The Impact of Information Quantity and Quality on Parameter Estimation for a Selection of Dynamic Bayesian Network Models with Latent Variables

by

Raymond E. Reichenberg

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2018 by the
Graduate Supervisory Committee:

Roy Levy, Co-chair
Natalie Eggum-Wilkens, Co-chair
Dawn DeLay
Masumi Iida

ARIZONA STATE UNIVERSITY

August 2018

ABSTRACT

Dynamic Bayesian networks (DBNs; Reye, 2004) are a promising tool for modeling student proficiency under rich measurement scenarios (Reichenberg, in press). These scenarios often present assessment conditions far more complex than what is seen with more traditional assessments and require assessment arguments and psychometric models capable of integrating those complexities. Unfortunately, DBNs remain understudied and their psychometric properties relatively unknown. If the apparent strengths of DBNs are to be leveraged, then the body of literature surrounding their properties and use needs to be expanded upon. To this end, the current work aimed at exploring the properties of DBNs under a variety of realistic psychometric conditions. A two-phase Monte Carlo simulation study was conducted in order to evaluate parameter recovery for DBNs using maximum likelihood estimation with the Netica software package. Phase 1 included a limited number of conditions and was exploratory in nature while Phase 2 included a larger and more targeted complement of conditions. Manipulated factors included sample size, measurement quality, test length, the number of measurement occasions. Results suggested that measurement quality has the most prominent impact on estimation quality with more distinct performance categories yielding better estimation. While increasing sample size tended to improve estimation, there were a limited number of conditions under which greater samples size led to more estimation bias. An exploration of this phenomenon is included. From a practical perspective, parameter recovery appeared to be sufficient with samples as low as $N = 400$ as long as measurement quality was not poor and at least three items were present at each measurement occasion. Tests consisting of only a single item required exceptional measurement quality in order to adequately recover model parameters. The study was

somewhat limited due to potentially software-specific issues as well as a non-comprehensive collection of experimental conditions. Further research should replicate and, potentially expand the current work using other software packages including exploring alternate estimation methods (e.g., Markov chain Monte Carlo).

DEDICATION

This work, much like my love, is dedicated to my wife, Erynn, without whose support this project would not have reached completion. Further dedication is owed to my children -- Addy, Emil, and Olly, who serve as a constant source of motivation.

ACKNOWLEDGMENTS

I would, first and foremost like to acknowledge Dr. Roy Levy for guiding me through this journey and encouraging me to be a more rigorous, productive, and decisive researcher. I would also like to thank all those that served on my committee at one point of another – Dr. Natalie Eggum-Wilkens, Dr. Masumi Iida, Dr. Dawn DeLay, Dr. Marilyn Thompson, and Dr. Samuel Green. The contributions of these individuals were vital in ensuring the scholarly quality of this project. Finally, I would like to extend gratitude to Dr. Samuel Green, specifically, for his contributions not only to my own work, but also to the field at large. His dedication to the academy and the integrity which he maintained in both in his personal and professional life continue to serve as an inspiration.

TABLE OF CONTENTS

Page

CHAPTER Page

CHAPTER Page

APPENDIX Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Advances in technological capacity and accessibility in recent years have opened up possibilities for authentic assessments couched within rich environments which may yield proficiency score estimates that are more predictive of an examinee's real-world performance capabilities than what are produced by more traditional assessments such as paper/pencil exams (Eseryel, Ge, Ifenthaler, & Law, 2011; Shute, Leighton, Jang, & Chu, 2016; Zapata-Rivera & Bauer, 2012). These assessments often take the form of games (e.g., Chung et al., 2010; Iseli, Koenig, Lee, & Wainess, 2010; Rowe & Lester, 2010; Shute, 2011), simulations (e.g., Almond, Mulder, Hemat, & Yan, 2009; Williamson, Mislevy, & Bejar, 2006), or intelligent tutoring systems (ITSs; e.g., Mislevy & Gitomer, 1995; Reye, 2004; Sao Pedro, de Baker, Gobert, Montalvo, & Nakama, 2013; VanLehn, 2008) and, in many cases can be easily embedded within the course of classroom activity.

Dynamic Bayesian networks (DBNs; Reye, 2004) are a promising tool for modeling student proficiency under these rich measurement scenarios (Reichenberg, in press). These scenarios often present assessment conditions far more complex than what is seen with more traditional assessments and require assessment arguments and psychometric models capable of integrating those complexities. An assessment embedded within a level of an educational game, for example, might include a single item or task that is repeated many times depending on whether the player responds correctly (i.e., passes the level). There is also an element of feedback inherent in such a scenario: the player knows with each attempt whether or not they completed the task successfully. In many cases, these applications might also require on-the-fly updating of student

proficiency estimates to facilitate task selection, creation, or augmentation. These conditions are not necessarily conducive to modeling the player's proficiency using more widely adopted assessment modeling frameworks such as item response theory (IRT). IRT models, for example, often assume latent constructs consisting of multiple, locally independent tasks each administered a single time with no feedback being given to the respondent. Furthermore, these games might pursue the goal of modeling the player's growth or change in proficiency, a goal not often represented in most applications of IRT. IRT models have been proposed for modeling growth (see Andersen, 1985; Culpepper, 2014; Embretson, 1991; and von Davier, Xu, & Carstensen, 2011 as examples). However, it is not clear that these models, as currently developed, would be suitable for certain complex assessment situations that are of interest here, characterized by feedback to students that case doubt on the conditional independence assumptions.

DBNs offer a number of strengths in the context of the complex assessment scenarios detailed previously. Almond et al. (Almond, Mislevy, Steinberg, Yan, & Williamson, 2015; pp. 14-16) presented reasons for considering the use of Bayesian networks (BNs), all of which apply to DBNs. Of particular importance in light of the demands presented by complex assessment scenarios are DBN's ability to handle very large and/or complex models while remaining computationally efficient (i.e., they are fast). This efficiency means not only that a researcher may save time in estimating model parameters, but also that the model can be queried in real-time for diagnostic updates (Almond et al., 2015). This real-time updating makes these models well-suited for use in computer adaptive testing (CAT; Almond & Mislevy, 1999), game/simulation-based assessment, diagnostic assessment, and, potentially in classroom-based formative

assessment scenarios where teachers need to make decisions on-the-fly (Almond, Shute, Underwood, & Zapata-Rivera, 2009). Though more prevalent methods such as IRT also have utility in some of these areas (CAT, for example), the computational efficiency advantages DBNs provide often set them apart and, in some cases, such as when dealing with very large and/or complex systems of variables, may position DBNs as the only feasible option. Finally, scores resulting from DBNs might be more easily understood by consumers (e.g., researchers, assessment designers, teachers, parents, students) given that latent proficiency variables in these models are often assumed to be categorical, a choice which simplifies the interpretation and representation of those proficiencies (Almond et al., 2009; Almond et al., 2015).

Research Questions

Unfortunately, the body of literature related to the use of DBNs in educational measurement and related fields such as psychological measurement is rather sparse (Reichenberg, in press). If DBNs are to gain wider use, thus leveraging their apparent strengths under conditions such as those previously mentioned, then the knowledge base surrounding their use must be made more robust and understanding of their structure, function, strengths, and potential utility among both researchers and practitioners must be increased. In this early stage of adoption, methodological investigations aimed at better understanding the psychometric properties of DBNs and providing practitioners with guidelines for use should be a primary focus of the additions to the literature. A central question for the use of any psychometric modelling approach is the quantity (e.g., sample size or test length) and quality (e.g., measurement quality) of information (e.g., examinee responses) needed to reliably calibrate the model. As will be discussed in a later chapter, there has, to date been very little research conducted to answer that question for Bayesian

networks (BNs) and DBNs. The purpose of the current study, then is to further our understanding of the impact of information quantity and quality on parameter estimation in dynamic Bayesian networks using Monte Carlo simulation methods. Specifically, the study aims to address the following questions:

1. What impact do information quantity and quality have on the quality of parameter estimation and classification accuracy for a variety of dynamic Bayesian network structures containing latent variables?
2. How are those effects moderated by factors such as the magnitude of the true values for other, incidental parameters such as the transition probability and initial probability of mastery and the number of time slices (i.e., measurement occasions) present in the model?

Specific hypotheses and operational definitions of relevant terms (e.g., information quantity/quality, classification accuracy) are presented in Chapter 3 while concepts such as transition probability, initial mastery probability, and time slice are covered in the next section.

Exploration of these topics will require a brief overview of BNs and DBNs focusing on aspects relevant to psychometrics and will include an illustrative example using real data. The current study will focus on DBNs which are specified using categorical (dichotomous, specifically) observed and latent variables and will not consider models using continuous variables (such as Kalman filters), though approaches for using DBNs with continuous variables exist and may retain advantages relative to other modeling frameworks under such specifications (see Gharamani & Hinton, 2000; Johns & Woolf [2006] applied a similar model in an educational research context). The

assumption of a fixed structure will also be inherent in subsequent discussions. Though the literature on machine learning and data mining offer methods for learning the structure of a model given some data, most psychometric applications are predicated on the notion of an assessment designed with some target structure in mind.

Overview of (Dynamic) Bayesian Networks

Any discussion of DBNs requires at least a basic understanding of Bayesian networks as the former represents an extension of the latter. As BNs are not the central focus of this paper, I will provide only such a description as is necessary for facilitating understanding of DBNs.

Bayesian networks. A BN is multivariate distribution of discrete variables, commonly depicted as an acyclic directed graph (aka directed acyclic graph, or DAG) to express the dependence and conditional independence assumptions in the model for the joint distribution (Jensen, 1996). More concretely, a BN models the probability of an event or state such as a latent proficiency conditioned on a set of observed states, events, or characteristics such as item responses. A BN consists of a set of variables (often represented as "nodes" in the graph) and a set of "edges" between the variables. These edges are directed (i.e., single-headed arrows) and define the structure of the network. Under the representations considered here, each variable included in the model may take on a finite set of mutually exclusive states (i.e., they are categorical). More comprehensive overviews of BNs can be found in Nielsen and Jensen (2009), Neapolitan (2004), and Pearl (1988). See Almond et al. (2015), Culbertson (2015), and González-Brenes, Mislevy, Behrens, Levy, & DiCerbo (2016) for didactic treatments and reviews of BNs in educational assessment.

Bayesian networks have garnered increased attention in educational assessment in recent years due to their flexibility and computational advantages under certain circumstances (Almond, et al., 2015). Figure 1, adapted from Almond et al. (2015), depicts an example Bayesian network for a five-item assessment as represented in the Netica software package (Norsys Software Corp., 1995-2017). The joint probability distribution for this model can be written as

$$P(\theta, X_1, X_2, X_3, X_4, X_5). \tag{1}$$

Using the general product rule (aka the chain rule), Equation 1 can be re-written as

$$\begin{aligned} P(\theta, X_1, X_2, X_3, X_4, X_5) = P(\theta)P(X_1|\theta)P(X_2|\theta, X_1)P(X_3|\theta, X_1, X_2) \\ P(X_4|\theta, X_1, X_2, X_3)P(X_5|\theta, X_1, X_2, X_3, X_4). \end{aligned} \tag{2}$$

Under that assumption that the observed variables (X_1-X_5) are independent given the latent proficiency variable (θ), the distribution simplifies to

$$P(\theta, X_1, X_2, X_3, X_4, X_5) = P(\theta)P(X_1|\theta)P(X_2|\theta)P(X_3|\theta)P(X_4|\theta)P(X_5|\theta). \tag{3}$$

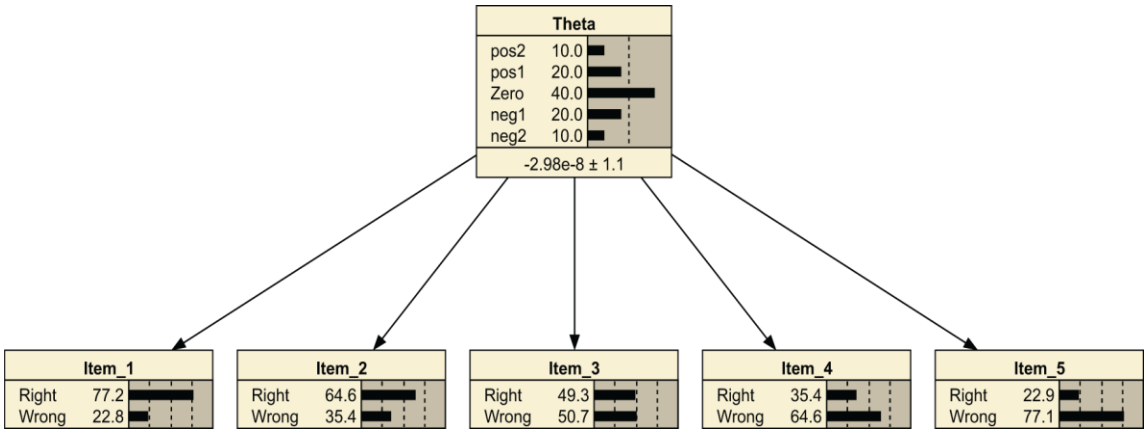

Figure 1. Sample Bayesian network representation of a five-item measurement model.

The relationships suggested by the right-hand side of Equation 3 reflect the structure depicted in Figure 1, which depicts the network in its initial state where no item responses have been observed. The probabilities shown for both the items as well as the latent proficiency represent prior beliefs in the form of population-level expectations. Note that the latent proficiency, "Theta" is defined in this example by five intervals or categories rather than by a continuum (the latter being the convention in IRT, for example). These five performance categories correspond to mean ability ("zero"), one/two standard deviations above the mean ("pos1"/"pos2"), and one/two standard deviations below the mean ("neg1"/"neg2"). These category designations are intended to be qualitative as opposed to strictly quantitative in that they represent a category that is only roughly aligned with the label. The directed arrows from the latent node to the observable item nodes represent conditional probability tables (CPTs) of a dimension determined by the number of categories defined for the respective latent/observable variables (5 x 2, in this case). Figure 2 shows the same network after the responses to items two and three have been observed for an individual. Notice the propagation of this

evidence through the network. The beliefs or posterior probabilities regarding the respondent's proficiency category membership have been updated as has the probability of correctly endorsing each the three remaining items. In this case, the respondent endorsed the correct response to both items two and three and, as would be expected, probability mass has been shifted from the lower proficiency categories to the higher proficiency categories and the probability of a correct response to any one of the remaining tasks has increased.

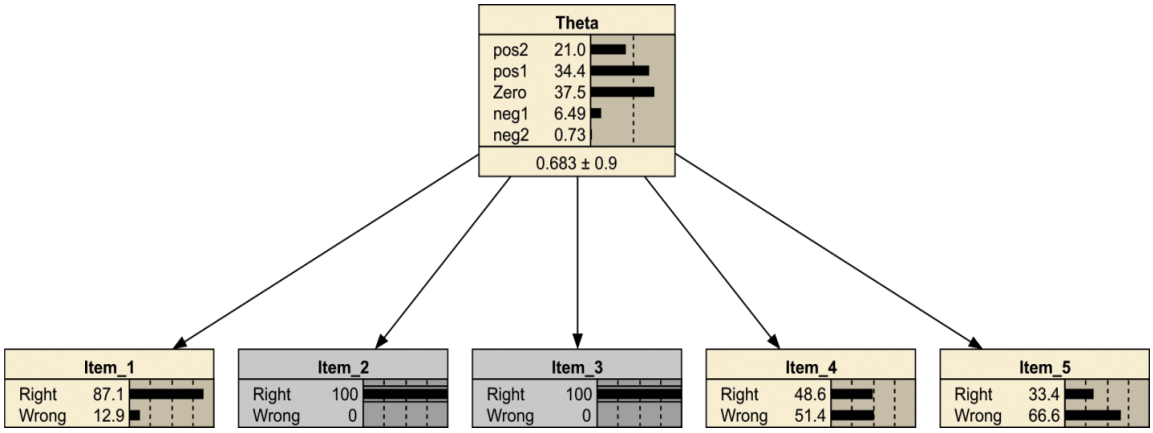

Figure 2. Sample Bayesian network representation of a five-item measurement model after two item responses have been observed.

