.95	0.90	0.94	0.95
		0.3	0.76	0.92	0.95	0.95	0.97	0.98	0.93	0.94	0.96	0.91	0.94	0.96

Note. ACCR=Attribute Correct Classification Rate; L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 5

Bias of the Initial Ability and Ability Change Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Bias of $\hat{\theta}_j^{(T_1)}$			Bias of $\Delta\hat{\theta}_j$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	-0.003	0.005	0.004	0.031	0.089	0.108
			0	-0.006	-0.002	-0.002	0.031	0.056	0.108
			0.3	-0.004	-0.002	-0.002	-0.188	-0.167	-0.126
		0.7	-0.3	-0.009	0.004	0.001	-0.115	0.061	0.115
			0	-0.011	-	-0.005	0.011	0.040	0.107
			0.3	-0.009	0.002	-0.007	-0.228	-0.179	-0.126
	0.8:0.2	0.3	-0.3	-0.002	0.003	0.003	0.072	0.114	0.122
			0	-0.004	0.002	-0.002	0.027	0.087	0.116
			0.3	-0.003	-	-0.002	-0.239	-0.165	-0.135
		0.7	-0.3	-0.005	0.003	0.001	-0.121	0.043	0.142
			0	-0.007	0.003	-0.005	-0.077	-0.008	0.092
			0.3	-0.005	-	-0.007	-0.307	-0.226	-0.147
800	0.6:0.4	0.3	-0.3	0.001	0.002	0.001	-0.109	-0.086	-0.025
			0	-	0.002	0.001	-0.099	-0.040	0.008
			0.3	-	0.001	-	-0.138	-0.067	-0.018
		0.7	-0.3	-	0.001	0.001	-0.190	-0.163	-0.025
			0	-0.001	0.001	0.002	-0.221	-0.144	0.002
			0.3	-0.001	-	-	-0.244	-0.165	-0.027
	0.8:0.2	0.3	-0.3	-	0.001	0.002	-0.075	-0.079	-0.020
			0	-	0.002	0.002	-0.073	-0.063	0.010
			0.3	-	0.001	-	-0.115	-0.081	-0.014
		0.7	-0.3	0.001	0.002	0.002	-0.168	-0.180	-0.016
			0	-	0.003	0.001	-0.174	-0.162	0.017
			0.3	-	0.001	-	-0.225	-0.197	-0.015

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM. Bias values that approaches 0 (i.e., -0.001<Bias<0.001) are represented with "-".

Table A. 6

SE of the Initial Ability and Ability Change Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	SE of $\hat{\theta}_j^{(T_1)}$			SE of $\Delta\hat{\theta}_j$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.294	0.200	0.236	0.402	0.243	0.264
			0	0.235	0.250	0.248	0.313	0.288	0.281
			0.3	0.242	0.261	0.288	0.252	0.239	0.268
		0.7	-0.3	0.297	0.218	0.240	0.584	0.248	0.262
			0	0.237	0.226	0.230	0.393	0.280	0.249
			0.3	0.236	0.253	0.245	0.289	0.228	0.220
	0.8:0.2	0.3	-0.3	0.331	0.181	0.254	0.390	0.233	0.274
			0	0.272	0.270	0.272	0.353	0.311	0.306
			0.3	0.241	0.265	0.282	0.272	0.249	0.257
		0.7	-0.3	0.336	0.229	0.238	0.607	0.271	0.258
			0	0.259	0.246	0.253	0.458	0.313	0.274
			0.3	0.240	0.245	0.254	0.303	0.240	0.231
800	0.6:0.4	0.3	-0.3	0.202	0.147	0.145	0.165	0.114	0.118
			0	0.206	0.148	0.147	0.185	0.132	0.138
			0.3	0.197	0.162	0.163	0.176	0.133	0.141
		0.7	-0.3	0.218	0.166	0.148	0.281	0.104	0.122
			0	0.217	0.160	0.145	0.304	0.116	0.135
			0.3	0.204	0.166	0.157	0.258	0.119	0.129
	0.8:0.2	0.3	-0.3	0.207	0.142	0.136	0.153	0.115	0.115
			0	0.214	0.144	0.136	0.175	0.122	0.129
			0.3	0.191	0.162	0.153	0.169	0.132	0.135
		0.7	-0.3	0.220	0.164	0.139	0.255	0.111	0.124
			0	0.213	0.164	0.141	0.271	0.126	0.136
			0.3	0.203	0.172	0.155	0.260	0.128	0.133

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 7

RMSE of the Initial Ability and Ability Change Estimates

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	RMSE of $\hat{\theta}_j^{(T_1)}$			RMSE of $\Delta\hat{\theta}_j$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.684	0.684	0.692	0.842	0.843	0.850
			0	0.687	0.659	0.656	0.722	0.702	0.707
			0.3	0.656	0.631	0.648	0.740	0.716	0.724
		0.7	-0.3	0.714	0.685	0.682	1.076	0.830	0.836
			0	0.696	0.644	0.644	0.870	0.714	0.706
			0.3	0.662	0.611	0.630	0.836	0.724	0.716
	0.8:0.2	0.3	-0.3	0.694	0.681	0.711	0.853	0.832	0.858
			0	0.697	0.676	0.688	0.749	0.718	0.719
			0.3	0.652	0.627	0.644	0.779	0.717	0.721
		0.7	-0.3	0.703	0.683	0.698	1.055	0.831	0.851
			0	0.704	0.663	0.668	0.926	0.727	0.713
			0.3	0.657	0.608	0.646	0.854	0.729	0.726
800	0.6:0.4	0.3	-0.3	0.751	0.756	0.758	0.755	0.750	0.769
			0	0.696	0.712	0.712	0.732	0.727	0.741
			0.3	0.659	0.647	0.645	0.737	0.697	0.703
		0.7	-0.3	0.764	0.747	0.754	0.913	0.733	0.766
			0	0.715	0.705	0.706	0.931	0.705	0.733
			0.3	0.673	0.640	0.638	0.881	0.689	0.690
	0.8:0.2	0.3	-0.3	0.738	0.756	0.759	0.764	0.752	0.769
			0	0.684	0.711	0.714	0.736	0.716	0.738
			0.3	0.641	0.650	0.655	0.724	0.697	0.702
		0.7	-0.3	0.759	0.750	0.757	0.881	0.731	0.769
			0	0.695	0.703	0.708	0.875	0.707	0.738
			0.3	0.655	0.641	0.643	0.869	0.695	0.696