Dynamic Bayesian networks. DBNs represent a longitudinal extension of BNs. In the simplest case, a DBN is a series of cross-sectional, time-specific BNs connected by a spine linking the latent proficiencies at each time point. Figure 3 represents this notion graphically. Following the method used previously for Equations 1-3, the joint probability distribution for this simple example can be expressed as

$$P(\theta_{t1}, \theta_{t2}, X_{t1}, X_{t2}) = P(\theta_{t1})P(\theta_{t2}|\theta_{t1})P(X_{t1}|\theta_{t1})P(X_{t2}|\theta_{t2}). \quad (4)$$

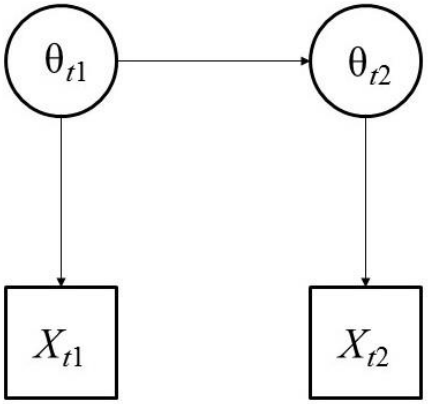

Figure 3. Graphical representation of a simple dynamic Bayesian network.

Note that this graph (as well as Equation 4) suggests three parameters that need to be either estimated or specified: (a) the prior state of the proficiency node at time *t*1 ($P(\theta_{t1})$), (b) the conditional probability distribution (CPD) of the proficiency node at time *t*1 given the evidence at time *t*1 ($P(X_{t1}|\theta_{t1})$), which is referred to as the *observation model*, and (c) the CPD for the proficiency node at time *t*2 given the proficiency node at time *t*1 ($P(\theta_{t2}|\theta_{t1})$), which is referred to as the *transition model*. When using DBNs, it is almost always the case that the observation model is assumed to be static across time slices (i.e., $P(X_{t1}|\theta_{t1}) = P(X_{t2}|\theta_{t2})$). The state of the proficiency node at time *t*2 is dependent on the state of the proficiency at time *t*1 and has encoded in it all the evidence that has been observed up to that point as well as the initial, or prior beliefs about the latent proficiency that were present before any observations were made. More generally, the CPD for the latent proficiency at any time point *t*2 depends only on the state of the proficiency at time

*t*1 and does not depend on the sequence of proficiency states that preceded time *t*1. That is to say that the value of the latent proficiency variable at time *t*2 is conditionally independent from the values of that same variable at any point preceding time *t*1. As with most psychometric models, applications of BNs almost always carry with them an exchangeability assumption (thus the lack of any subscripts referring to individual cases in Equation 1) in that we assume, *a priori* that each examinee is no different from any other examinee. That is to say that the model structure is assumed to hold for each respondent.

In many classic applications, the state of the proficiency variable can take on one of two values, usually interpreted as defining higher and lower proficiency groups. Similar to how these latent variables have been conceived of in the diagnostic classification literature (Rupp, Templin, & Henson, 2010), the levels are sometimes ascribed names such as "master" or "non-master" conveying interpretations with respect to the mastery of a skill the latent variable is intended to represent. Using these designations, the probability of a correct response given non-mastery is sometimes referred to as the "guess" parameter while the probability of an incorrect response given mastery is referred to as the "slip" parameter. This binary classification yields a 2x2 conditional probability table (CPT) for the probability of transitioning from one state to the next between two adjacent time points. It is the values in this table, or transition matrix, that defines the transition model discussed previously. Table 1 presents an example transition matrix for two adjacent time points. The values in the matrix represent conditional probabilities, such that each row gives the conditional probability distribution for mastery or non-mastery at time *t*2 given mastery or non-mastery at time *t*1. Note that

in Table 1, $P(\theta_{t+1} = NM \mid \theta_t = M) = 0$ and that $P(\theta_{t+1} = M \mid \theta_t = M) = 1$. This embodies a "once a master, always a master" assumption, where a respondent never regresses to non-mastery once mastery is achieved.

Table 1.

Sample transition matrix for a dynamic Bayesian network.

	$\theta_{t2} = NM$	$\theta_{t2} = M$
$\theta_{t1} = NM$	0.7	0.3
$\theta_{t1} = M$	0	1

Note. NM = non-master; M = master.

If we further assume an observed variable (e.g., a performance task presented to an examinee) consisting of two outcomes (i.e., "correct" and "incorrect"), then we end up with a 2x2 CPT for the probability of the examinee demonstrating either of the outcomes on the performance task conditioned on their membership in either of the categories on the proficiency variable. These probabilities define the observation model. Table 2 presents an example CPT for a situation such as that described above. In this example, the probability of a student who has mastered the content correctly endorsing the item is 0.70 or 70% while the probability of a non-master correctly endorsing the item is only 0.20 or 20%.

Table 2.

Sample conditional probability table (CPT) for a dynamic Bayesian network.

	$\theta_t = NM$	$\theta = M$
$P(X_{t1} = 1 \mid \theta_{t1})$	0.20	0.70
$P(X_{t1} = 0 \mid \theta_{t1})$	0.80	0.30

Note. NM = non-master; M = master.

An illustrative example. Figure 4 presents a simplified version of a DBN resulting from the *Save Patch* game (Chung et al., 2010; the scoring model by Levy [2014] is described in a later section) in which players are asked to place sections of rope in order to navigate an avatar (named *Patch*) across an expanse. The game assesses a variety of mathematical skills across its various levels. The example presented here is adapted from an early version of a level from the game which was designed to assess understanding of whole numbers. Response data from $N = 852$ students over a maximum of 10 attempts was used to calibrate the model. Only the first three time slices are presented in Figure 4 (and subsequent figures) for the sake of simplicity. Student responses were categorized as the expected successful solution (*StandardSolution*), an unexpected successful solution (*AlternateSolution*), a complete attempt that did not successfully navigate *Patch* across the expanse (*Error*), or an incomplete attempt (*IncompleteSolution*). In terms of proficiency estimates, students were characterized as having either mastered (*Master*) or not mastered (*Nonmaster)* the assessed skill. Tables 3 and 4 present the estimated CPTs for the measurement model and transition model, respectively for the calibrated network. Note that "once a master, always a master" assumption is not encoded in this model due software limitations.

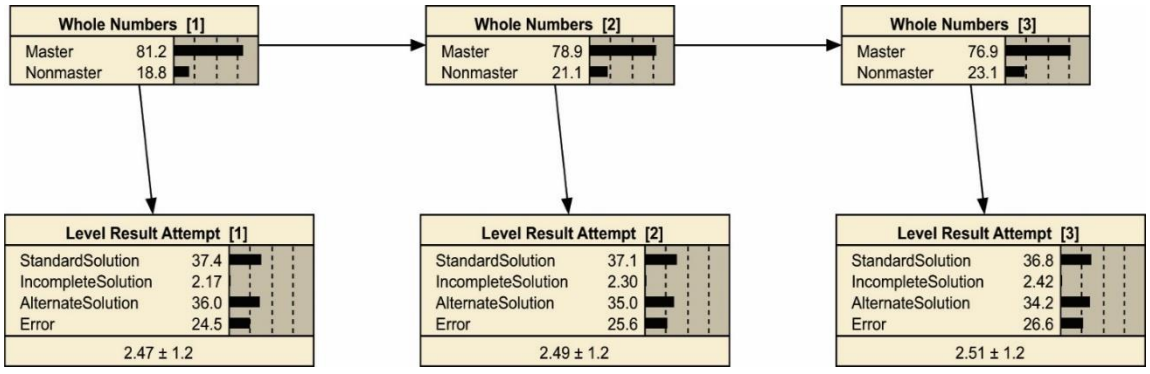

Figure 4. Netica representation of the first three time points for a calibrated DBN.

Table 3.

Measurement model CPT for the illustrative example.

	$\theta_t = NM$	$\theta_t = M$
$P(X_t = SS \mid \theta_t)$	0.263	0.400
$P(X_t = IS \mid \theta_t)$	0.069	0.011
$P(X_t = AS \mid \theta_t)$	0.021	0.438
$P(X_t = E \mid \theta_t)$	0.647	0.151

Note. SS = Standard Solution; IS = Incomplete Solution; AS = Alternate Solution; E = Error.

Table 4.

Transition model CPT for the illustrative example.

	$\theta_{t2} = NM$	$\theta_{t2} = M$
$\theta_{t1} = NM$	0.919	0.081
$\theta_{t1} = M$	0.047	0.953

Once the parameter estimation step is complete, the model can be employed for conducting inference at the level of the individual examinee. Figures 5, 6, and 7 present the example model after task performance at Time 1, Time 2, and Time 3, respectively have been observed for an examinee. The incorrect (*Error*) response at Time 1 dramatically reduces the probability of mastery in our estimate the student's proficiency at the first time point as well as our estimates of the probability that they will have mastered the content following their attempts at two future time points. Correspondingly, our beliefs about the student's likelihood of providing a correct response (*StandardSolution* or *AlternateSolution*) at either of the two future time points also trend towards skepticism. By the time we reach the conclusion of the observed task performance at Time 3, however, our beliefs about the student's probability of having

mastered the content have improved significantly due to the two correct responses observed at Time 2 and 3. The solution provided by the student at Time 3, in particular (*AlternateSolution*) has a dramatic effect on the posterior distribution for the latent variable due to the disparity in the probability of a master ($P(X_t = AS \mid \theta_t = M) = 0.438$) versus a non-master ($P(X_t = AS \mid \theta_t = NM) = 0.021$) providing such a response.

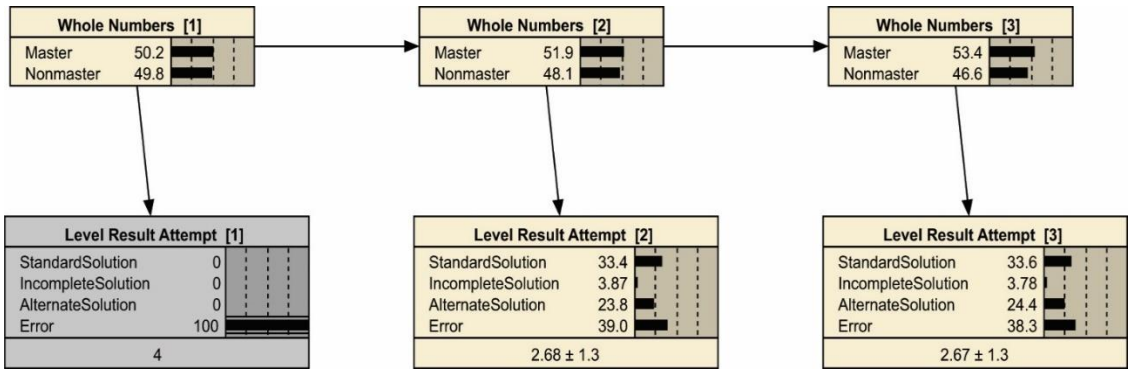

Figure 5. Netica representation of a DBN with one observed item response.

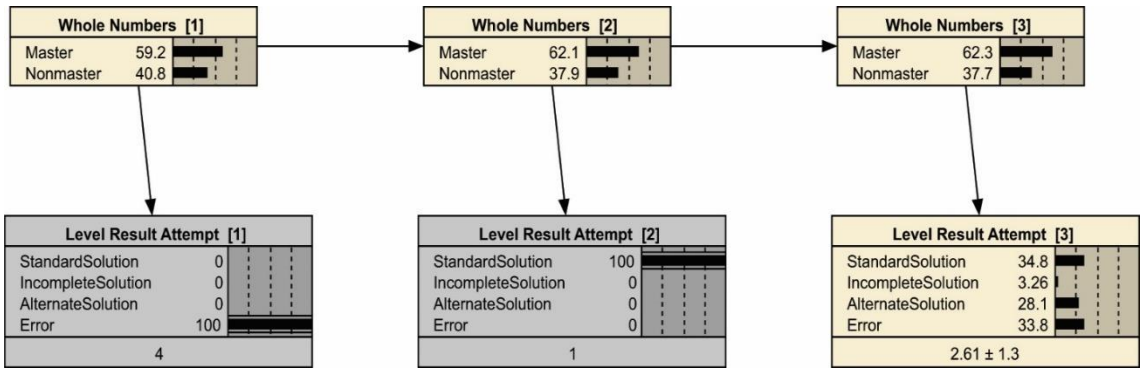

Figure 6. Netica representation of a DBN with two observed item responses.

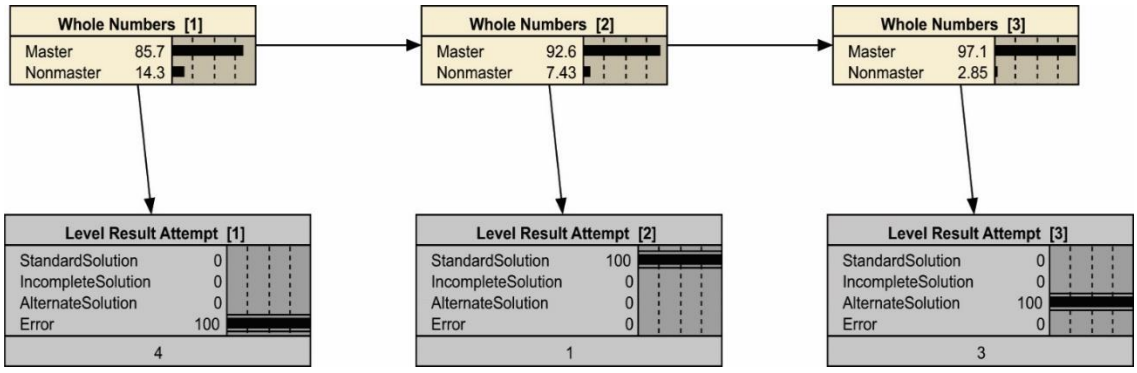

Figure 7. Netica representation of a DBN with observed responses to the first three items.

***Parameter estimation/specification*.** A DBN, in the simplest sense is defined by three parameters, or sets of parameters – prior distribution for the initial state latent variables, an observation model which defines the within-time relationships between the observed evidence and the latent variables, and a transition model which defines the between-time relationship between the latent variables. The latter two sets of parameters are specified in the form of probability tables. These tables contain either marginal or conditional probabilities depending on whether the variable in question is exogenous or endogenous, respectively. An example CPT for a transition model corresponding to two adjacent time points for that same, single latent variable consisting of two proficiency categories has been presented in Table 1 and an example CPT for a binary observable governed by a binary latent variable has been presented in Table 2. Much like with structural equation modeling (SEM) or IRT, the crux of employing DBNs in research lies in the accuracy of these parameters. Since these values are unknown in most cases, one must estimate them. Three approaches are most common with the use of DBNs – eliciting probabilities from content experts, employing some algorithm to estimate or "learn" the parameters using empirical data, or some combination of the two.

Expert opinion. One of the advantages of BNs and DBNs is that the parameters of the model (the values in the CPTs) may be more intuitive and easy to understand for content experts and end-users alike as compared to the parameters of, say a structural equation model. That is to say that it may be easier for a content expert to estimate the probability of a student being in a particular performance category based on their performance on some task then to estimate, for example a regression coefficient in an SEM model. As the name suggests, expert opinion or expert elicitation involves querying content-area experts to fill in the values in the CPTs such as those presented in Tables 1 and 2. This process is often highly structured to mitigate the bias inherent in human probabilistic reasoning (e.g., Tversky & Kahneman, 1975). Though the structure of the process varies from project to project, it typically involves the following steps (Renooij, 2001):

1. Selecting experts (i.e., those with high domain, or content knowledge). Ideally, these experts would be involved in both the development of the network or graph (e.g., defining the variables and specifying the causal structure) as well as the completion of the CPTs.
2. Training the experts on the specifics of the task as well as providing training in probability theory.
3. A context-specific, structured format for elicitation is developed. This includes developing detailed descriptions for all the of the variables in the model as well as questions to be used for elicitation (e.g., "What is the probability of Performance X on Task Y given that a student has [or has not] mastered the content?").