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 8

Bias of the Mean and Variance Estimates of Ability Change and Covariance Estimates between the Initial Ability and Ability Change

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Bias of $\hat{\mu}_{\Delta\theta}$			Bias of $\hat{\sigma}^2_{\Delta\theta}$			Bias of $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$		
				L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.027	0.083	0.102	0.970	0.561	0.526	0.435	0.613	0.621
			0	0.029	0.051	0.103	0.844	0.599	0.577	0.295	0.401	0.406
			0.3	-0.189	-0.168	-0.127	0.226	0.069	0.086	0.076	0.036	0.062
		0.7	-0.3	-0.120	0.055	0.110	2.436	0.530	0.515	0.392	0.571	0.602
			0	0.009	0.035	0.104	1.602	0.687	0.543	0.357	0.365	0.384
			0.3	-0.225	-0.180	-0.125	0.651	0.030	0.058	0.202	-0.051	0.029
	0.8:0.2	0.3	-0.3	0.068	0.109	0.116	0.639	0.544	0.546	0.510	0.578	0.597
			0	0.023	0.081	0.112	0.913	0.697	0.626	0.386	0.399	0.406
			0.3	-0.239	-0.166	-0.137	0.310	0.086	0.095	0.178	0.035	0.067
		0.7	-0.3	-0.126	0.038	0.136	1.869	0.537	0.585	0.508	0.520	0.600
			0	-0.080	-0.012	0.089	1.910	0.812	0.541	0.463	0.368	0.383
			0.3	-0.305	-0.227	-0.146	0.665	0.102	0.104	0.268	-0.029	0.066
800	0.6:0.4	0.3	-0.3	-0.110	-0.086	-0.026	-0.355	-0.461	-0.370	0.515	0.585	0.677
			0	-0.099	-0.041	0.007	-0.165	-0.293	-0.225	0.463	0.469	0.526
			0.3	-0.138	-0.068	-0.018	-0.187	-0.351	-0.299	0.294	0.208	0.253
		0.7	-0.3	-0.190	-0.163	-0.025	0.544	-0.582	-0.395	0.808	0.426	0.658
			0	-0.220	-0.145	0.001	0.967	-0.479	-0.250	0.772	0.284	0.507
			0.3	-0.243	-0.165	-0.027	0.578	-0.481	-0.343	0.619	0.077	0.221
	0.8:0.2	0.3	-0.3	-0.075	-0.080	-0.021	-0.386	-0.416	-0.341	0.608	0.608	0.689
			0	-0.073	-0.064	0.009	-0.270	-0.341	-0.218	0.484	0.421	0.523
			0.3	-0.115	-0.081	-0.015	-0.219	-0.336	-0.282	0.291	0.203	0.254
		0.7	-0.3	-0.168	-0.180	-0.017	0.259	-0.566	-0.349	0.855	0.408	0.678
			0	-0.174	-0.163	0.016	0.512	-0.487	-0.225	0.807	0.252	0.522
			0.3	-0.225	-0.197	-0.016	0.572	-0.460	-0.309	0.655	0.071	0.244

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 9

SE of the Mean and Variance Estimates of Ability Change and Covariance Estimates between the Initial Ability and Ability Change

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	SE of $\hat{\mu}_{\Delta\theta}$			SE of $\hat{\sigma}^2_{\Delta\theta}$			SE of $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$		
				L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.112	0.101	0.097	0.487	0.142	0.142	0.112	0.055	0.060
			0	0.085	0.097	0.099	0.229	0.153	0.147	0.081	0.080	0.089
			0.3	0.073	0.105	0.112	0.097	0.074	0.071	0.052	0.061	0.058
		0.7	-0.3	0.126	0.098	0.095	1.301	0.098	0.135	0.124	0.057	0.054
			0	0.108	0.109	0.092	0.493	0.222	0.151	0.092	0.072	0.079
			0.3	0.087	0.090	0.101	0.136	0.096	0.061	0.073	0.081	0.063
	0.8:0.2	0.3	-0.3	0.143	0.088	0.100	0.292	0.119	0.138	0.085	0.065	0.064
			0	0.101	0.100	0.109	0.304	0.200	0.199	0.092	0.098	0.101
			0.3	0.089	0.112	0.113	0.121	0.063	0.066	0.069	0.067	0.047
		0.7	-0.3	0.158	0.107	0.081	1.304	0.105	0.132	0.118	0.081	0.055
			0	0.113	0.106	0.108	0.845	0.271	0.203	0.087	0.080	0.094
			0.3	0.094	0.110	0.116	0.164	0.118	0.066	0.067	0.093	0.047
800	0.6:0.4	0.3	-0.3	0.029	0.032	0.030	0.052	0.058	0.057	0.046	0.054	0.048
			0	0.029	0.031	0.031	0.091	0.077	0.088	0.056	0.056	0.060
			0.3	0.032	0.035	0.036	0.063	0.086	0.097	0.047	0.058	0.063
		0.7	-0.3	0.035	0.029	0.030	0.183	0.031	0.071	0.083	0.047	0.055
			0	0.032	0.031	0.034	0.250	0.051	0.090	0.090	0.050	0.062
			0.3	0.031	0.034	0.035	0.173	0.061	0.088	0.082	0.057	0.058
	0.8:0.2	0.3	-0.3	0.033	0.029	0.026	0.046	0.056	0.051	0.044	0.049	0.042
			0	0.034	0.027	0.028	0.069	0.058	0.066	0.051	0.038	0.044
			0.3	0.029	0.030	0.033	0.074	0.074	0.089	0.049	0.050	0.056
		0.7	-0.3	0.035	0.030	0.029	0.148	0.035	0.069	0.086	0.053	0.056
			0	0.039	0.038	0.033	0.188	0.061	0.082	0.073	0.064	0.056
			0.3	0.033	0.036	0.034	0.161	0.065	0.089	0.068	0.063	0.056

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 10

RMSE of the Mean and Variance Estimates of Ability Change and Covariance Estimates between the Initial Ability and Ability Change

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B \mid M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	RMSE of $\hat{\mu}_{\Delta\theta}$			RMSE of $\hat{\sigma}^2_{\Delta\theta}$			RMSE of $\hat{\sigma}_{\theta^{(T_1)}\Delta\theta}$		
				L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM	L-LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.115	0.130	0.141	1.085	0.579	0.545	0.449	0.615	0.624
			0	0.090	0.109	0.143	0.874	0.619	0.595	0.306	0.409	0.416
			0.3	0.203	0.198	0.169	0.246	0.101	0.111	0.092	0.071	0.085
		0.7	-0.3	0.174	0.112	0.145	2.761	0.539	0.532	0.411	0.574	0.604
			0	0.108	0.115	0.138	1.676	0.722	0.563	0.368	0.372	0.392
			0.3	0.241	0.201	0.160	0.665	0.101	0.084	0.215	0.096	0.070
	0.8:0.2	0.3	-0.3	0.158	0.140	0.153	0.703	0.557	0.563	0.517	0.582	0.600
			0	0.103	0.129	0.156	0.962	0.725	0.657	0.396	0.411	0.418
			0.3	0.255	0.200	0.177	0.333	0.107	0.116	0.191	0.075	0.082
		0.7	-0.3	0.202	0.114	0.158	2.279	0.547	0.600	0.521	0.526	0.603
			0	0.139	0.107	0.140	2.089	0.856	0.578	0.471	0.377	0.394
			0.3	0.319	0.252	0.186	0.685	0.156	0.123	0.277	0.097	0.081
800	0.6:0.4	0.3	-0.3	0.113	0.092	0.039	0.358	0.465	0.374	0.517	0.587	0.679
			0	0.103	0.052	0.032	0.189	0.303	0.242	0.467	0.472	0.530
			0.3	0.142	0.076	0.040	0.197	0.361	0.315	0.298	0.216	0.260
		0.7	-0.3	0.193	0.165	0.039	0.574	0.583	0.402	0.812	0.428	0.660
			0	0.223	0.148	0.034	0.999	0.481	0.266	0.777	0.288	0.510
			0.3	0.245	0.169	0.044	0.604	0.485	0.354	0.625	0.096	0.229
	0.8:0.2	0.3	-0.3	0.082	0.085	0.033	0.388	0.420	0.344	0.610	0.610	0.690
			0	0.081	0.069	0.030	0.279	0.346	0.228	0.487	0.423	0.525
			0.3	0.119	0.087	0.036	0.231	0.344	0.296	0.295	0.209	0.260
		0.7	-0.3	0.172	0.182	0.034	0.298	0.567	0.355	0.859	0.412	0.680
			0	0.179	0.167	0.036	0.546	0.490	0.240	0.810	0.260	0.525
			0.3	0.228	0.201	0.037	0.594	0.465	0.322	0.658	0.095	0.250

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 11

Bias, SE and RMSE of the Estimates of the Initial Mixing Proportion of Strategy A

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Bias of $\hat{\pi}_{M_A}^{(T_1)}$		SE of $\hat{\pi}_{M_A}^{(T_1)}$		RMSE of $\hat{\pi}_{M_A}^{(T_1)}$	
				L-MCDM	LTA-L-MCDM	L-MCDM	LTA-L-MCDM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	-0.093	-0.032	0.029	0.036	0.097	0.048
			0	-0.082	-0.034	0.033	0.039	0.089	0.052
			0.3	-0.106	-0.039	0.041	0.035	0.114	0.052
		0.7	-0.3	-0.245	-0.041	0.041	0.038	0.249	0.056
			0	-0.219	-0.044	0.047	0.043	0.224	0.062
			0.3	-0.260	-0.056	0.057	0.040	0.266	0.068
	0.8:0.2	0.3	-0.3	-0.115	-0.046	0.030	0.030	0.119	0.055
			0	-0.110	-0.035	0.040	0.036	0.117	0.050
			0.3	-0.131	-0.053	0.033	0.035	0.135	0.064
		0.7	-0.3	-0.283	-0.055	0.050	0.033	0.288	0.064
			0	-0.248	-0.043	0.059	0.039	0.255	0.058
			0.3	-0.277	-0.065	0.055	0.035	0.282	0.074
800	0.6:0.4	0.3	-0.3	-0.101	-0.016	0.013	0.013	0.102	0.020
			0	-0.082	0.001	0.012	0.014	0.083	0.014
			0.3	-0.087	-0.002	0.012	0.014	0.088	0.014
		0.7	-0.3	-0.260	-0.019	0.017	0.012	0.261	0.022
			0	-0.238	-0.001	0.018	0.014	0.238	0.014
			0.3	-0.241	-0.004	0.020	0.013	0.242	0.014
	0.8:0.2	0.3	-0.3	-0.121	-0.012	0.013	0.011	0.121	0.016
			0	-0.107	0.001	0.014	0.011	0.108	0.011
			0.3	-0.103	0.001	0.014	0.010	0.104	0.010
		0.7	-0.3	-0.294	-0.012	0.018	0.011	0.295	0.017
			0	-0.272	-	0.017	0.011	0.273	0.011
			0.3	-0.277	-0.001	0.022	0.011	0.278	0.011

Note. L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM. Bias values that approaches 0 (i.e., -0.001<Bias<0.001) are represented with "-".

Table A. 12

Bias, SE and RMSE of the Estimates of the Latent Transition Probability from Strategy A to Strategy B of the LTA-longitudinal-MCDM

J	$\pi_{M_A}^{(1)} : \pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	$\hat{\tau}_{M_B\|M_A}^{(T_1)}$ of the LTA-longitudinal-MCDM		
				Bias	SE	RMSE
100	0.6:0.4	0.3	-0.3	-0.018	0.078	0.080
			0	0.021	0.058	0.062
			0.3	-0.049	0.049	0.069
		0.7	-0.3	0.003	0.086	0.086
			0	0.078	0.073	0.107
			0.3	-0.012	0.069	0.070
	0.8:0.2	0.3	-0.3	-0.058	0.055	0.080
			0	0.045	0.065	0.079
			0.3	0.017	0.060	0.062
		0.7	-0.3	-0.046	0.065	0.079
			0	0.031	0.058	0.066
			0.3	-0.028	0.054	0.060
800	0.6:0.4	0.3	-0.3	-0.027	0.025	0.037
			0	-0.020	0.030	0.036
			0.3	-0.009	0.030	0.031
		0.7	-0.3	-0.006	0.018	0.019
			0	0.009	0.022	0.024
			0.3	0.027	0.022	0.035
	0.8:0.2	0.3	-0.3	-0.010	0.014	0.017
			0	-	0.016	0.016
			0.3	0.012	0.020	0.023
		0.7	-0.3	-0.030	0.015	0.034
			0	-0.015	0.018	0.023
			0.3	0.014	0.016	0.021

Note. Bias values that approaches 0 (i.e., -0.001<Bias<0.001) are represented with "-".