4. Probabilities are elicited from experts using the materials developed in Step 3. This often takes the form of individual elicitations that are then aggregated or in a panel format where consensus is targeted.
5. Some process for verifying the information gathered in Step 4 is undertaken.

The verification conducted in Step 5 can take on a variety of forms. One might use test-retest reliability (assuming at least two rounds of elicitation) to examine the stability of expert opinions or, at the very least review the values to ensure that they conform to the basic laws of probabilities. If data has been collected on the variables in the model, then the elicited probabilities can be compared to observed frequencies or cross-validation might be employed to evaluate agreement. Pitchforth and Mengersen (2013) provide an overview and extension of methods for validating expert opinions for belief networks such as DBNs.

Parameter learning. Alternatively, a researcher may opt to eschew the use of expert opinions and, instead employ an entirely data-driven approach to parameter estimation. In that case, data would first be collected using the observed variables in the network (or proposed network). Those data would then be used to estimate the relationships among the variables for a given structure (though methods for using available data to estimate the structure of the network also exist; see Murphy, 2002). This process is not unlike what is often done in confirmatory factor analysis (CFA) or IRT. Note that the use of a pre-specified network structure (or factor model in the case of CFA) implies some level of expert opinion in the parameter specification process. Two methods of parameter estimation or "learning" (the latter being more the more common term in the fields of artificial intelligence and computer science) pervade the BN/DBN

literature – maximum likelihood estimation (MLE) using the expectation-maximization (EM) algorithm and Markov chain Monte Carlo (MCMC) estimation.

Maximum likelihood estimation (Eliason, 1993; Enders, 2010 pp. 56-85; Myung, 2003) is an approach whose goal is to maximize some likelihood function. This function is a representation of agreement between sample observations and expectations based on estimated values for the model parameters. In other words, the method seeks to find parameter values that maximize the probability of producing the sample data (i.e., the likelihood of the parameters given the data). In some simple cases, this problem can be solved directly. In most cases, though, particularly when there are unobserved (i.e., latent) or missing values, a search algorithm needs to be employed to arrive at a solution. There are a variety of algorithms available to accomplish this task. The EM algorithm (Dempster, Laird, & Rubin, 1977) is the most commonly applied among these in the realm of Bayesian networks. EM is an iterative algorithm involving two alternating steps – the *expectation* step and the *maximization* step. In the expectation step, the values of the parameters are treated as known quantities which are then used to compute the best values for the unobserved variables given the parameter estimates. In the maximization step, the parameter value estimates are updated given the latent variable estimates produced in the most recent expectation step. This process iterates until some convergence criterion is met. This criterion is typically defined as a minimal change in the value of the likelihood function from one step to the next. For a BN or DBN, the goal is to maximize the likelihood of the values of the variables in the model conditioned on the observed data (X_{t1} and X_{t2} in Figure 3). Almond, et al. (2015, Chapter 9.4) provide a thorough treatment of EM estimation for Bayesian network models.

Markov chain Monte Carlo (MCMC; Gilks, Richardson, & Spiegelhalter, 1996 estimation is a routine typically employed with fully Bayesian analytic techniques. MCMC estimation involves sampling from a particular distribution that is intended to be an analog for the posterior distribution of the model parameters. The characteristics of the values drawn from this distribution are equivalent to those of the true joint posterior distribution in the long run (i.e., given a large enough number of draws; Gelfand & Smith, 1990). Arriving at that particular distribution can be accomplished through a variety of means. In the case of conjugacy (Gill, 2014), the posterior distribution is of a known form and can be sampled from directly. In other cases, an algorithm needs to be used to approximate and sample from the posterior. For univariate scenarios, an adaptive rejection sampling algorithm (Gilks, Best, & Tan, 1995) might be used. For higher dimensional problems such as DBNs with more than a few nodes, the Metropolis-Hastings algorithm (Chib & Greenberg, 1995) or the Gibbs sampler (Gilks, Richardson, & Spiegelhalter, 1996) are the most common choices, though new approaches such as the Hamiltonian Monte Carlo method (Hoffman & Gelman, 2014) are becoming more prevalent with advances in computing power.

Combined approach. The third option might be the closest analogue to common psychometric practice where subject matter experts help to define features of the model (e.g., levels of the latent variable, task scoring rules, model structure, hypothesized causal relationships between the latent variables, etc.), and then pilot data are used to estimate parameters and critique the model. This approach may also be more useful in the event that the amount of available pilot data (e.g., small sample size) is potentially insufficient for estimating model parameters with the desired level of precision. The aforementioned

work by Levy (2014) provides an applied example of the melding of evidence from subject matter experts and pilot data.

In the case of the illustrative example, the values presented in Table 3 and 4 were estimated using the EM algorithm (see Levy, 2014 for an applied example using MCMC estimation) in the Netica software package (Norsys, www.norsys.com). The parameter estimates for the measurement and transition models were constrained to be static (i.e., invariant with respect to time) in order to aid with model identification. Minimally informative start values were supplied for the measurement model CPT in order to help avoid issues with label-switching (Rodriguez & Walker, 2014; Stephens, 2000).

Identifiability. Very little is known about model identification for BNs relative to other modeling frameworks such as IRT and SEM but the flexibility of BNs may make it easier for researchers to specify models which cannot be identified from the data (Almond, et al., 2015). Work done using similar, or related models (discussed further on in this chapter) may shed light on best practices for aiding model identification with BNs or DBNs. In the latent transition analysis (LTA; Collins & Lanza, 2013) literature for example, it is recommended that researchers apply constraints to models such as specifying that the relationship between the items and the latent variable as well as the probability of transitioning from one latent state to another be equal across time. In the Bayesian knowledge tracing (BKT; Corbett & Anderson, 1995; Corbett, Koedinger, & Anderson, 1997) literature, some recommendations have been offered for dealing with model identification issues (Beck & Chang, 2007). It is not clear to what extent these solutions might apply to a more general or complex DBN, however.

Evaluation of data-model fit. Evaluating data-model fit is an important step in critiquing a model regardless of how the elements of that model were arrived upon (e.g., data-driven approaches, expert opinion). Many of the general classes of options for model criticism of BNs are similar those found in the IRT and SEM literature – graphical evaluation, fit indices, data generation, or bootstrapping routines. It should be noted that almost the entirety of the literature in this area has focused on BNs and not necessarily DBNs. It stands to reason that many of the methods discussed below should be adequate for DBNs to the extent that they are adequate for BNs but that assumption has not been evaluated directly to date.

Fit indices. As is the case with other modeling frameworks, a variety of indices of data-model fit have been proposed for use with BNs. Williamson, Almond, and Mislevy (2000) reviewed three indices commonly applied to weather prediction under a variety of model misspecification conditions and found two -- Weaver's Surprise Index (Weaver, 1948) and the Ranked Probability Score (Epstein, 1969) – to be useful for assessing global fit while Good's Logarithmic Score (Good, 1952) provided some utility as a measure of node (i.e., local) fit. In the case of model comparison (i.e., comparing the fit or utility of two or more candidate models), Spiegelhalter, Best, Carlin, and van der Linde's (2002) deviance information criterion (DIC) provides a model complexity penalized option while Gilula and Haberman (2001) put forth two methods based on entropy reduction.

Posterior predictive model checking (PPMC). PPMC focuses on discrepancies between the observed data and replicate sets of model-implied, or model-*predicted* data. Discrepancy measures are used to assess the discrepancy between the data and the model.

Large differences between the *realized* values of the chosen discrepancy measure and the *model-implied* values are a potential indicator of data-model misfit. Gelman, Meng, and Stern (1996) recommended the use of a posterior predictive *p* value (PPP) to summarize information in PPMC. PPP represents the degree of overlap between the distribution of the discrepancy measure derived from the observed data and that of the replicate data. PPMC provides a flexible platform for assessing data-model fit and conducting model criticism. As Levy, Mislevy, and Sinharay (2009) point out, the PPMC framework may offer a number of advantages over other model checking approaches. Specifically, PPMC does not necessarily rely on asymptotic theory, nor does it rely on measures with known sampling distributions. Furthermore, as Rubin (1984) points out, simple summary statistics can be used to monitor data-model fit regardless of the complexity of the models themselves.

Crawford (2014) and Sinharay (2006) reviewed options for diagnosing misfit in simple BNs including a presentation of various choices for discrepancy metrics that prove useful in the context of PPMC. These included metrics based on class membership (e.g., proportion of examinees in a particular class that responded to items in the same way), raw score (e.g., the proportion of examinees with similar raw scores that responded to items in the same way), correlations and odds ratios between item pairs, and summary statistics based on χ^2 or G^2 (Agresti, Mehta, & Patel, 1990). Other discrepancy measures have been proposed and tested in the context of Bayesian approaches to psychometric modeling, though not necessarily using BNs or DBNs (Levy & Svetina, 2011; Levy, Xu, Yel, & Svetina, 2015; Reckase, 1997; Yen, 1984; see Levy, Mislevy, & Sinharay, 2009 for a review of some of these metrics).

Graphical options. Many researchers have suggested examination of a variety of graphical representations of data-model fit as a precursor to the use of quantitative techniques. Graphical options include Bayesian residual plots which display the difference in the observed score and the PPMC-derived expected score, item fit plots which can be thought of as an empirical item characteristic curve (ICC) based on examinee equivalence classes (see Sinharay, Almond, & Yan, 2004 for a review of these two options), and direct data display (Sinharay & Almond, 2007) which display actual item responses and posterior predictive responses (using a small sample of the poster predictive replications) in a grid with individuals and items ordered by some rough indicator of ability (e.g., raw score) and difficulty (e.g., proportion correct), respectively. These plots can be compared to find areas of the testing space (e.g., low difficulty/low ability, high difficulty/low ability, etc.) which demonstrate some misfit. Echoing an earlier caveat, these data display options have only been investigated using BNs. It is not clear how they might be adapted for use with DBNs. In particular, these options may not be well-suited to examining local fit for the temporal portion of the DBN model.

Reliability. The classical test theory (CTT) notion of reliability is generally defined as the proportion of an examinee's true score (*T*) variance that is associated with variance in the observed score (*X*). This is often represented in equation form as

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} \tag{5}$$

or alternatively,

$$\rho_{XX} = 1 - \frac{\sigma_E^2}{\sigma_X^2} \tag{6}$$

where the observed score variance is the sum of the true score variance and error variance (i.e., measurement error; *E*). The greater the error variance relative to observed score variance, the lower the reliability of the test. Reliability has also been characterized as the weight of evidence for inferences about examinee abilities or claims about what an examinee can, or cannot do based on the observed evidence (Mislevy et al., 2015). The former conceptualization does not lend itself well to categorical representations of latent proficiency variables. When using DBNs, individuals are often assigned membership into performance categories or represented by a distribution over a set of performance categories (e.g., the distribution for θ in Figure 2). Each of these approaches necessitates a different method for capturing the reliability of the measure. In the case of students being placed into a performance category based upon some criteria (e.g., the category for which they have the highest posterior probability of membership), one can use a classification accuracy approach similar to what might be used with logistic regression – a confusion matrix is generated and some index of classifier effectiveness is computed. Cohen's κ (Cohen, 1968) and Goodman and Kruskal's (1954) λ have both seen use for this purpose with BNs (Almond et al., 2015, Ch. 7.5). This approach does not differentiate between an individual with a very clear category membership (i.e., one category with a very high posterior probability) and an individual with two or more categories that are close to being equally likely. For this reason, Almond et al. (2015, Ch.

7.5) recommend using an expected weight of evidence approach which weights individuals based on their posterior probability of category membership.

Related Models

There exist a variety of models which have seen relatively wide use (though often only in specific fields) and that can be considered as special instances of DBNs. The BKT model as an example, is graphically similar or, in most cases identical to what is depicted in Figure 3. Other examples include LTA models, hidden Markov models (HMM; Ghahramani, 2001; Murphy, 2002), Markov decision processes (MPD; Barber, 2012; Boutilier, Dean, & Hanks, 1999), and longitudinal diagnostic classification models (L-DCM; Kaya & Leite, 2017; Li et al., 2015; Madison & Bradshaw, in press). Table 5 presents a comparison of these models, as well as others, in terms of their relevant features. Though methodological research focused on these models may not necessarily inform the use of DBNs in the general sense, that work remains important to the goal(s) of the current study. This section will present a brief overview of each of these classes of models. Methodological literature related to these models will be summarized in Chapter 2.

Table 5.

Comparison of DBN and related models.

	Features					
	Time Points (*t*)	LVs per *t*	OVs per *t*	Nature of LVs	Nature of OVs	Decision node(s)?
Cross-sectional Models						
Generalized BN	One	Multiple	Multiple	Categorical or Continuous	Categorical or Continuous	N/A
Latent class analysis	One	One	Multiple	Categorical	Categorical	N/A
Longitudinal Models						
Generalized DBN	Multiple	Multiple	Multiple	Categorical or Continuous	Categorical or Continuous	Yes
Latent transition analysis	Multiple	One	Multiple	Categorical	Categorical	No
Longitudinal DCM	Multiple	Multiple	One/Multiple	Categorical	Categorical	N/A
Markov decision process	Multiple	One	One	Categorical	Categorical	Yes
State Space Models (SSMs)						
Particle/Kalman filter, etc.	Multiple	One	One	Continuous	Categorical or Continuous	No
Hidden Markov model	Multiple	One	One	Categorical	Categorical or Continuous	No
BKT model	Multiple	One	One	Dichotomous	Dichotomous	No

State-space models. State-space models (SSM; Ghahramani, 2001; Murphy, 2002) represent a family of models all of which can be thought of as a special case of a DBN. An SSM is a graphical model which captures the probabilistic dependence of an unobserved "state" (i.e., a latent variable) variable and observed variables. These models typically consist of only one observable and are often used with time series data to model the "true," latent state of several observations of the same variable. Both the latent and observed variables can be either continuous or categorical. Several common variants are used in practice. Arguably the most widely used is the Kalman filter (Harvey, 1990) which models a continuous state variable and specifies Gaussian error distributions for the observed variables (Chen & Brown, 2013). The discrete-state version of a Kalman filter is known as a hidden Markov model (HMM; Ghahramani, 2001). HMMs differ from DBNs in that they model only one hidden state (Ghahramani, 2001). An HMM with categorical observables is identical to the DBN discussed earlier and presented in Figure 3. The BKT model is a special case of the HMM model wherein the hidden state is always binary (e.g., "master," "non-master"; Yudelson, Koedinger, & Gordon, 2013).

Markov decision processes (MDP). MDPs represent an extension Markov chain models (e.g., HMMs) wherein there is some agent interaction with the model at each time slice. That is to say that, between two adjacent time points there might be some action, or "decision" which augments the transition probabilities. For example, a student might receive some intervention or specialized feedback between two attempts at a performance task. MDPs for which not all of the variables are observables are called partially observed Markov decision processes (POMPDs). In a way, these models might be thought of informally as a latent growth curve model (LGCM) with covariates in that you have some

growth process which is specified as a function of some predictor. In the case of MDPs, however, the effect of this predictor may change from one time point to the next whereas with LGCMs the predictor is typically static. The utility of MDPs lies in the ability to plan optimal paths to desired outcomes. For example, one might use an MDP to map out the optimal sequence of instructional modules to present to a student to lead them to content mastery in the minimal number of time points. Almond (2007a; 2007b), LaMar (2018), and Rafferty et al. (Rafferty, Brunskill, Griffiths, & Shafto, 2011) provide an overview of MPDs being used in an educational context.