Table A. 13

Bias of the Item Intercept and Attribute Main Effect Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Bias of $\hat{\lambda}_{i,0}$			Bias of $\hat{\lambda}_{i,1,(k)}$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.505	0.208	0.175	-0.659	-0.252	-0.201
			0	0.431	0.172	0.145	-0.562	-0.194	-0.146
			0.3	0.427	0.171	0.152	-0.570	-0.198	-0.134
		0.7	-0.3	0.590	0.260	0.172	-0.764	-0.340	-0.188
			0	0.498	0.205	0.132	-0.654	-0.273	-0.155
			0.3	0.491	0.222	0.139	-0.671	-0.300	-0.144
	0.8:0.2	0.3	-0.3	0.437	0.205	0.169	-0.539	-0.291	-0.226
			0	0.354	0.178	0.131	-0.457	-0.236	-0.162
			0.3	0.387	0.205	0.143	-0.487	-0.247	-0.159
		0.7	-0.3	0.540	0.279	0.169	-0.670	-0.383	-0.194
			0	0.444	0.234	0.125	-0.558	-0.303	-0.141
			0.3	0.451	0.245	0.138	-0.576	-0.308	-0.143
800	0.6:0.4	0.3	-0.3	0.368	0.054	0.022	-0.525	-0.090	-0.011
			0	0.345	0.044	0.012	-0.489	-0.078	-0.007
			0.3	0.339	0.053	0.013	-0.478	-0.077	-0.011
		0.7	-0.3	0.444	0.112	0.016	-0.622	-0.203	-0.009
			0	0.420	0.107	0.009	-0.581	-0.196	-0.006
			0.3	0.406	0.110	0.009	-0.565	-0.191	-0.015
	0.8:0.2	0.3	-0.3	0.281	0.062	0.019	-0.391	-0.126	-0.022
			0	0.248	0.049	0.011	-0.354	-0.104	-0.003
			0.3	0.243	0.061	0.013	-0.344	-0.103	-0.005
		0.7	-0.3	0.368	0.110	0.010	-0.513	-0.238	-0.010
			0	0.341	0.105	0.007	-0.473	-0.223	-0.005
			0.3	0.340	0.127	0.007	-0.466	-0.232	-0.013

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 14

SE of the Item Intercept and Attribute Main Effect Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	SE of $\hat{\lambda}_{i,0}$			SE of $\hat{\lambda}_{i,1,(k)}$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.318	0.346	0.344	0.385	0.471	0.480
			0	0.333	0.355	0.351	0.395	0.489	0.491
			0.3	0.342	0.345	0.346	0.406	0.482	0.499
		0.7	-0.3	0.320	0.352	0.344	0.395	0.474	0.472
			0	0.328	0.360	0.346	0.412	0.494	0.487
			0.3	0.333	0.337	0.339	0.415	0.480	0.496
	0.8:0.2	0.3	-0.3	0.342	0.365	0.367	0.403	0.519	0.528
			0	0.365	0.364	0.360	0.418	0.516	0.526
			0.3	0.354	0.352	0.354	0.413	0.507	0.526
		0.7	-0.3	0.334	0.365	0.356	0.401	0.494	0.489
			0	0.346	0.362	0.356	0.413	0.503	0.496
			0.3	0.349	0.352	0.346	0.415	0.495	0.505
800	0.6:0.4	0.3	-0.3	0.133	0.143	0.145	0.150	0.187	0.196
			0	0.139	0.147	0.149	0.156	0.194	0.204
			0.3	0.138	0.146	0.149	0.157	0.196	0.207
		0.7	-0.3	0.136	0.141	0.146	0.157	0.183	0.192
			0	0.136	0.143	0.149	0.158	0.184	0.196
			0.3	0.137	0.146	0.149	0.158	0.192	0.200
	0.8:0.2	0.3	-0.3	0.145	0.152	0.151	0.163	0.209	0.214
			0	0.149	0.157	0.158	0.166	0.225	0.232
			0.3	0.142	0.153	0.153	0.161	0.222	0.228
		0.7	-0.3	0.141	0.146	0.147	0.156	0.191	0.196
			0	0.143	0.150	0.155	0.159	0.200	0.207
			0.3	0.141	0.151	0.153	0.161	0.205	0.205

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 15

RMSE of the Item Intercept and Attribute Main Effect Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	RMSE of $\hat{\lambda}_{i,0}$			RMSE of $\hat{\lambda}_{i,1,(k)}$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.678	0.420	0.400	0.888	0.549	0.534
			0	0.627	0.405	0.388	0.810	0.541	0.524
			0.3	0.623	0.398	0.387	0.824	0.540	0.530
		0.7	-0.3	0.758	0.455	0.399	0.997	0.600	0.519
			0	0.684	0.426	0.378	0.919	0.587	0.520
			0.3	0.682	0.416	0.375	0.943	0.596	0.525
	0.8:0.2	0.3	-0.3	0.616	0.435	0.418	0.770	0.616	0.593
			0	0.572	0.425	0.393	0.717	0.589	0.565
			0.3	0.602	0.421	0.391	0.750	0.585	0.564
		0.7	-0.3	0.714	0.489	0.415	0.905	0.648	0.539
			0	0.646	0.458	0.388	0.824	0.617	0.524
			0.3	0.660	0.449	0.384	0.856	0.620	0.534
800	0.6:0.4	0.3	-0.3	0.432	0.156	0.152	0.637	0.217	0.200
			0	0.419	0.157	0.154	0.609	0.221	0.208
			0.3	0.419	0.158	0.152	0.601	0.220	0.210
		0.7	-0.3	0.512	0.186	0.150	0.762	0.295	0.195
			0	0.490	0.190	0.152	0.725	0.296	0.200
			0.3	0.482	0.191	0.152	0.707	0.294	0.204
	0.8:0.2	0.3	-0.3	0.356	0.173	0.155	0.499	0.257	0.218
			0	0.337	0.172	0.162	0.466	0.265	0.236
			0.3	0.335	0.174	0.157	0.457	0.265	0.232
		0.7	-0.3	0.438	0.205	0.149	0.625	0.343	0.199
			0	0.417	0.209	0.157	0.591	0.345	0.211
			0.3	0.422	0.221	0.156	0.592	0.347	0.208