Latent transition analysis. Latent class analysis (LCA) and latent transition analysis (LTA; see Collins & Lanza [2013] for a detailed treatment of both) share much of the same relationship as BNs and DBNs. LCA models, which can be considered a special case of a BN, include categorical latent variables that govern a set of categorical observables. The five-item BN presented in Figure 1 can be considered an ordinal LCA model. LTA models represent a longitudinal extension of LCA and, as such can be considered as a special case of a DBN. In an LTA model, the structure from an LCA is repeated for a number of time steps. The categorical latent variables from adjacent time steps are connected by a directed arrow suggesting that an individual's state at any particular time is dependent on their state at the previous time point. Figure 8 depicts a generalized path diagram for an LTA model spanning T time points with K indicators per time point. Note the lack of directed arrows that span more than one time point which suggests a Markov process (i.e., the future is independent of the past conditional on the present). Being a less general model, LTA models are necessarily less flexible when compared to DBNs. Examples of LTA models in the literature to date include just one

latent proficiency variable making them more like an HMM with multiple categorical observables whereas many examples of DBNs with multiple proficiencies exist (see the next chapter for examples). Furthermore, the typical application of LTAs do not consider the model from a Bayesian network perspective. Were such a perspective adopted, these LTA models would share the same advantages that DBNs offer – quick updating, an ability to handle very complex systems of variables, and suitability to situations requiring real-time diagnostic inferences.

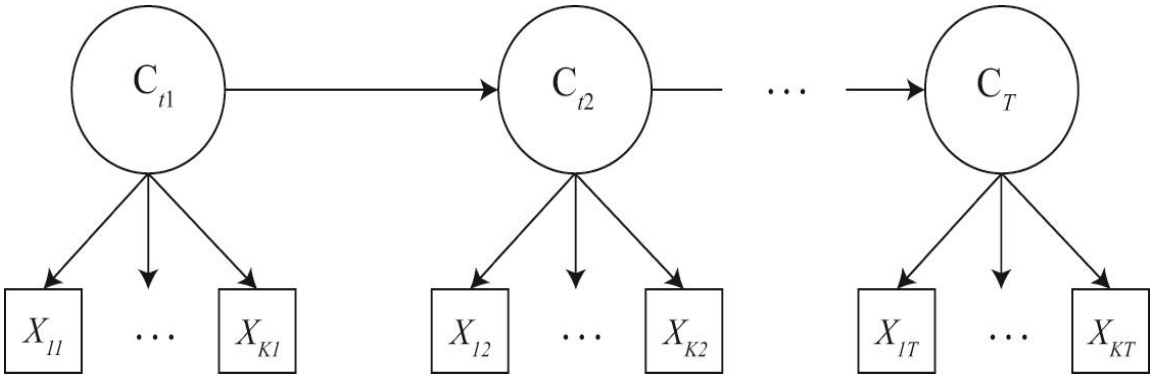

Figure 8. General Latent Transition Analysis model.

Longitudinal Diagnostic Classification Models. Diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010) are a class of models used to diagnose an examinee's mastery status (e.g., "master"/"non-master") on a collection of skills for the purpose of providing diagnostic feedback to teachers, for example, to help in guiding further instruction and/or remediation. DCMs typically contain multiple latent skills and multiple tasks, or items. The relationship between the skills and tasks is specified using a Q-matrix in which a "1" is coded at the intersection of a task and skill if that task requires the skill to complete, while a "0" is coded if the task does not require the skill. Most applications contain multiple tasks per latent skill, but this is not necessarily the case. Longitudinal DMCs (L-DCMs), then represent a longitudinal extension wherein the focus

is on modeling the change in skill profile across multiple time points. The relationship between DCMs and L-DCMs is similar to the relationship between LCA and LTA and between BN and DBN. L-DCMs represent a relatively new area of research but several recent examples exist in the literature (Kaya & Leite, 2017; Li et al., 2016; Madison & Bradshaw, in press)

Chapter 2

LITERATURE REVIEW

Though still relatively unused in educational measurement, DBNs have seen some limited investigation and application. This section will summarize the methodological literature related to DBNs that can inform the current study. In addition, a selection of publications related to the application of DBNs in practice will be presented for references purposes as well as to provide justification for design choices made in Chapter 3. Classes of models for which there are very few (if any) methodological examples in the literature (MDPs, L-DCMs) will not be covered in this chapter.

Methodological Literature

Generalized DBNs. For the purposes of the current work, a "generalized DBN" is defined as any DBN which does not fit into one of classes of models presented below (e.g., state-space model, latent transition model, etc.). To date, the author is not aware of any methodological studies in the literature that focused on generalized DBNs in an educational or psychological context (or any other context, for that matter). Furthermore, the research using models that can be considered as a special case of a DBN do not acknowledge that relationship and/or do not present the models in the context of the Bayesian network framework. There are examples of literature that focus on BNs (see Almond, Yan, & Hemat, 2007; Culbertson, 2014; Guo, Gao, Di, & Yang, 2015 as examples) but it is not clear to what extent the conclusions drawn in those studies can be applied to DBNs. In particular, those studies necessarily ignore the aspects of DBNs that separate them from BNs (i.e., the transition model and related issues such as rollup). Additionally, though one can find a few examples of methodological investigations using the Bayesian knowledge tracing model (BKT; Corbett & Anderson, 1995; Corbett,

Koedinger, & Anderson, 1997; see Coetzee, 2014 [sample size recommendations], Pardos, Bergner, Seaton, & Pritchard, 2013 [meeting MOOC challenges]; Qui, Qi, Lu, Pardos, & Heffernan, 2010 [modeling delay in attempts]; Yudelson, Koedinger, & Gordon, 2013 [modeling student-specific variability] for examples of methodological investigations of the BKT model) – a special case of a DBN (Murphy, 2002) – it may be the case that those conclusions do not generalize to the more flexible DBN framework. For example, the BKT model typically includes a single latent ability with two performance categories (i.e., binary) identified by a limited number of tasks. This specification suggests that the methodological work using BKT probably ignores model complexity as a factor influencing model performance. That work would then have limited utility to a researcher using a more complex design such as that presented by Levy (2014) which contains many latent proficiencies as well as hidden nodes representing misconceptions the examinee may harbor.

State-space models. From the descriptions presented in Chapter 1, one can see that methodological work involving SSMs such as HMMs or the BKT model may be relevant to the use of the more general DBN, though necessarily lacking in some areas (e.g., multiple and/or polytomous hidden states). Unfortunately, very little of this research has been conducted to date. This is most likely a product of the areas in which these models have been most commonly applied in the context of education and learning – ITSs and data mining. Much of the ITS literature involves models specified using expert opinion. This process is not dependent on notions such as sample size, missing data, estimation algorithm, etc. Educational data mining applications, on the other hand, involve some structure and/or parameter learning almost without exception. Like most

data mining applications, however, these studies often involve extremely large sample sizes which mitigate the importance of some of the common estimation concerns (missing data, data sparseness, insufficient information, etc.). Furthermore, data mining tends to be an exploratory endeavor which is not often as concerned with model critiquing in the way that SEM researchers, for example, might be. That said, there are a limited number of methodological investigations which may be of interest to the educational researcher interesting in DBNs. Many of these studies were aimed at proposing and/or validating extensions or modifications to the HMM/BKT model (e.g., Klingler, Käser, Solenthaler, & Gross, 2015; Pardos, Bergner, Seaton, & Pritchard, 2013; Pardos, Dailey, & Heffernan, 2010; Pardos & Heffernan 2010; 2011; Pavlik, Cen, & Koedinger, 2009; Qiu, Qi, Lu, Pardos, & Heffernan, 2011; van de Sande, 2013; Yudelson, Koedinger, & Gordon, 2013) or improving the computational efficiency of the estimation of such models (e.g., Fisher, Walsh, Blaha, Gunzelmann, & Veksler, 2016). Very few parameter recovery studies or investigations of the effect of sample size and other modeling factors have been carried out.

Coetzee (2014) examined the sample size requirements for the BKT model under a limited set of conditions including the varying true values for the guess/slip (i.e., measurement quality), learn (i.e., transition probability), and prior (i.e., initial probability of mastery) parameters. He found that the most conditions required a sample size of at least $N = 1{,}000$ in order to achieve a desirable level of estimation accuracy where "accuracy" was operationalized as the standard deviation of the parameter estimates across multiple replication (termed "efficiency" in Chapter 3 of the current work). The findings further suggested that parameter estimation is better for true parameter values

closer to their upper/lower bounds (zero and one, respectively in this case) and that the patterns of interactions between the conditions (differing true parameter values) were "complex."

Nooraei, Pardos, Heffernan, and de Baker (2011) investigated the number of time slices necessary for accurately estimating model parameters in a BKT model. Their results suggest that using only the most recent five to 15 data points can yield parameter estimates that are within 1% of estimates coming from the use of 40 or more time slices in terms of accuracy (defined by root mean squared error, in this case). It's worth noting that, in the design of this study the most recent five to 15 time slices were used out of a sequence of 60+. It may be the case that data coming from a learner having only five to 15 repetitions as opposed to using data from the final few in a sequence of 60 or more repetitions may result in less accurate parameter estimates due to factors such as learner familiarity with the system (i.e., a practice effect) or some other phenomena.

Lastly, Choi (2012) evaluated two models in the context of modeling learning progressions – a typical DBN as well as a DBN with covariates which is very similar in structure to the Markov decision process models discussed in the next section. This evaluation explored sample size, test length (termed "task size"), and various types of CPTs, priors, and transition matrices in terms of the resulting parameter estimation accuracy. The study included only two sample size and test length conditions (i.e., a 2x2 design). For the simple DBN, 30 tasks (versus 9) and a sample size of N = 1,000 (versus N = 100) yielded statistically significant improvement in estimation accuracy. The same held true for the DBN w/ covariate model. For both models, the quality of the task, where

more highly discriminating tasks are considered to be of higher quality, was a significant contributor to estimation accuracy.

Latent transition analysis. The literature on LTA is fairly extensive and dates back almost a quarter-century. Very little of this literature, however, has put forth any guidelines for practitioners wishing to implement LTA in terms of sample size recommendations, estimation routine suggestions, ensuring model identification, or other common methodological concerns.

Several studies have examined the issue of sparseness with respect to the contingency tables. These tables often become unmanageably large even for relatively simple models. These studies have concluded that estimation of LTA model parameters using the EM algorithm is rather robust even in the presence of sparse contingency tables (Collins & Tracy, 1997; Collins, Wugalter, & Fidler, 1996). This somewhat contradicts the findings of Levy (2014) which found that estimation became rather cumbersome in the presence of high levels of sparseness using DBNs for data from a game-based assessment. Collins and Wugalter (1992) also studied the trade-off between the added benefits of additional indicators and the cost of increasing the size of the contingency tables by doing so. In keeping with later studies, they determined that the increase in sparseness presented little issue relative to the increased precision of estimates for the latent variable that go along with adding additional indicators (under the assumption that these additional indicators fit well within the context of the model). On the subject of data-model fit, Collins, Fidler, Wugalter, and Long (1993) examined the performance of three fit indices – χ^2, G^2, and the *power-divergence* statistic (Read & Cressie, 1988) in the presence of sparseness. Their results suggested that none of the indices are ideal and,

instead argued for the use of re-sampling techniques (e.g., bootstrapping). Finally, Baldwin (2015) explored the issue of power to detect effects of transition probabilities with LTA for different sample sizes, transition matrix specifications, levels of model misspecification, class size ratios (i.e., equal versus unequal), and threshold magnitudes. As one would expect, the level of model misspecification and the homogeneity of class membership (i.e., few individuals with similar membership probabilities for more than one latent class) had a significant impact on statistical power. Beyond those factors, sample size and class size were the primary drivers of power with larger samples and larger class sizes yielding higher power. The results suggest that sample sizes as low as N = 100 might be adequate under ideal conditions (e.g., well-specified model with large thresholds) but deviations from the best-case scenario can quickly escalate the sample size needed to return adequate power. Some conditions suggested that sample sizes in the 500-5,000 range might be needed even when the model is well-specified. Sample sizes as large as N = 10,000 were tested and did not yield power above 0.80 under some of the less favorable conditions (e.g., when using poorly specified models). It is not clear to what extent these results apply to DBNs but many of the issues that plague LTA, such as sparseness and category separation are also present in DBN models. When combined with the often more complex nature of DBNs, these studies suggest that very large samples may be needed to adequately calibrate DBN model parameters even under favorable conditions.

Applied Examples

This section aims to review examples of DBNs being applied to novel problems in educational assessment. This review is not intended to be comprehensive but rather is

intended to provide readers new to DBNs with examples of how DBNs might be deployed in practice. Several examples are briefly summarized here. Relevant information from each example, if available is summarized and presented in Table 6.

Table 6.

Summary of applied examples using DBNs.

	Carmona et al. (2008)	Conati & Maclaren (2009)	Iseli, et al. (2010)	Rowe & Lester (2010)	Sabourin, Mott, & Lester (2013)	Levy (2014)
LVs Per Time	1	?	2	4	3	1 to 15
OVs Per Time	≤ 4	?	3	?	16	1 to 18
Response Categories for Latent Variables	Polytomous	Binary	Polytomous	Binary	Polytomous	Polytomous
Time Points	?	?	~20	?	4	19 levels
Lag Between Time Points	?	3-10 seconds	?	?	Short (< 1 hour)	?
Sample Size	?	66	36	116	260	851
Parameter Estimation	?	Specified?	Specified	Specified	Estimated (EM)	Estimated (MCMC)

Levy (2014) detailed a DBN-based scoring model for the educational game *Save Patch* – a modified excerpt of which was presented earlier as an illustrative example of a DBN. During the game, players (examinees) complete levels of the game by using math skills to navigate from the beginning of the level to the end. Results of the study suggested that the DBN framework using MCMC estimation is suitable for use with game-based assessment but noted issues with estimation resulting from data sparseness due, in part to the fact that not all variables were assessed in each level. This suggests a need for games that are designed with robust psychometric analyses in mind such that there is a synergy between the conditions that make the game-experience engaging and educational and conditions that produce data that are suited for the available analytic techniques.

Carmona et al. (2008) developed a DBN for characterizing students' learning styles based on the learning objects that those students choose to interact with as well as their reported rating of those objects (scored 1-4 "stars"). The work was exploratory in that little validation or model critiquing was conducted. Parameters were specified using expert opinions and only a small sample of student data was collected.

Conati and Maclaren (2009) designed a DBN-based model for detecting the emotional state of users interacting with an educational game (*Prime Climb*; developed by the EGEMS group at the University of British Columbia) with the goal of improving student outcomes based on the theory that increased emotional engagement leads to increased attention and learning (Conati, 2002; Ortony, Clore, & Collins, 1988). The model was developed using data collected over several rounds of user studies. These data provided the basis for specifying the structure and parameter values for the model.

Results suggested that the predictive portion of the model yielded predictions that were more accurate than what would be expected by simply choosing the most likely emotional state (i.e., the population mode). The diagnostic portion of the model was not specifically evaluated in the paper.

Interactive Narrative Environments is another area of research in which DBNs have been applied. These narrative environments might be found in role-playing games (RPGs) centered on learning or exploration. Rowe and Lester (2010) present a DBN for updating beliefs about the user's knowledge based on their interactions with the environment. These beliefs are used to tailor the narrative elements that the user is presented with. Posterior category membership was compared to the results of a knowledge post-test for the purposes of assessing the accuracy of the model. A model which assigned a uniformly distributed, random probability to each of the knowledge nodes was used for comparison. As one would expect, the target model significantly outperformed the random model in terms of accuracy. The authors note that accuracy might be improved by collecting data from a larger sample to learn the model parameters as opposed to "hand-authoring" (Rowe & Lester, 2010, pg. 5) the model.