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 16

Bias of the Attribute Easiness and Attribute Discrimination Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	Bias of $\hat{\beta}_k$			Bias of $\hat{\xi}_k$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	-0.234	0.125	0.083	0.286	0.138	0.148
			0	-0.288	-0.064	-0.088	0.553	0.225	0.220
			0.3	-0.184	0.016	-0.035	0.631	0.393	0.389
		0.7	-0.3	-0.266	0.167	0.088	0.448	0.133	0.116
			0	-0.327	0.027	-0.126	0.725	0.250	0.246
			0.3	-0.219	0.110	-0.081	0.806	0.404	0.350
	0.8:0.2	0.3	-0.3	-0.121	0.139	0.139	0.130	0.097	0.101
			0	-0.171	0.029	-0.048	0.354	0.184	0.208
			0.3	-0.074	0.039	-0.009	0.592	0.425	0.395
		0.7	-0.3	-0.154	0.197	0.152	0.237	0.109	0.097
			0	-0.199	0.122	-0.081	0.491	0.190	0.213
			0.3	-0.106	0.147	-0.052	0.688	0.436	0.367
800	0.6:0.4	0.3	-0.3	-0.133	0.203	0.151	0.699	0.699	0.650
			0	-0.162	0.153	0.110	0.496	0.646	0.610
			0.3	-0.093	0.074	0.032	0.645	0.459	0.429
		0.7	-0.3	-0.287	0.231	0.153	0.957	0.672	0.604
			0	-0.280	0.200	0.102	0.864	0.661	0.560
			0.3	-0.199	0.140	0.023	0.902	0.467	0.382
	0.8:0.2	0.3	-0.3	-0.087	0.236	0.170	0.293	0.656	0.643
			0	-0.093	0.178	0.117	0.289	0.627	0.639
			0.3	-0.045	0.116	0.050	0.408	0.475	0.484
		0.7	-0.3	-0.145	0.303	0.177	0.418	0.687	0.603
			0	-0.134	0.247	0.112	0.377	0.650	0.564
			0.3	-0.089	0.205	0.041	0.515	0.487	0.417

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 17

SE of the Attribute Easiness and Attribute Discrimination Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	SE of $\hat{\beta}_k$			SE of $\hat{\xi}_k$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.233	0.173	0.181	0.628	0.355	0.499
			0	0.247	0.215	0.227	0.381	0.495	0.514
			0.3	0.261	0.231	0.259	0.419	0.566	0.717
		0.7	-0.3	0.258	0.191	0.192	0.570	0.390	0.510
			0	0.246	0.186	0.200	0.311	0.352	0.399
			0.3	0.255	0.204	0.199	0.331	0.530	0.484
	0.8:0.2	0.3	-0.3	0.237	0.171	0.175	0.704	0.313	0.519
			0	0.246	0.235	0.248	0.482	0.556	0.564
			0.3	0.268	0.258	0.262	0.423	0.562	0.645
		0.7	-0.3	0.267	0.199	0.164	0.710	0.446	0.476
			0	0.245	0.217	0.208	0.322	0.434	0.496
			0.3	0.256	0.231	0.200	0.334	0.440	0.476
800	0.6:0.4	0.3	-0.3	0.101	0.066	0.066	0.195	0.150	0.161
			0	0.099	0.068	0.069	0.191	0.159	0.165
			0.3	0.102	0.070	0.069	0.193	0.215	0.229
		0.7	-0.3	0.124	0.085	0.073	0.221	0.267	0.204
			0	0.111	0.080	0.064	0.199	0.221	0.169
			0.3	0.115	0.073	0.064	0.223	0.197	0.194
	0.8:0.2	0.3	-0.3	0.104	0.073	0.071	0.207	0.137	0.140
			0	0.106	0.075	0.067	0.260	0.148	0.125
			0.3	0.088	0.077	0.070	0.140	0.204	0.193
		0.7	-0.3	0.116	0.084	0.063	0.213	0.245	0.164
			0	0.108	0.089	0.064	0.214	0.249	0.161
			0.3	0.101	0.083	0.065	0.192	0.189	0.196

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Table A. 18

RMSE of the Attribute Easiness and Attribute Discrimination Estimates

J	$\pi_{M_A}^{(1)}:\pi_{M_B}^{(1)}$	$p_{M_B\|M_A}$	$\rho_{\theta^{(T_1)}\Delta\theta}$	RMSE of $\hat{\beta}_k$			RMSE of $\hat{\xi}_k$		
				L- LLM	L-MCDM	LTA-L-MCDM	L- LLM	L-MCDM	LTA-L-MCDM
100	0.6:0.4	0.3	-0.3	0.392	0.229	0.242	0.707	1.062	1.077
			0	0.481	0.315	0.328	0.845	0.643	0.653
			0.3	0.417	0.293	0.323	0.976	0.809	0.912
		0.7	-0.3	0.463	0.258	0.236	0.953	1.030	1.033
			0	0.539	0.315	0.342	1.024	0.576	0.637
			0.3	0.483	0.270	0.272	1.162	0.731	0.813
	0.8:0.2	0.3	-0.3	0.283	0.252	0.252	0.846	1.057	1.146
			0	0.363	0.282	0.330	0.759	0.663	0.715
			0.3	0.324	0.325	0.335	1.023	0.806	0.892
		0.7	-0.3	0.339	0.283	0.240	0.823	1.014	1.121
			0	0.421	0.277	0.332	0.918	0.557	0.692
			0.3	0.371	0.299	0.248	1.074	0.694	0.873
800	0.6:0.4	0.3	-0.3	0.175	0.218	0.175	0.909	1.188	1.198
			0	0.210	0.187	0.154	0.628	1.101	1.094
			0.3	0.148	0.107	0.103	0.694	0.604	0.602
		0.7	-0.3	0.369	0.257	0.175	1.359	1.001	1.141
			0	0.375	0.256	0.142	1.222	0.973	1.035
			0.3	0.295	0.160	0.092	1.108	0.529	0.512
	0.8:0.2	0.3	-0.3	0.138	0.254	0.195	0.452	1.178	1.232
			0	0.146	0.208	0.162	0.415	1.059	1.159
			0.3	0.106	0.140	0.115	0.435	0.610	0.684
		0.7	-0.3	0.222	0.331	0.193	0.740	1.030	1.186
			0	0.205	0.288	0.145	0.605	0.903	1.049
			0.3	0.171	0.225	0.099	0.623	0.533	0.566

Note. L-LLM=Longitudinal LLM; L-MCDM=Longitudinal MCDM; LTA-L-MCDM=LTA-longitudinal MCDM.