Using the same gaming environment as Rowe and Lester (2010), Sabourin, Mott, and Lester (2013) developed an early detection system for a learner's self-regulated learning (SRL) capabilities using a DBN guided by research showing that student with low SRL abilities may need scaffolding when operating in a largely self-guided environment such as *Crystal Island*. Early detection of a student's SRL status provides an opportunity for that scaffolding to occur.

Iseli et al. (Iseli, Koenig, Lee, & Wainess, 2010) validated a DBN used for automated scoring of complex tasks. They were interested in comparing the performance of DBN-based performance scoring software and subject matter experts trained in scoring the tasks in terms of their ability to identify satisfactory performance with the goal of determining whether the automated scoring approach might eventually be able to replace human raters. Their network was specified in collaboration with subject matter experts. All told, there was a high degree of agreement between the automated and judge-scored simulations, though the automated scoring algorithm seemed unable to view the examinee's performance from a holistic perspective. This, as the authors note, is evidence of the difficulty in developing a DBN that approaches a full representation of human knowledge even for a very specific domain. A more general domain would likely result in very long lead times to develop the DBN scoring model.

When viewed together, one can see that the groundwork is currently being laid for fully adaptive and automated versions of games or simulations. In such an environment, the content relevant aspects of the game experience such as the context, domain content, performance tasks, and even the scoring of complex tasks might be adapted to the user's ability level and interest set to increase the efficiency of knowledge assessment as well as to maximize learner engagement. This, in many cases might be accomplished with a need to collect only a relatively small amount of background information on the learner (via a short survey, for example) and log data from a short period of the learner's gameplay experience. Further development of these applications will be supported by methodological research on various aspects of the use of these models, including

parameter estimation, reliability analyses, and model criticism. The next chapter describes a proposed study pursuing the first of these aspects.

Chapter 3

METHOD

The goal of the current work was to use Monte Carlo simulation methods to evaluate parameter recovery characteristics for DBNs (and models which can be considered as a special case of DBNs) under a variety of testing conditions. Factors hypothesized to impact the quality of parameter recovery included information quantity, information quality, model features (number of time slices, number of items per time slice), and true values for the transition probability and initial mastery probability.

For the purposes of the current work, information quantity was represented as the number of cases, or "examinees" in the data set (i.e., sample size; *N*) while information quality, or measurement quality was represented by the true values of the parameters in the CPT for the relationship between an item and a latent variable. More specifically, the latter was represented by $P(X=1|\theta=M)$, or the probability of a "master" providing a correct response to an item and by $P(X=1|\theta=NM)$, or the probability of a "non-master" providing a correct response. Larger values for the former and smaller values for the latter represent better measurement quality (i.e., the item provides more information as to the distinction between the two mastery classes). For the sake of simplicity, all conditions in the study assumed symmetry in these values. That is to say that $P(X=1|\theta=NM)=1-P(X=1|\theta=M)$ and vice versa.

The remainder of this chapter will describe the design choices made for the current study. The study was structured in multiple phases with Phase 1 being an exploratory pilot study intended to identify salient design factors and appropriate true

parameter values for maximizing the utility of Phase 2. This first phase included a limited number of conditions and was aimed at honing in on the best combination of conditions for identifying points at which parameter recovery becomes unreliable. Phase 2, then included the full complement of experimental conditions. All features of the study other than the number of experimental conditions (e.g., data analytic approach, outcome criteria, data generation procedures) remain unchanged from Phase 1 to Phase 2.

Phase 1

Design. The current manipulated six factors hypothesized to impact parameter recovery for DBNs with latent variables – information quantity and quality (discussed above), the true values for initial probability of mastery and transition probability, and model structure (number of items per time slice and number of time slices). In Phase 1, two values for each factor were used for a total of $2^6 = 64$ experimental conditions. The proposed values for use in Phase 1 were guided by the literature reviewed in Chapter 2 and were chosen to represent realistic upper and lower bounds for what has been used in practice. Table 7 presents the chosen values as well as any relevant guiding citations.

Model specifications. All of the proposed models for use in Phases 1 and 2 contained only one latent variable per time slice. All latent and observed variables were specified as having two categories (e.g., *Master*/*Non-master* for latent variables and *Correct*/*Incorrect* for observed variables). Given the proposed values for the number of items per time slice, it can be noted that these models can be classified as either hidden Markov models (when $J = 1$) or, more specifically a Bayesian knowledge tracing model

given the dichotomous nature of the latent and observed variables, or latent transition models (when $J = 5$).

Table 7.

Model condition values for study design (Phase 1).

	Value(s)	Reference(s)
N	{200, 1000}	Baldwin (2015); Coetzee (2014)
$P(X_{j,t}=1\|\theta_t=M)$	{"Low" = 0.60, "High" = 0.90}	Baldwin (2015)
$P(X_{j,t}=1\|\theta_t=NM)$	{"Low" = 0.40, "High" = 0.10}	Baldwin (2015)
$P(\theta_{t1})$	{"Low" = 0.20, "High" = 0.40}	Coetzee (2014)
$P(\theta_{t2}=M\|\theta_{t1}=NM)$	{"Low" = 0.20, "High" = 0.40}	Baldwin (2015); Coetzee (2014)
J (per t)	{1, 5}	Baldwin (2015)
T	{5, 10}	Nooreai et al. (2010)

Note. NM = non-master; M = master.

In order to aide with model identification, the transition matrices were constrained to equality for all time slice pairings (i.e., $P(\theta_{t+1} = M \mid \theta_t = NM) = P(\theta_{t+2} = M \mid \theta_{t+1} = NM) = ... = P(\theta_T = M \mid \theta_{T-1} = NM)$) and all CPTs describing the relationship between the item(s) and their respective latent variables were also constrained to equality (i.e., $P(X_{j1,t1} = 1 \mid \theta_{t1} = M) = P(X_{j1,t2} = 1 \mid \theta_{t2} = M) = ... = P(X_{j1,T} = 1 \mid \theta_T = M)$). These choices are in keeping with commonly applied model constraints (see Collins & Lanza [2013] for a discussion of such constraints in the context of latent transition models). Finally, the "once a master, always a master" or "no skill regression" assumption described in Chapter 1 was encoded into the transition matrices.

Data generation. Data were generated using the "rbn" function included in the "bnlearn" package (Scutari, 2010) in the R computing environment (R Core Team, 2017). Several large-sample (N = 10,000) data sets generated using this function were fit to their respective models using complete data (i.e., the values for the latent variables were not obscured) in various software packages (Netica, "bnlearn," and the "catnet" package in R [Balov & Salzman, 2017]) to ensure that the estimated parameters matched the values supplied to the data generation function. These tests suggested that the data generation process is valid.

As was mentioned previously, the assumption of no skill regression was applied in the analyses. This is a *partial* constraint, in the sense that one row of the CPT for the transition matrix is fixed, while the other row is freely estimated. Netia, however, does not allow for partial constraints to be applied to CPTs; the user must either fix all of the cells or none of the cells in a CPT. To enact the partial constraint, a workaround was applied by including two "dummy" variables in the model (Dummy1 and Dummy2). Dummy2 is specified as being dependent on Dummy1 and, during the data generation process, both Dummy1 and Dummy2 are represented by columns consisting entirely of codes corresponding to examinee mastery. The CPT for the relationship between Dummy1 and Dummy2 is then constrained to be equal to the transition matrices describing the relationship between the latent variables at adjacent time slices. These steps, essentially provide a great deal of information to the estimation of $P(\theta_{t+1} = NM \mid \theta_t = M)$ and $P(\theta_{t+1} = M \mid \theta_t = M)$, which should be 0 and 1, respectively given the model assumptions, while providing no information to the estimation of

$P(\theta_{t+1} = M \mid \theta_t = NM)$, which is a parameter of interest in the current study. Pilot testing indicated that this workaround successfully implemented the desired partial constraint.

Data analysis. Models were fit using the C-Netica API (Norsys Software Corp., 1995-2017) interfaced with R using the RNetica package (Almond, 2017). All model parameters were estimated using the expectation-maximization algorithm (see Chapter 1) with the maximum number of iterations set to 10,000. All conditions analyzed a total of *R* = 1,000 replicate data sets.

Initial tests indicated the likelihood of label-switching under some conditions. Given that there is no natural order to the latent proficiency categories, it is not uncommon for the category assigned to the more proficient examinees (with proficiency being an arbitrary distinction from the perspective of the model) to change from one estimation attempt to the next (i.e., the labels switch). This is caused by the likelihood function for the estimation routine being bimodal and results in inconsistent results when aggregated across multiple replications. Two measures were taken to correct this. First, minimally informative start values were supplied to the software for the expected probability of a correct response for "master" and "non-master" examinees. These start values were set to $P(X = 1 \mid \theta = M) = .51$ and $P(X = 1 \mid \theta = NM) = .49$ with node experience (akin to the number of prior cases observed) set to one. This serves to encode the minimal amount of prior evidence allowed by the software in order to guide the estimation algorithm towards the desired mode in the likelihood function. If one thinks of label-switching as being an issue of bimodality (i.e., there are two solutions that are equally as likely) then this measure essentially starts the estimation routine closer to one

of the two solutions for the sake of consistency across replications. Second, a data integrity check was performed after every estimation attempt to ensure that the plausible assumption of "masters" being more likely to correctly endorse an item than "non-masters" was met. In the event that the assumption was not met, that replication was discarded and re-attempted. Initial trials suggested that these measures were able to completely remove any instances of label-switching.

Parameter recovery criteria. Parameter recovery was assessed via four commonly applied criteria – bias (or "raw" bias), relative bias, accuracy, and efficiency (Bandalos & Leite, 2013). Values for these indices were recorded for a total of four parameters – the initial probability of mastery ($P(\theta_{t1} = M)$), the transition probability ($P(\theta_{t+1} = M \mid \theta_t = NM)$), and the probability of a correct response to any one item associated with the latent variable at any one time slice for both a master ($P(X_{j1,t1} = 1 \mid \theta_{t1} = M)$) and a non-master ($P(X_{j1,t1} = 1 \mid \theta_{t1} = NM)$). As mentioned above, the CPTs from which the conditional probabilities of a correct response are derived were constrained to equality across all items and time points (i.e., $P(X_{j1,t1} = 1 \mid \theta_{t1}) = P(X_{j2,t1} = 1 \mid \theta_{t1}) = P(X_{j1,t2} = 1 \mid \theta_{t2}) = ... = P(X_{J,T} = 1 \mid \theta_T)$). The practical implications of parameter recovery quality (or lack thereof) was assessed via classification accuracy.

Raw bias captures deviations between an estimator and the parameter it is intended to estimate. This quantity, $B(\hat{y})$ is defined as

$$B(\hat{y}) = (\hat{y}_r - y) \tag{7}$$

where *y* is the true value for the parameter and $\hat{y}_r$ is the estimated value for a particular replication *r*. This quantity can then be averaged across *R* replications to produce $\bar{B}(\hat{y})$, defined as

$$\bar{B}(\hat{y}) = \frac{\sum_{r=1}^{R}(\hat{y}_r - y)}{R} \tag{8}$$

where *R* is the total number of Monte Carlo replications conducted. Bias is sign-dependent and captures systematic overestimation (positive) or underestimation (negative). Dividing the bias value through by the true parameter value yields the relative bias ($B(\hat{y})_{REL}$). This value is typically, then multiplied by 100 and interpreted as bias as a percentage of the true parameter value. Averaging across the *R* replications yields $\bar{B}(\hat{y})_{REL}$. Equation (9) presents this notion mathematically.

$$\bar{B}(\hat{y})_{REL} = \frac{\sum_{r=1}^{R}\left(\frac{(\hat{y}_r - y)}{y} x100\right)}{R} \; . \tag{9}$$

As an example, a true parameter value of 0.2 and a bias estimate of .01 can be interpreted as 5% positive bias. Muthén & Muthén (2002) argued that relative bias values less than 5% (positive or negative) are ignorable, values between 5% and 10% are moderately biased, and values greater than 10% are substantially biased.

Root mean squared error ($RMSE(\hat{y})$, referred to henceforth as RMSE) will be used to assess the accuracy of the parameter estimates. RMSE is defined as

$$RMSE(\hat{y}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{y}_r - y)^2}{R-1}} \quad . \tag{10}$$

Lower values indicate greater precision (i.e., lower variability) in the parameter estimates and/or less bias in the parameter estimates (Bandalos & Gagne, 2012). That is to say that RMSE is an indicator of both the variability and the bias of the estimates (Mood, Graybill, & Boes, 1974).

Efficiency ($Eff(\hat{y})$, referred to henceforth as efficiency) is defined as

$$Eff(\hat{y}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{y}_r - \bar{\hat{y}})^2}{R-1}} \tag{11}$$

where $\bar{\hat{y}}$ is the mean of the parameter estimates, $\hat{y}_r$ is the parameter estimate for a particular replication, and *R* is the total number of replications. This quantity, then represents the standard deviation of the parameter estimates across the *R* replications. Values closer to zero suggest a more efficient (or more stable) parameter estimation process.

Finally, classification accuracy was used to assess the potential impact of the quality of the parameter estimates in terms of the ability to correctly estimate a student's proficiency classification ("master" and "non-master" in the current study). It is defined as the proportion of total cases (*N*) for which the estimated proficiency category is equal to the true proficiency category. As an example, Table 8 presents a confusion matrix for estimated and true classifications for a scenario with *N* = 1,000 cases. The upper-left and

lower-right cells represent correct classifications. For this scenario, the classification accuracy (*CA*) would be calculated as

$$CA = ((187 + 748) / (187 + 748 + 13 + 52)) * 100 = 93.5\%. \quad (12)$$

In the current study, the estimated proficiency category was defined as the posterior mode for the latent proficiency variable at the terminal time point of the model (i.e., posterior mode of θ_{t5} for a model with $T = 5$ time slices). Classification accuracy as recorded under two scenarios – one using the training data (i.e., the same data used to estimate model parameters) to conduct inference and another using a newly generated validation data set. The validation set was generated using the same procedure as was used to generate the training data thus making it data from the same population, but not from the same specific cases (see James, Witten, Hastie, & Tibshirani [2013] for an overview of cross-validation terminology and methods). The classification accuracy resulting from the use of the training set should exceed that resulting from using the validation set (given a sufficient sample size) due to overfitting. Classification accuracy was aggregated by calculating the mean of the values across the *R* replications. The acceptability of a particular classification accuracy value is highly dependent on context (e.g., the source of the data and the stakes of the application). As such, the current study focused only on patterns of results under the assumption that higher classification accuracy values were preferable to lower values.

Table 8.

Example confusion matrix for calculating classification accuracy.

	True = *NM*	True = *M*
Estimate = *NM*	187	52
Estimate = *M*	13	748

Note. NM = non-master; M = master.

Simulation workflow. Figure 9 portrays the steps involved in the Monte Carlo study by replication as a flow chart.

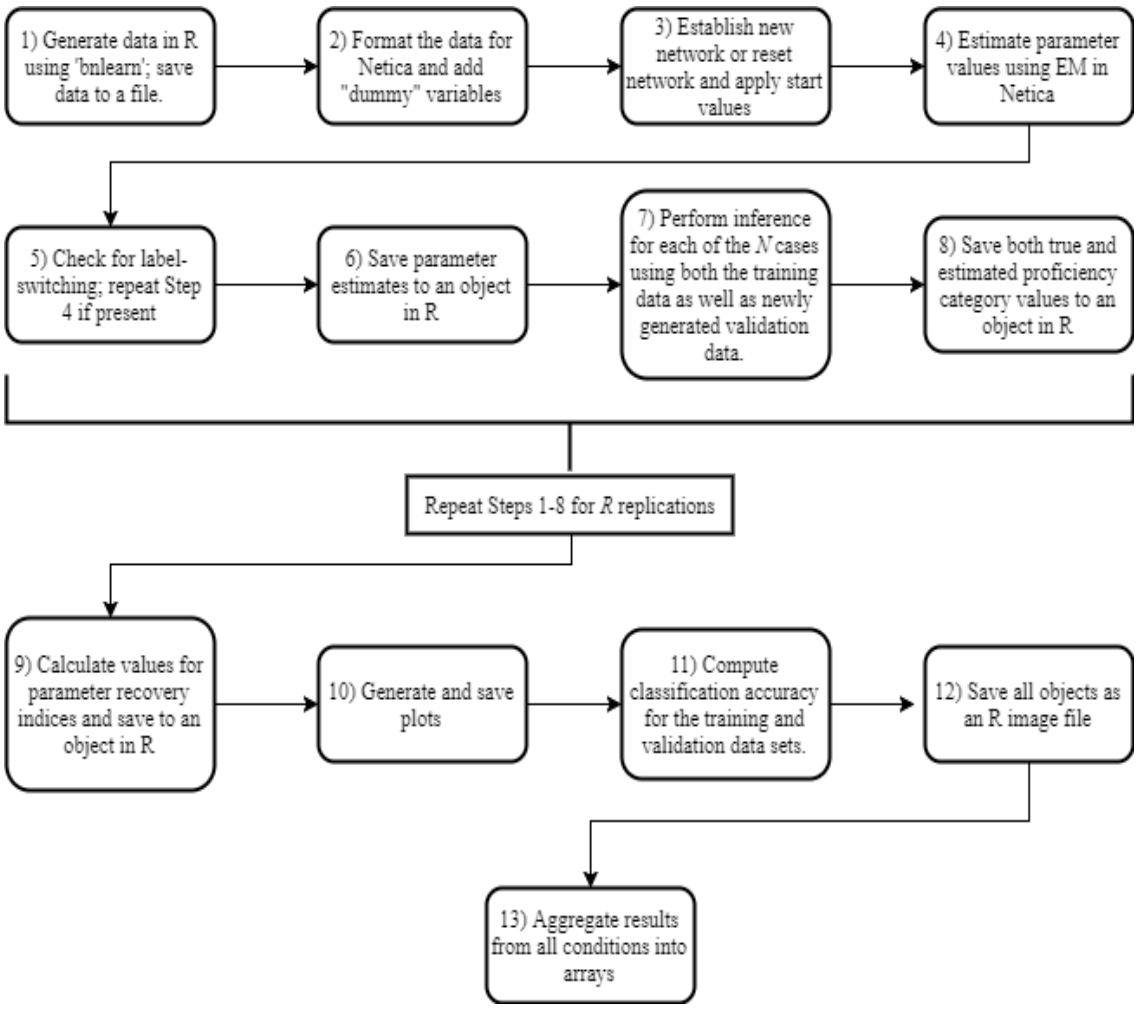

Figure 9. Flowchart for Monte Carlo study procedure.

Research hypotheses. The hypotheses related to parameter recovery and performance indices are as follows:

1. It was expected that parameter recovery index values would improve (i.e., raw and relative bias approach zero, RMSE and efficiency decrease, classification accuracy approaches 100%) as information quantity (sample size) increased.

2. It was expected that parameter recovery index values would improve as information quality (measurement quality) increased. Furthermore, it was expected that classification accuracy, in particular would be extremely poor (relative to other conditions) as item information approached zero. In the context of the current study, this occurs when both masters and non-masters have the same probability of correctly endorsing an item.
3. It was expected that parameter recovery index values would improve as the number of time slices and the number of items per time slice increased.
4. It was expected that parameter recovery index values would improve as the true values for the transition probability and initial mastery probability approached either one or zero.
5. Finally, it was expected that there would be noticeable interactions between the effect of these design facets on the values of the parameter recovery indices. For example, the impact of decreased measurement quality would likely be more noticeable as the number of items per time point decreased.

Phase 2

Phase 2 of this work expanded the number of experimental conditions, to pursue combinations of conditions under which the values for the parameter recovery criteria degrade. The conditions for Phase 2 were chosen based on the findings from Phase I. As such, the conditions for Phase 2 are described following a presentation of the results from Phase I.

Summary

The current work used a Monte Carlo simulation study to investigate the impact of information quality, information quantity, model features, and the true values for certain parameters on parameter recovery for DBNs with latent variables. The study was broken into two phases – the first an exploratory attempt at informing the most efficient combination of values for the manipulated design facets to be included in the second. Parameter recovery was captured by bias, relative bias, RMSE, and efficiency while the implications of the parameter recovery were assessed via classification accuracy. It was hypothesized that bias, relative bias, RMSE, efficiency, and classification accuracy would improve as sample size and measurement quality increased. Furthermore, it was expected that adding more items per time point and more time points would also improve parameter recovery.

Chapter 4

RESULTS

Phase 1 Results

The goal of Phase 1 was to explore the broad parameter space with regard to the manipulated design facets (sample size, measurement quality, test length, number of time points, etc.) in order to identify a more targeted set of conditions to be explored in Phase 2. A total of 64 experimental conditions were examined in the first phase (see Table 7) with six indices (raw bias, relative bias, root mean squared error, efficiency, classification accuracy – training, and classification accuracy – validation) being used to capture the quality of parameter recovery. The main effects (i.e., mean outcome for a particular facet marginalized over all other facets) for each of the four parameters of interest across each of the six outcomes are presented in Tables 9-14 while the disaggregated results for Phase 1 are presented in Appendix A. The values in these tables were generated using the "DC-100" dummy coding strategy which will be covered in the next section. Examination of these tables reveals some insights which will simplify the presentation of the remaining results. First, it can be noted that the difference between the validation and training set classification accuracy values was negligible. As such, only the validation set values, the more conservative of the two approaches, will be mentioned going forward. Second, recovery of the measurement model parameters was sufficient-to-excellent across all conditions while recovery of the initial probability of mastery and the transition probability proved more problematic. The presentation of results going forward, then will focus primarily on the latter two parameters. The remainder of this section is organized by results pertaining to each design facet and supported by relevant evidence.

Table 9.

Marginal means for bias by experimental factor and parameter (Phase 1).

		Marginal Mean Bias			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	-0.011	0.020	-0.003	0.006
	1000	-0.001	0.009	-0.001	-0.001
Measurement Quality (*MQ*)	Low	-0.010	0.029	-0.002	-0.002
	High	-0.002	0.000	-0.001	0.007
Test Length (*J*)	1	-0.004	0.027	-0.003	0.001
	5	-0.008	0.003	0.000	0.004
Time Points (*T*)	5	0.007	0.024	-0.004	-0.002
	10	-0.019	0.006	0.000	0.007
Transition Probability (*TP*)	Low	0.005	0.022	-0.004	-0.002
	High	-0.017	0.008	0.000	0.007
Initial Mastery Probability (*IP*)	Low	0.010	0.016	-0.003	-0.001
	High	-0.022	0.014	-0.001	0.006

Table 10.

Marginal means for relative bias by experimental factor and parameter (Phase 1).

		Marginal Mean Relative Bias			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	-1.70%	8.74%	-0.38%	5.99%
	1000	1.14%	4.10%	-0.15%	0.57%
Measurement Quality (*MQ*)	Low	0.29%	12.71%	-0.36%	-0.53%
	High	-0.86%	0.13%	-0.16%	7.09%
Test Length (*J*)	1	1.97%	11.74%	-0.50%	3.97%
	5	-2.53%	1.09%	-0.02%	2.60%
Time Points (*T*)	5	5.00%	10.17%	-0.59%	1.44%
	10	-5.56%	2.66%	0.07%	5.12%
Transition Probability (*TP*)	Low	3.84%	10.89%	-0.60%	1.04%
	High	-4.40%	1.94%	0.08%	5.53%
Initial Mastery Probability (*IP*)	Low	4.96%	6.74%	-0.43%	1.94%
	High	-5.53%	6.09%	-0.09%	4.62%

Table 11.

Marginal means for RMSE by experimental factor and parameter (Phase 1).

		Marginal Mean RMSE			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	0.080	0.073	0.021	0.033
	1000	0.049	0.034	0.010	0.019
Measurement Quality (*MQ*)	Low	0.103	0.089	0.023	0.036
	High	0.026	0.018	0.009	0.016
Test Length (*J*)	1	0.082	0.077	0.020	0.033
	5	0.048	0.030	0.012	0.020
Time Points (*T*)	5	0.066	0.065	0.021	0.028
	10	0.063	0.042	0.010	0.024
Transition Probability (*TP*)	Low	0.061	0.051	0.018	0.024
	High	0.068	0.056	0.013	0.029
Initial Mastery Probability (*IP*)	Low	0.064	0.050	0.017	0.024
	High	0.065	0.056	0.015	0.028

Table 12.

Marginal means for Efficiency by experimental factor and parameter (Phase 1).

		Marginal Mean Efficiency			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	0.073	0.067	0.021	0.031
	1000	0.043	0.031	0.010	0.017
Measurement Quality (*MQ*)	Low	0.090	0.080	0.022	0.034
	High	0.026	0.018	0.009	0.014
Test Length (*J*)	1	0.070	0.069	0.019	0.029
	5	0.046	0.029	0.012	0.019
Time Points (*T*)	5	0.059	0.057	0.020	0.026
	10	0.057	0.041	0.010	0.022
Transition Probability (*TP*)	Low	0.056	0.044	0.017	0.022
	High	0.061	0.054	0.013	0.026
Initial Mastery Probability (*IP*)	Low	0.058	0.045	0.016	0.022
	High	0.059	0.053	0.014	0.026

Table 13.

Marginal means for Classification Accuracy (training set) by experimental factor and parameter (Phase 1).

Factor	Level	Marginal Mean Classification Accuracy (Training)
Sample Size (*N*)	200	93.66%
	1000	94.05%
Measurement Quality (*MQ*)	Low	89.04%
	High	98.66%
Test Length (*J*)	1	92.46%
	5	95.25%
Time Points (*T*)	5	90.36%
	10	97.35%
Transition Probability (*TP*)	Low	90.64%
	High	97.07%
Initial Mastery Probability (*IP*)	Low	93.08%
	High	94.63%

Table 14.

Marginal means for Classification Accuracy (validation set) by experimental factor and parameter (Phase 1).

Factor	Level	Marginal Mean Classification Accuracy (Validation)
Sample Size (*N*)	200	93.64%
	1000	94.05%
Measurement Quality (*MQ*)	Low	89.02%
	High	98.67%
Test Length (*J*)	1	92.46%
	5	95.23%
Time Points (*T*)	5	90.33%
	10	97.37%
Transition Probability (*TP*)	Low	90.62%
	High	97.07%
Initial Mastery Probability (*IP*)	Low	93.08%
	High	94.61%

Investigation of various "dummy coding" strategies. As covered in Chapter 3, implementing the desired "once a master, always a master" assumption within Netica required appending the response data matrix with two dummy variables, coded *M* for *Master,* then constraining the relationship between those dummy variables to be equal to the relationship between the latent proficiency variables at adjacent time points. This would, in theory supply ample information to the estimation of the skill regression parameter ($P(\theta_{t+1} = NM \mid \theta_t = M)$; intended to be equal to zero) while providing no information to the estimation of the transition probability. Examination of initial results, however revealed that this approach may have been contributing to unintended, positive bias in the estimation of $P(\theta_{t1} = M)$ under certain conditions. Furthermore, this effect seemed to increase with sample size. That this parameter was the only one affected was not necessarily surprising. The staticity constraints applied to the transition matrix and measurement model meant that information across all five or ten time points could be leveraged in estimating those parameters. For the initial probability of mastery, however, only information from the first time point was relevant to estimation. Under conditions when only one item was available at that time point (i.e., $J = 1$), only five time points were available in total (i.e., $T = 5$), and the quality of the information provided by the item was poor (i.e., *MQ* = Low), the estimation of $P(\theta_{t1} = M)$ proved problematic. The limited information available to estimate the initial probability of mastery allowed the information provided by the dummy variables to inflate the bias in the estimates. The specific mechanism through which this inflation occurred within the software remains unclear.

As bias induced by incidental characteristics of a study's design is obviously undesirable, further investigation was needed to arrive at a dummy coding approach that would (a) implement the desired parameter constraint while (b) eliminating (or limiting, at the least) unintended consequences (i.e., bias inflation). To this end, five dummy coding strategies were tested. All 64 of the Phase 1 conditions were run under each of the four strategies. The strategies were evaluated based on (a) the mean estimate of the skill regression parameter across all conditions where $J = 1$, $T = 5$, and MQ = Low (a value of zero being ideal) and (b) the difference in mean absolute estimation bias for the initial probability of mastery between the $N = 200$ and $N = 1{,}000$ conditions when $J = 1$, $T = 5$, and MQ = Low. Both criteria were marginalized across the *TP* and *IP* conditions. The latter criterion was chosen under that assumption that increasing sample size should yield bias closer to zero or, at the least, should not impact bias. A positive difference in bias when $N = 1{,}000$ and $N = 200$, then, was considered to be undesirable.

Dummy coding strategies. Five strategies were considered. The first strategy, Dummy Variable – *N* (DV-*N*) represented the initial strategy wherein two dummy variables were added to the response data. The *N*, here, suggests that the amount of information added by the dummy variables was proportional to the sample size (i.e., each of the *N* cases received values for each of the dummy variables). The next three strategies eschewed the dummy variable approach in favor of a dummy case approach. Under this method, a certain number of dummy cases (rows) were added to the data set in addition to the dummy variables (columns). Each dummy case received the typical *M* code for the dummy variables. The item responses for the dummy cases were coded as missing (Netica uses an asterisk to represent missing data) as were the values on the dummy

variables for the "real" data. These three new strategies were defined by the amount of information added in the form of dummy cases. Either $N_{Dummy} = 1$ (DC-1), $N_{Dummy} = 100$ (DC-100), or $N_{Dummy} = N$ (DC-*N*) cases were added. Finally, a strategy with no dummy variables was included as a point of reference. Tables 15 and 16 provide a comparison of the DV-*N* and DC-*N* approaches via small, hypothetical data sets.

Table 15.

Example data set using the DV-N approach where N = 5.

ID	$X_{1, t1}$	$X_{1, t2}$	$X_{1, t3}$	$X_{1, t4}$	$X_{1, t5}$	$Dummy_1$	$Dummy_2$
1	0	0	1	1	0	*M*	*M*
2	1	0	1	0	1	*M*	*M*
3	0	1	1	0	0	*M*	*M*
4	1	1	0	1	1	*M*	*M*
5	1	1	1	1	1	*M*	*M*

Table 16.

*Example data set using the DC-N approach where N = 5; * indicates a missing value.*

ID	$X_{1, t1}$	$X_{1, t2}$	$X_{1, t3}$	$X_{1, t4}$	$X_{1, t5}$	$Dummy_1$	$Dummy_2$
1	0	0	1	1	0	*	*
2	1	0	1	0	1	*	*
3	0	1	1	0	0	*	*
4	1	1	0	1	1	*	*
5	1	1	1	1	1	*	*
D1	*	*	*	*	*	*M*	*M*
D2	*	*	*	*	*	*M*	*M*
D3	*	*	*	*	*	*M*	*M*
D4	*	*	*	*	*	*M*	*M*
D5	*	*	*	*	*	*M*	*M*

Results and conclusions. Table 17 presents the mean skill regression parameter estimate when $J = 1$ and *MQ* = Low as well as the mean absolute estimation bias for the

initial probability of mastery parameter where $N = 200$ and $N = 1{,}000$ when $J = 1$, $T = 5$, and MQ = Low across the five dummy coding strategies.

Table 17.

Comparison of dummy coding strategy outcomes.