Appendix B: Item Parameter and Higher-Order Structural Parameter Estimates in the Empirical Data Analysis

Table B. 1

Item Parameter Estimates of the LTA-longitudinal-MCDM in the Empirical Data Analysis and the Derived Conditional Item Correct Response Probability Given Successful Strategy Application and Skill Implementation Difficulty

Item	Item Intercept ($\lambda_{i,0}$)	Attribute main effect parameters ($\lambda_{i,1,(k)}$)				Conditional probability of correct response given the successful strategy application		Probability of individuals with $\theta = 0$ mastering all the required attributes of a strategy	
		RPR	MD	NSF	GG	Complex Strategy	Simple Strategy	Complex Strategy	Simple Strategy
1	-2.61	1.77	2.04			0.77	0.30	0.18	0.25
	(0.17)	(0.22)	(0.20)						
2	-0.49		2.25			0.85	0.85	0.74	0.74
	(0.10)		(0.14)						
3	-3.69	0.55	1.57	0.69	1.41	0.63	0.19	0.02	0.15
	(0.26)	(0.29)	(0.28)	(0.18)	(0.29)				
4	-2.12		1.97	1.21		0.74	0.29	0.15	0.20
	(0.13)		(0.18)	(0.16)					
5	-1.86		2.71	3.54		0.99	0.84	0.15	0.20
	(0.15)		(0.26)	(0.27)					
6	-0.89			5.20		0.99	0.99	0.20	0.20
	(0.09)			(0.67)					
7	-4.69	0.61		4.97	0.49	0.80	0.57	0.03	0.20
	(0.40)	(0.35)		(0.39)	(0.32)				
8	-1.18			1.09	1.16	0.75	0.48	0.12	0.20
	(0.09)			(0.14)	(0.17)				
9	-1.07		2.57			0.82	0.82	0.74	0.74
	(0.11)		(0.14)						

10	-2.19		4.57			0.92	0.92	0.74	0.74
	(0.19)		(0.22)						
11	-1.41		4.92			0.97	0.97	0.74	0.74
	(0.14)		(0.26)						
12	-2.17		3.83			0.84	0.84	0.74	0.74
	(0.17)		(0.19)						
13	-1.11	3.44	3.01			0.87	0.91	0.74	0.25
	(0.14)	(0.44)	(0.21)						
14	-3.22	3.31	3.00	1.91		0.85	0.52	0.15	0.25
	(0.25)	(0.31)	(0.31)	(0.26)					
15	-1.21				2.18	0.72	0.72	0.62	0.62
	(0.12)				(0.15)				
16	-0.84				2.01	0.76	0.76	0.62	0.62
	(0.11)				(0.15)				
17	-4.49	3.40		0.06	0.05	0.25	0.27	0.25	0.03
	(0.55)	(0.57)		(0.06)	(0.05)				
18	-1.06				5.00	0.98	0.98	0.62	0.62
	(0.13)				(0.58)				
19	-4.49				4.98	0.62	0.62	0.62	0.62
	(0.59)				(0.60)				
20	-2.12	0.87			2.08	0.70	0.49	0.15	0.62
	(0.17)	(0.20)			(0.20)				
21	-2.55	2.47	0.32		1.36	0.78	0.30	0.15	0.46
	(0.18)	(0.27)	(0.20)		(0.21)				

Note. A blank entry in the main effect parameters indicates that an attribute does not affect the correct item response probability in complex strategy or simple strategy as, based on the empirical Q-matrices, the attribute is not required to solve the item by either strategy. RPR=ratios and proportional relationships; MD=measurement and data; NSF=number system – fractions; GG=geometry – graphing.

Table B. 2

Higher-Order Structural Parameter Estimates of the LTA-longitudinal-MCDM in the Empirical Data Analysis

Attribute	Attribute easiness parameter (β_k)			Attribute discrimination parameter (ξ_k)		
	Estimate	SE	95% CI	Estimate	SE	95% CI
Ratio & proportion relations (RPR)	-1.12	0.35	[-1.90, -0.53]	2.39	0.43	[1.66, 3.42]
Measurement & data (MD)	1.03	0.22	[0.63, 1.53]	3.01	0.52	[2.22, 4.31]
Number system – fractions (NSF)	-1.39	0.19	[-1.77, -1.06]	1.94	0.27	[1.46, 2.49]
Geometry – graphing (GG)	0.48	0.15	[0.19, 0.78]	1.85	0.24	[1.41, 2.38]

Note. 95% CI=95% Bayesian Credible Interval.

References

Abele, S., & von Davier, M. (2019). CDMs in vocational education: Assessment and usage of diagnostic problem-solving strategies in car mechatronics. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages* (pp. 461–488). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_22

Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, *61*(5), 364–373.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Alder, A. G., Adam, J., & Arenberg, D. (1990). Individual-differences assessment of the relationship between change in and initial level of adult cognitive functioning. *Psychology and Aging*, *5*(4), 560–568. https://doi.org/10.1037//0882-7974.5.4.560

Alexander, P. A., Graham, S., & Harris, K. R. (1998). A perspective on strategy research: Progress and prospects. *Educational Psychology Review*, *10*(2), 129–154.

Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, 93–114.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*(1), 3–16.

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. Springer Science & Business Media.

Azevedo, C. L. N., Fox, J.-P., & Andrade, D. F. (2016). Bayesian longitudinal item response modeling with restricted covariance pattern structures. *Statistics and Computing*, *26*(1–2), 443–460. https://doi.org/10.1007/s11222-014-9518-5

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55.

Beheshti, B., Desmarais, M. C., & Naceur, R. (2012). Methods to find the number of latent skills. *Proceedings 5th International Conference on of Educational Data Mining*, 81–86. http://www.professeurs.polymtl.ca/michel.desmarais/Papers/EDM2012/nskills-edm2012.pdf

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (Vol. 2). University of Wisconsin Press.

Berkhof, J., Van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*(3), 337–354.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*. http://ci.nii.ac.jp/naid/10015088440/

Bottge, B. A. (2001). Reconceptualizing mathematics problem solving for low-achieving students. *Remedial and Special Education, 22*(2), 102–112.

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. D., Butler, M., & Cho, S.-J. (2014). Effects of blended instructional models on math performance. *Exceptional Children, 80*(4), 423–437.

Bottge, B. A., Rueda, E., Serlin, R. C., Hung, Y.-H., & Kwon, J. M. (2007). Shrinking achievement differences with anchored math problems: Challenges and possibilities. *The Journal of Special Education, 41*(1), 31–49.

Bottge, B. A., Toland, M. D., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., & Ma, X. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Exceptional Children, 81*(2), 158–175.

Box, G. E. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics, 25*(2), 290–302.