	DV-*N*	DC-1	DC-100	DC-*N*	NoDummy
Mean Skill Regression Estimate	0.000	0.192	0.000	0.000	0.280
Mean Absolute Bias in $P(\theta_{t1} = M)$ - ($N = 200$)	0.069	0.091	0.069	0.068	0.065
Mean Absolute Bias in $P(\theta_{t1} = M)$ - ($N = 1000$)	0.111	0.082	0.070	0.102	0.069

Note. Values are marginalized over *TP* and *IP* while $J = 1$, $T = 5$, and MQ = Low.

From these results we can see that, while the original DV-*N* approach did implement the desired parameter constraint, it also inflated bias as sample size increased. The same conclusion holds for the DC-*N* approach which did not appear to be appreciably different from the DV-*N* method. Though the DC-1 approach yielded bias values more in line with expectations (i.e., bias decreased with an increase in sample size), it did not impart enough information to constrain the skill regression parameter to a value near zero. The best compromise of the evaluation criteria was represented by the DC-100 approach. Under this approach, the skill regression parameter was negligibly different from zero across all conditions. Additionally, there appeared to be a minimal amount of bias inflation present using this approach. Based on these results, the DC-100 approach was used for all conditions in both Phase 1 and Phase 2 in the current study. Secondarily, these results also illustrate the negligible impact of sample size on estimation bias for the initial probability of mastery even without the use of any dummy coding strategy.

Impact of measurement quality (*MQ*). As can be seen in Tables 9-14, measurement quality had the most dramatic effect of the quality of parameter recovery across all indices. When measurement quality was "High," parameter recovery was satisfactory across all other conditions. "Low" measurement quality resulted in far less desirable parameter recovery particularly as it pertains to the initial probability of mastery and the transition probability. This impact is perhaps most readily apparent in the classification accuracy values (Table 13 and Table 14). Figure 10 presents the classification accuracy results graphically for conditions where $N = 200$. In this plot, classification accuracy is shown on the Y-axis while measurement quality is plotted along the X-axis. The panels are separated by the four combinations of test length (*J*) and the number of measurement occasions (*T*). The individual lines being plotted are indicated by the four combinations of the true value for the initial probability of mastery (*IP*) and the transition probability (*TP*). The $N = 200$ condition plot is presented instead of the $N = 1{,}000$ condition plot as the former represents a worst-case scenario from a sample size perspective within the confines of the study design. We can see the large, positive slope of the line plots going from "LowQ" to "HighQ." Furthermore, we can see that, when measurement quality is high, the classification accuracy values are very close to 100% regardless of the combination of the other design facets.

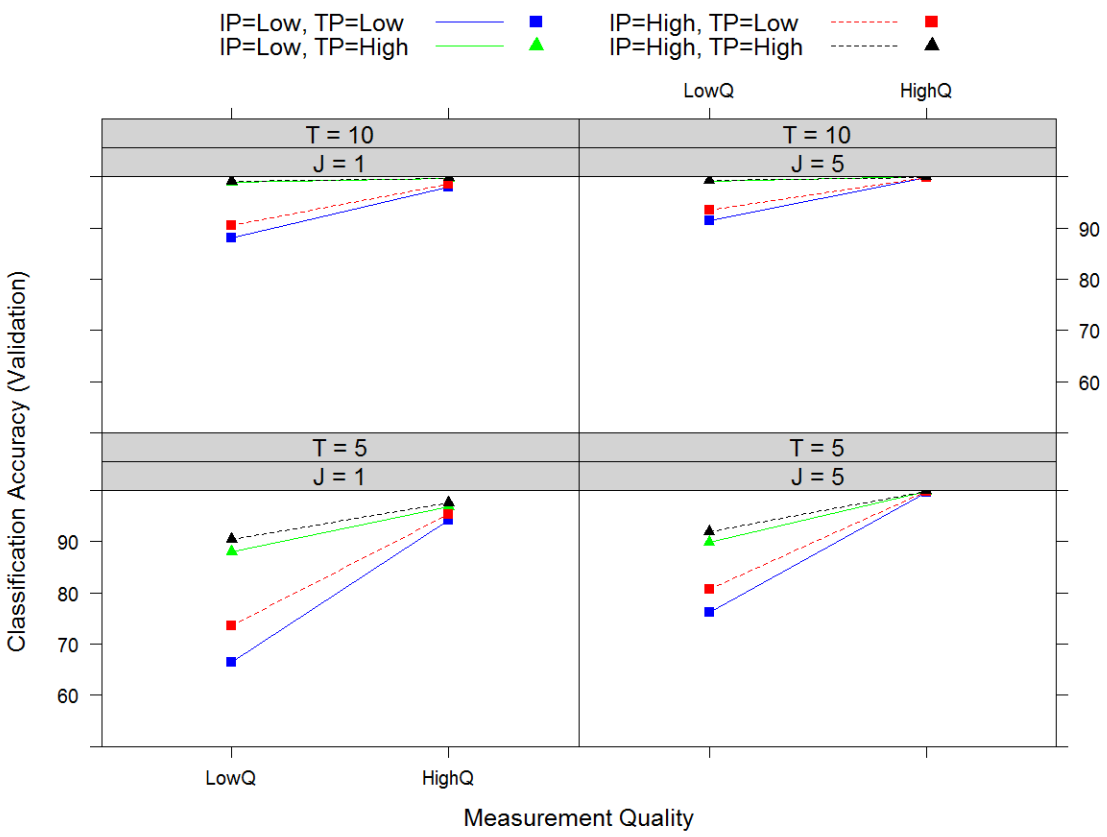

Figure 10. Classification accuracy (validation) when N = 200 (Phase 1).

Figure 11 and Figure 12 present the same type of plot with bias substituted for classification accuracy and results presented for the recovery of $P(\theta_{t1} = M)$ and $P(\theta_{t+1} = M \,|\, \theta_t = NM)$, respectively. Again, we can see the line plots converge very close to zero (indicated by the light grey, horizontal line) when measurement quality is high. Finally, Figures 13 and 14 present the RMSE values for these same two parameters when N = 200. Note that the RMSE values when measurement quality is high are approximately equal to the estimation efficiency given the decomposition of RMSE (Mood, Graybill, & Boes, 1974) and that bias is essentially zero under these conditions.

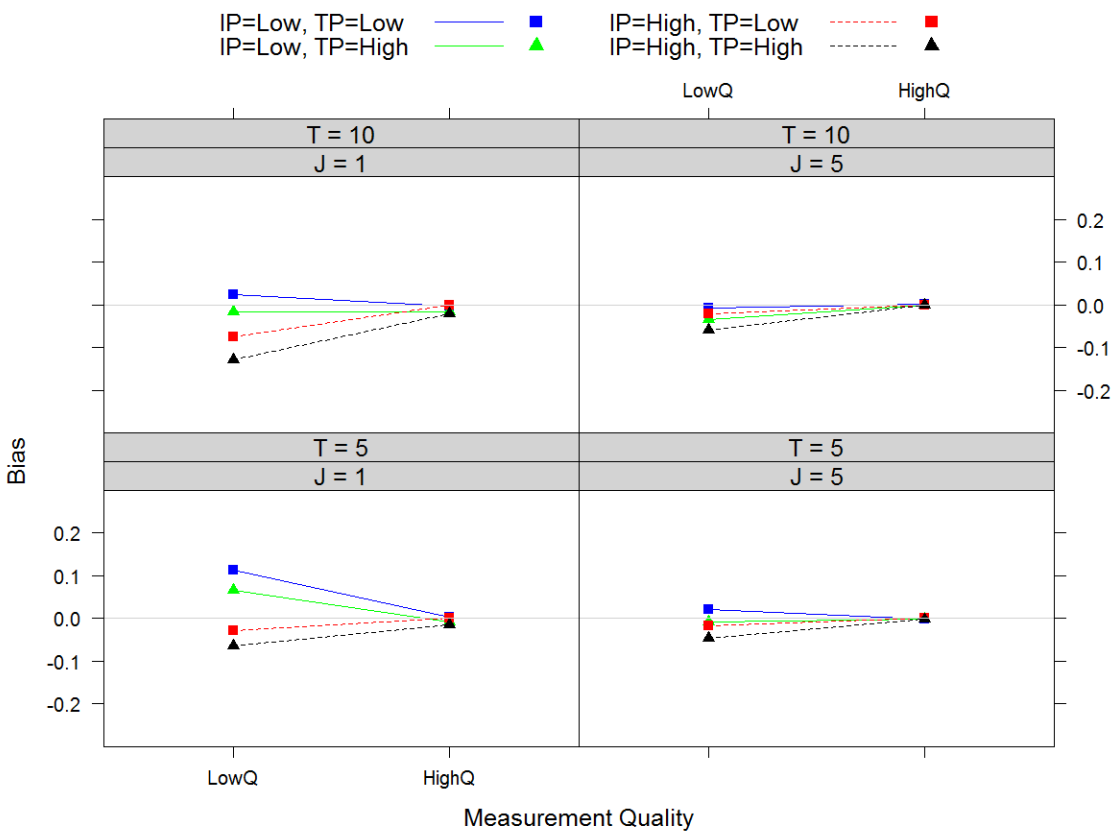

Figure 11. Bias in the initial probability of mastery parameter when N = 200 (Phase 1).

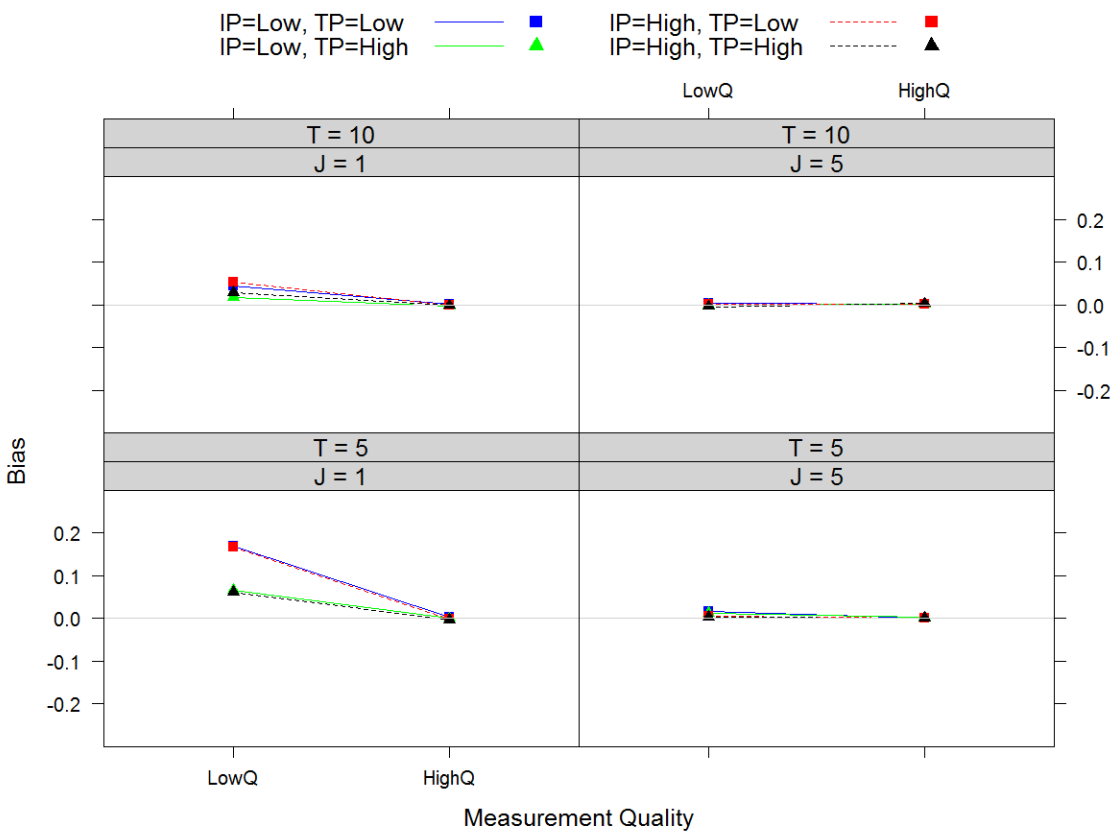

Figure 12. Bias in the transition probability parameter when N = 200 (Phase 1).

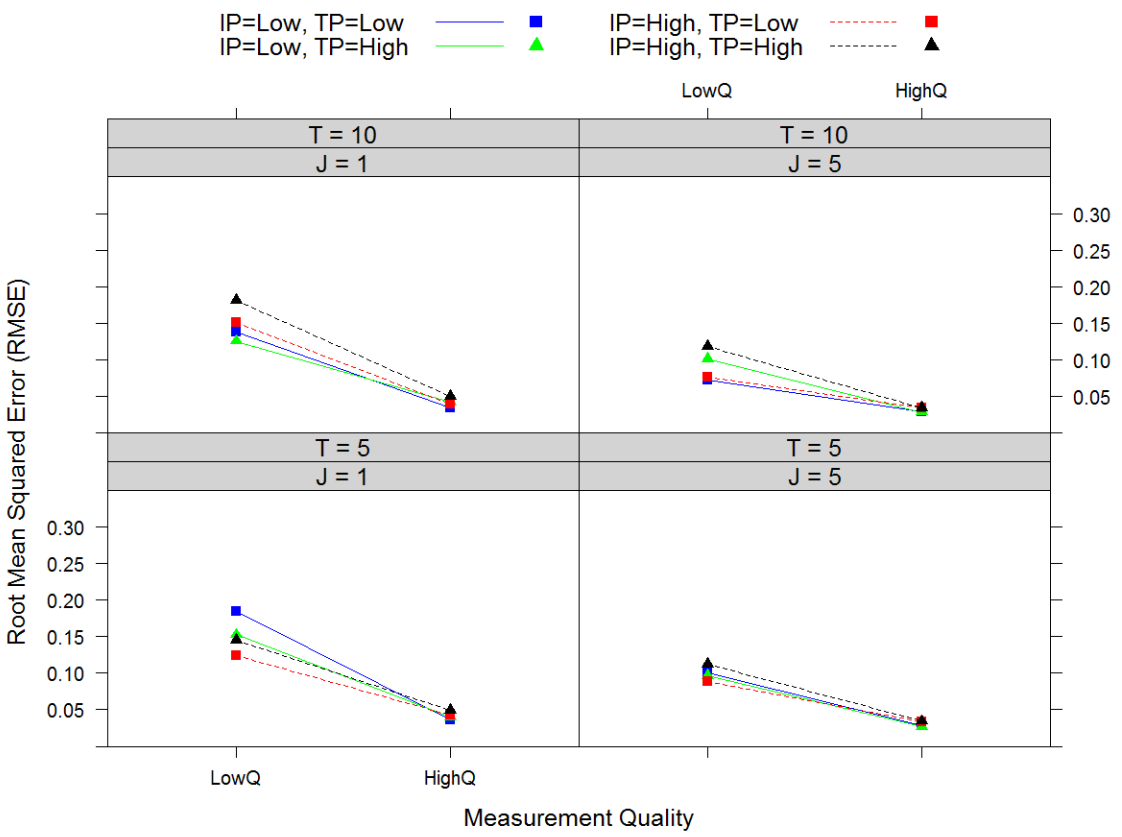

Figure 13. RMSE in the initial probability of mastery parameter when N = 200 (Phase 1).