Bransford, J. D., & Stein, B. S. (1993). *The ideal problem solver: A guide for improving thinking, learning, and creativity* (2nd ed.). W.H. Freeman.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16*(3), 221.

Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis* (1st ed.). Chapman & Hall.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis, 1*(4), 651–673.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13*(2), 195–212.

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37*(8), 598–618.

Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*(4), 633.

Cho, S.-J., Cohen, A. S., & Bottge, B. (2013). Detecting intervention effects using a multilevel latent transition analysis with a mixture IRT model. *Psychometrika, 78*(3), 576–600.

Cho, S.-J., Cohen, A. S., Kim, S.-H., & Bottge, B. (2010). Latent transition analysis with a mixture item response theory measurement model. *Applied Psychological Measurement, 34*(7), 483–504.

Cohen, J. (1965). Some statistical issues in psychological research. *Handbook of Clinical Psychology*, 95–121.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research*, *27*(1), 131–157.

Congdon, P. (2003). *Applied Bayesian modelling*. Wiley.

Coughlin, J., & Montague, M. (2011). The effects of cognitive strategy instruction on the mathematical problem solving of adolescents with spina bifida. *The Journal of Special Education*, *45*(3), 171–183.

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

de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*(4), 595–624.

DeCarlo, L. T. (2012). Recognizing uncertainty in the Q-matrix via a Bayesian extension of the DINA model. *Applied Psychological Measurement*, *36*(6), 447–468. https://doi.org/10.1177/0146621612449069

DeIorio, M., & Robert, C. P. (2002). Discussion of Spiegelhalter et al. *Journal of the Royal Statistical Society, Series B*, *64*, 629–630.

Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based Q-matrices. *16th Conference on Artificial Intelligence in Education*, 441–450. http://www.professeurs.polymtl.ca/michel.desmarais/Papers/aied2013.pdf

Desmarais, M. C., Xu, P., & Beheshti, B. (2015). Combining techniques to refine item to skills Q-matrices with a partition tree. *8th Conference on Educational Data Data Mining*, 29–36. http://www.professeurs.polymtl.ca/michel.desmarais/Papers/EDM2015/desmarais-xu-beheshti.pdf

Doornik, J. A. (2009). *An Object-Oriented Matrix Programming Language Ox 6.*

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495–515.

Embretson, S. E. (1997). Structured ability models in tests designed from cognitive theory. *Objective Measurement: Theory into Practice*, *4*, 223–236.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–760.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gelman, Andrew., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (2nd ed.). CRC Press.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.

Griffin, J. E. (2010). Default priors for density estimation with mixture models. *Bayesian Analysis*, *5*(1), 45–64.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *29*(1), 83–100.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Log-linear panel, trend, and cohort analysis*. Sage Publications, Inc.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Sage.

Hancock, G. R., & Choi, J. (2006). A vernacular for linear latent growth models. *Structural Equation Modeling*, *13*(3), 352–377.

Hansen, M. (2013). *Hierarchical item response models for cognitive diagnosis*. https://escholarship.org/uc/item/6ps9d3fd.pdf

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210.

Hickey, D. T., Moore, A. L., & Pellegrino, J. W. (2001). The motivational and academic consequences of elementary mathematics environments: Do constructivist innovations and reforms make a difference? *American Educational Research Journal*, *38*(3), 611–652.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54.

Horsnell, G. (1953). The effect of unequal group variances on the F-test for the homogeneity of group means. *Biometrika*, *40*(1/2), 128–136.

Houts, C. R., & Cai, L. (2015). *flexMIRT®: Flexible multilevel multidimensional item analysis and test scoring (version 3.0)*. Vector Psychometric Group.

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement*, *54*(4), 440–480.

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*(6), 464–485.

IBM Corporation. (2011). *IBM SPSS Statistics 20*.

Im, S., & Corter, J. E. (2011). Statistical consequences of attribute misspecification in the rule space method. *Educational and Psychological Measurement*, *71*(4), 712–731. https://doi.org/10.1177/0013164410384855

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*(1), 32–60.

Jitendra, A., DiPipi, C. M., & Perron-Jones, N. (2002). An exploratory study of schema-based word-problem—Solving instruction for middle school students with learning disabilities: An emphasis on conceptual and procedural understanding. *The Journal of Special Education*, *36*(1), 23–38.

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

Kadengye, D. T., Ceulemans, E., & van den Noortgate, W. (2013). A generalized longitudinal mixture IRT model for measuring differential growth in learning environments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-013-0413-3

Kaya, Y., & Leite, W. L. (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.

Kelly, S., & Ye, F. (2017). Accounting for the relationship between initial status and growth in regression models. *The Journal of Experimental Education*, *85*(3), 353–375.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, *99*(3), 422.

Koehler, A. B., & Murphree, E. S. (1988). A comparison of the Akaike and Schwarz criteria for selecting model order. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *37*(2), 187–195.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *The Journal of Experimental Education*, *43*(1), 61–69.

Lao, H. (2016). *Estimation of diagnostic classification models without constraints: Issues with class label switching* [PhD Thesis]. University of Kansas.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton, Mifflin. http://www.gbv.de/dms/hbz/toc/ht000685628.pdf

Lee, S. Y. (2017). *Growth curve cognitive diagnosis models for longitudinal assessment* [PhD Thesis]. UC Berkeley.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*(3), 205–237. https://doi.org/10.1111/j.1745-3984.2004.tb01163.x

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin & H. Hotelling (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Stanford University Press.

Li, F., Cohen, A., Bottge, B., & Templin, J. L. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*(2), 181–204. https://doi.org/10.1177/0013164415588946

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3–21.

Liu, R., & Huggins-Manley, A. C. (2016). The specification of attribute structures and its effects on classification accuracy in diagnostic test design. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, & J. A. Douglas (Eds.), *Quantitative psychology research* (pp. 243–254). Springer.

Ma, W., & Guo, W. (2019). Cognitive diagnosis models for multiple strategies. *British Journal of Mathematical and Statistical Psychology*.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational and Behavioral Statistics*, *2*(2), 99–120.

Madison, M. J., & Bradshaw, L. (2018a). Evaluating Intervention Effects in a Diagnostic Classification Model Framework. *Journal of Educational Measurement*, *55*(1), 32–51.

Madison, M. J., & Bradshaw, L. P. (2018b). Assessing growth in a diagnostic classification model framework. *Psychometrika*, *83*(4), 963–990.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 439–476.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*(2), 204–209.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Wadsworth.

McCulloch, R. E., Polson, N. G., & Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, *99*(1), 173–193.

McGrory, C. A., & Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, *51*(11), 5352–5367.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. Marcel Dekker.

Mercer, C. D., & Mercer, A. R. (2001). *Teaching students with learning problems* (6th ed.). Merrill/Prentice Hall.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New Methods for the Analysis of Change*. American Psychological Association.

Miller, J. W., & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, *113*(521), 340–356.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*(4), 379–416.

Mislevy, R. J., & Huang, C.-W. (2007). Measurement models as narrative structures. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and Mixture Distribution Rasch Models: Extensions and Applications* (pp. 15–35). Springer. https://doi.org/10.1007/978-0-387-49839-3_2

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*(2), 195–215.

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(3), 193–211.

Muthén, L. K., & Muthén, B. O. (2007). Mplus. *Statistical Analysis with Latent Variables. Version, 3*. http://www.statmodel.com/virg_nov_course.shtml

Nasserinejad, K., van Rosmalen, J., de Kort, W., & Lesaffre, E. (2017). Comparison of criteria for choosing the number of classes in Bayesian finite mixture models. *PloS One*, *12*(1), e0168838. https://doi.org/10.1371/journal.pone.0168838

Pan, Q. (2018). *Growth modeling in a diagnostic classification model (DCM) framework* [PhD Thesis]. University of Kansas.

Paris, S. G., Cross, D. R., & Lipson, M. Y. (1984). Informed strategies for learning: A program to improve children's reading awareness and comprehension. *Journal of Educational Psychology*, *76*(6), 1239.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, *8*(3), 293–316.

Paris, S. G., Wasik, B. A., & Turner, J. C. (1991). The development of strategic readers. In R. Barr, M. L. Kamil, P. B. Monsenthal, & P. D. Pearson (Eds.), *Handbook of Reading Research: Vol. II* (pp. 609–640). Longman.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika*, 114–133.

Plummer, M. (2015). *JAGS Version 4.0. 0 user manual*. https://sourceforge. net/projects/mcmc-jags/files/Manuals/4. x

Pressley, M. (2000). What should comprehension instruction be the instruction of? In M. L. Kamil, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 545–561). Erlbaum.

R Development Core Team. (2013). *R: A language and environment for statistical computing*. http://cran.fiocruz.br/web/packages/dplR/vignettes/timeseries-dplR.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Educational Research Institute.

Rijkes, C. P., & Kelderman, H. (2007). Latent-response Rasch models for strategy shifts in problem-solving processes. In *Multivariate and mixture distribution Rasch models* (pp. 311–328). Springer.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282.

Rousseau, J., & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(5), 689–710.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, *6*(4), 377–401.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 1151–1172.

Rubin, D. B. (1996). Comment: On posterior predictive p-values. *Statistica Sinica*, 787–792.

Rupp, A. A., & Templin, J. L. (2008a). The effect of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78–96. https://doi.org/10.1177/0013164407301545

Rupp, A. A., & Templin, J. L. (2008b). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist*, *6*(2), 461–464.

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied Psychological Measurement*.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626–633.

Siegler, R. S., Strauss, S., & Levin, I. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, 1–84.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–407.

Steele, R. J., & Raftery, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. *Frontiers of Statistical Decision Making and Bayesian Analysis, 2*, 113–130.

Su, Y.-S., & Yajima, M. (2015). R2jags: Using R to run 'JAGS'. R package version 0.5-7. *Available: CRAN. R-Project. Org/Package= R2jags.(September 2015)*.

Swanson, H. L. (2001). Research on interventions for adolescents with learning disabilities: A meta-analysis of outcomes related to higher-order processing. *The Elementary School Journal, 101*(3), 331–348.

Tatsuoka, K. K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55–73.

Tatsuoka, K. K. (1987). Validation of cognitive sensitivity for item response curves. *Journal of Educational Measurement, 24*(3), 233–245.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic Monitoring of Skill and Knowledge Acquisition*, 453–488.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287.

Templin, J. L., & Hoffman, L. (2013). Obtaining diagnostic classification model estimates using Mplus. *Educational Measurement: Issues and Practice, 32*(2), 37–50.

Thorndike, R. L. (1966). Intellectual status and intellectual growth. *Journal of Educational Psychology, 57*(3), 121.

Tiku, M. L. (1964). Approximating the general non-normal variance-ratio sampling distributions. *Biometrika, 51*(1–2), 83–95.

Toprak, T. E., Aryadoust, V., & Goh, C. (2019). The log-linear cognitive diagnosis modeling (LCDM) in second language listening assessment. *Quantitative Data Analysis for Language Assessment Volume II: Advanced Methods*, 56.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (RR-05-16; ETS Research Series). Educational Testing Service.

von Davier, M. (2007). Mixture distribution diagnostic models. *ETS Research Report Series*, *2007*(2), i–21. https://doi.org/10.1002/j.2333-8504.2007.tb02074.x

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

von Davier, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Research Report Series*, *2014*(2), 1–13.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*(2), 318–336.

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In *Multivariate and mixture distribution Rasch models* (pp. 99–115). Springer.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, *43*(3), 476–496.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics*, *43*(1), 57–87.

Wang, W.-C. (2014). Multidimensional Rasch models: Theories and applications. In *Advancing methodologies to support both summative and formative assessments* (pp. 215–241). Information Age Publishing Inc.

Wang, W.-C., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with multidimensional Rasch models. *Journal of Outcome Measurement*, *2*(3), 240–265.

Wang, W.-C., & Wu, C.-I. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement*, *64*(5), 758–780.

Willett, J. B. (1997). Measuring change: What individual growth modeling buys you. *Change and Development: Issues of Theory, Method, and Application*, *213*, 243.

Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, *20*(1/2), 32–52.

Xu, P., & Desmarais, M. (2016). Boosted decision tree for Q-matrix refinement. *The 9th International Conference on Educational Data Mining*, 551–555. http://www.professeurs.polymtl.ca/michel.desmarais/Publications-Michel/EDM_2016_paper_134.pdf

Yamamoto, K. (1987). *A model that combines IRT and latent class models* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models.* (RR-89-41; ETS Research Report). Educational Testing Service.

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics*, 1076998619827593. https://doi.org/10.3102/1076998619827593

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998619826040

Zheng, Y., Chiu, C. Y., & Douglas, J. A. (2014). NPCD: Nonparametric methods for cognitive diagnosis. *R Package Version 1.0*, *5*.