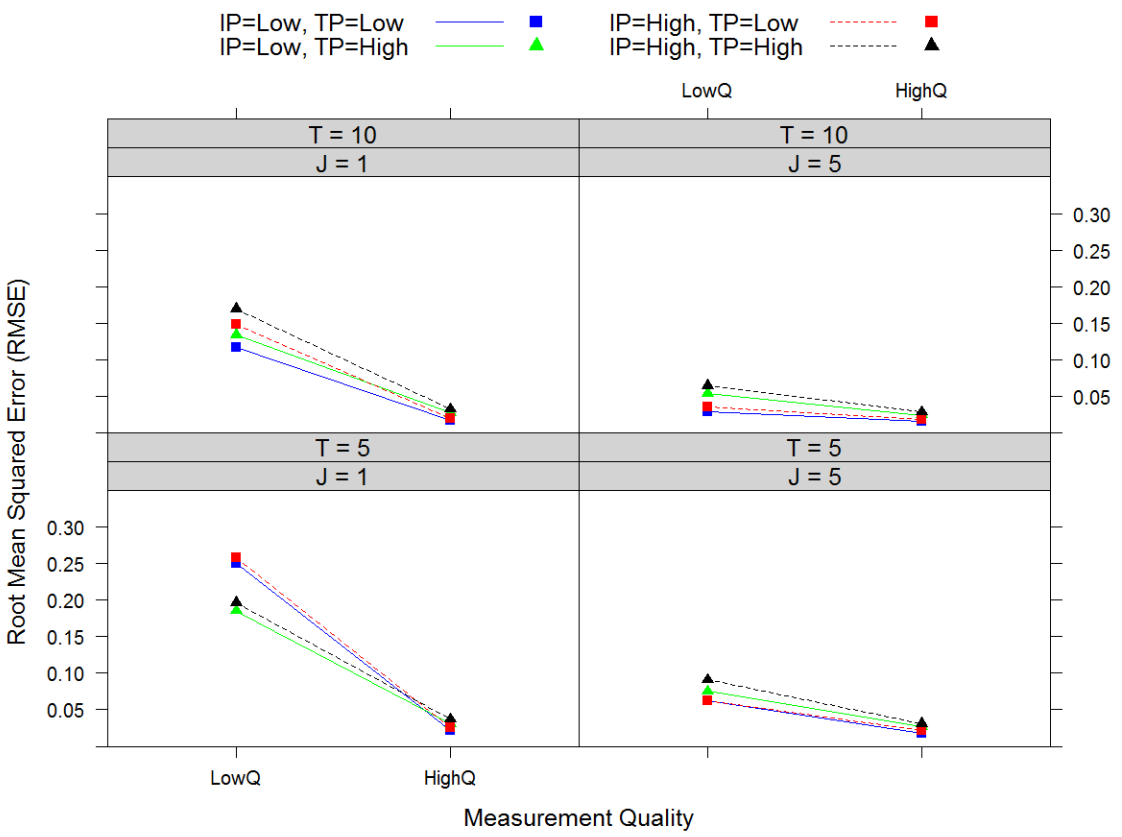

Figure 14. RMSE in the transition probability parameter when N = 200 (Phase 1).

Impact of sample size (*N*). Two sample size conditions were included in Phase 1 – N = 200 and N = 1,000. In general, increasing sample size resulted in better parameter recovery (higher classification accuracy, less bias, etc.). Figure 15 demonstrates the impact of increasing sample size from 200 to 1,000 on the RMSE of the estimates for $P(\theta_{t1} = M)$ (note that this effect is moderated by the amount of estimation bias present). As one would expect, increasing sample sizes yields more efficient estimates which, in turn yields lower RMSE.

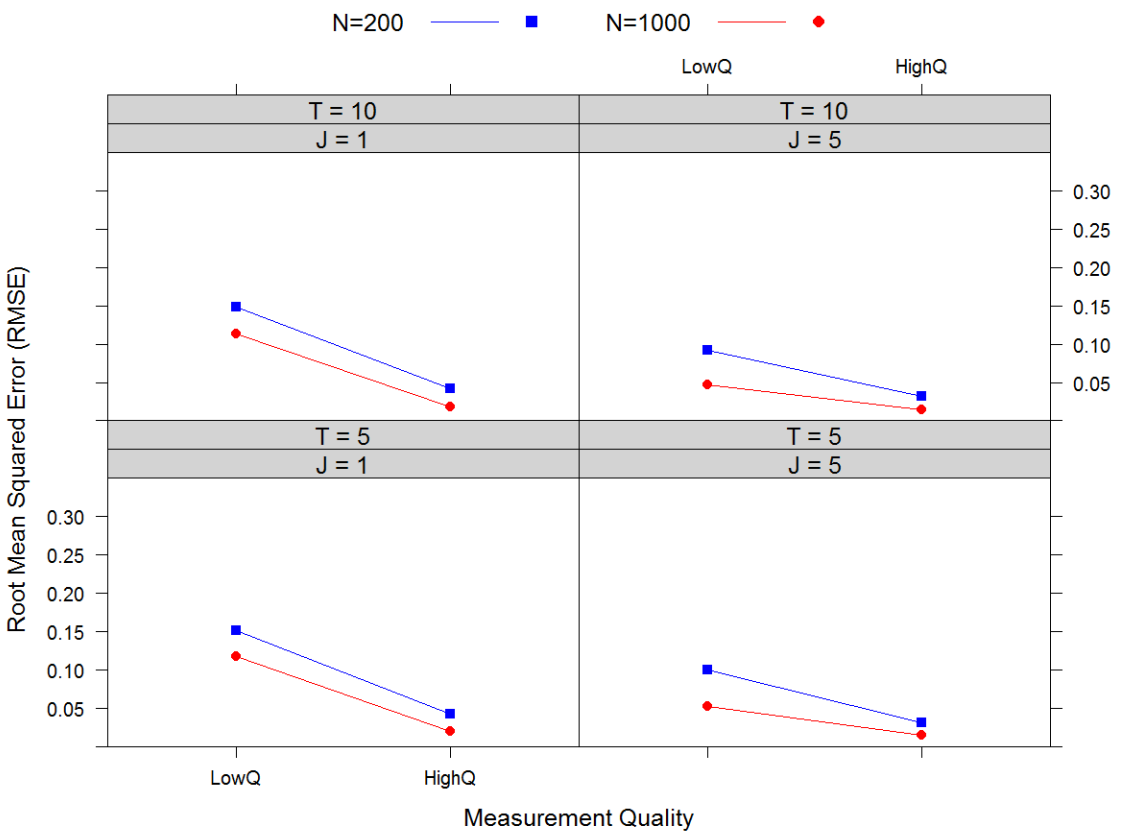

Figure 15. Impact of sample size on RMSE for the initial probability of mastery parameter (Phase 1).

Figures 16-19 allow for the comparison of the impact of sample size as a function of all the other design facets. Figure 16 (a repeat of Table 11 above presented here for ease of comparison) and Figure 17 show bias for $P(\theta_{t1} = M)$ when N = 200 and N = 1,000, respectively while Figure 18 (a repeat of Table 10 above) and Figure 19 show classification accuracy under the same conditions. From these figures we can see that increasing sample size yields lower bias in all conditions save for when J = 1, T = 5, and MQ = Low (as addressed in the prior section on dummy coding strategies) and higher classification accuracy, though perhaps marginally so, across all conditions.

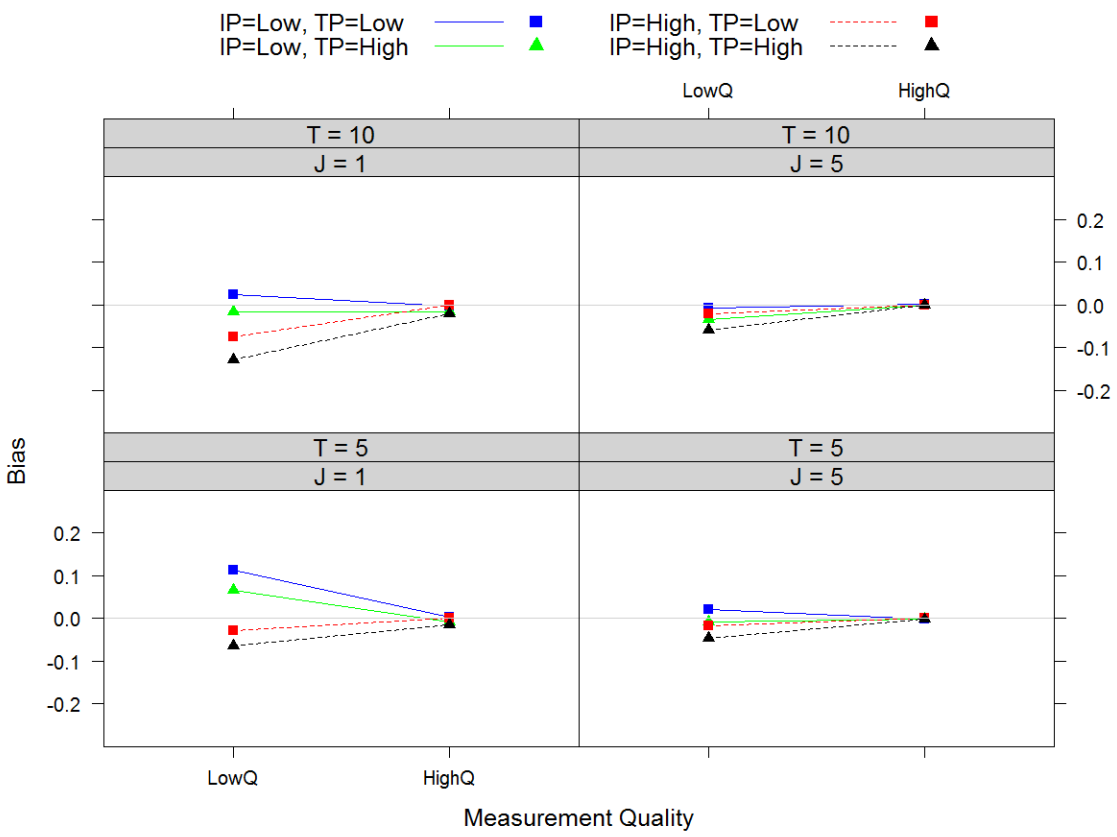

Figure 16. Bias in the initial probability of mastery parameter when N = 200 (Phase 1).

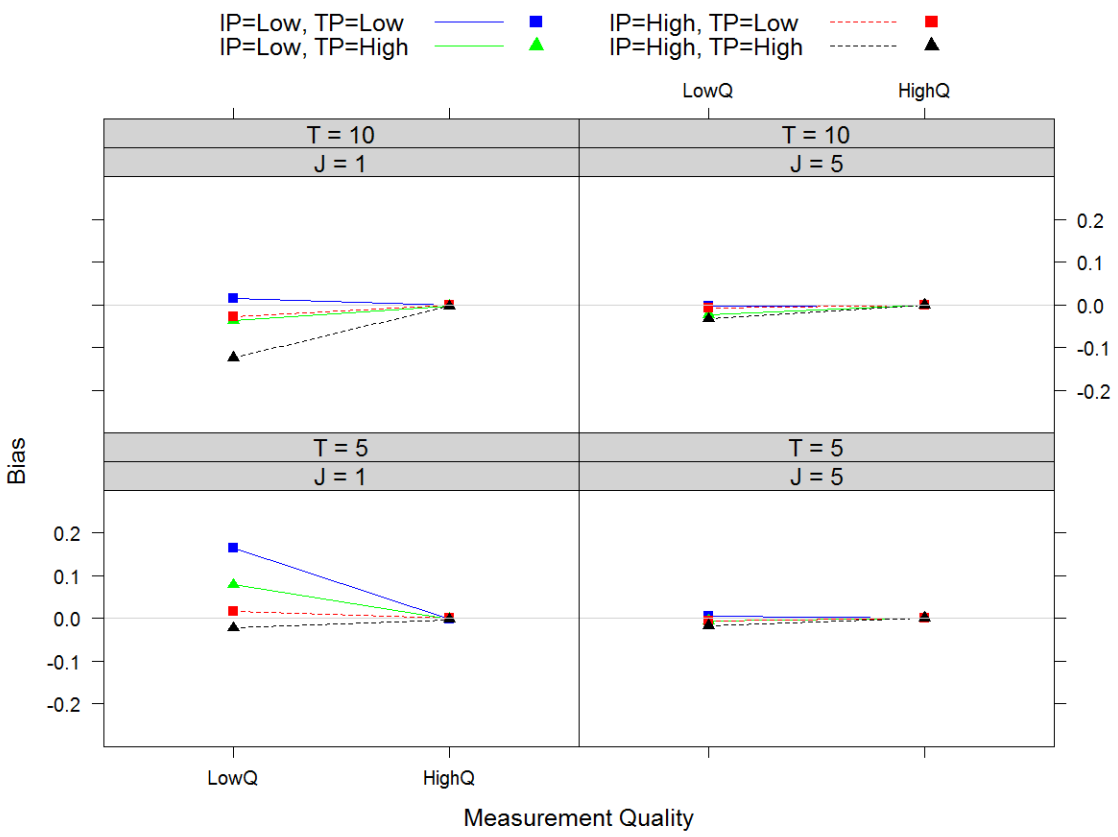

Figure 17. Bias in the initial probability of mastery parameter when N = 1,000 (Phase 1).

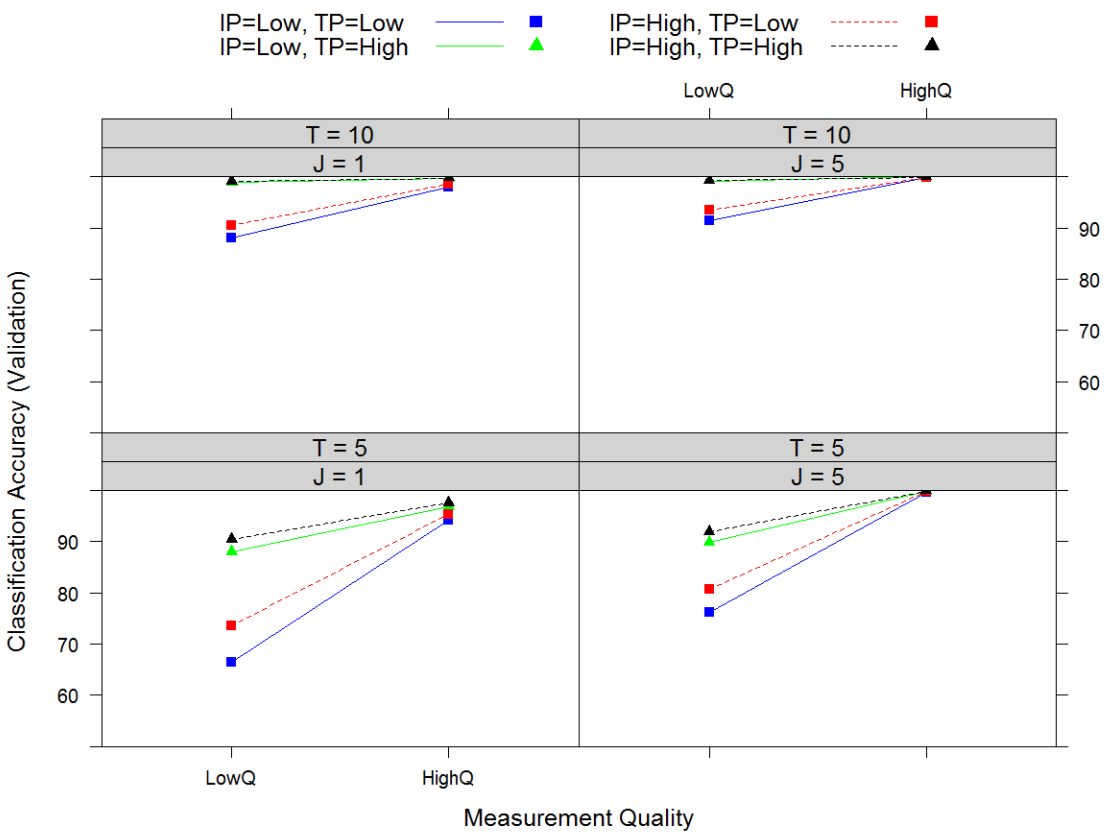

Figure 18. Classification accuracy (validation) when N = 200 (Phase 1).

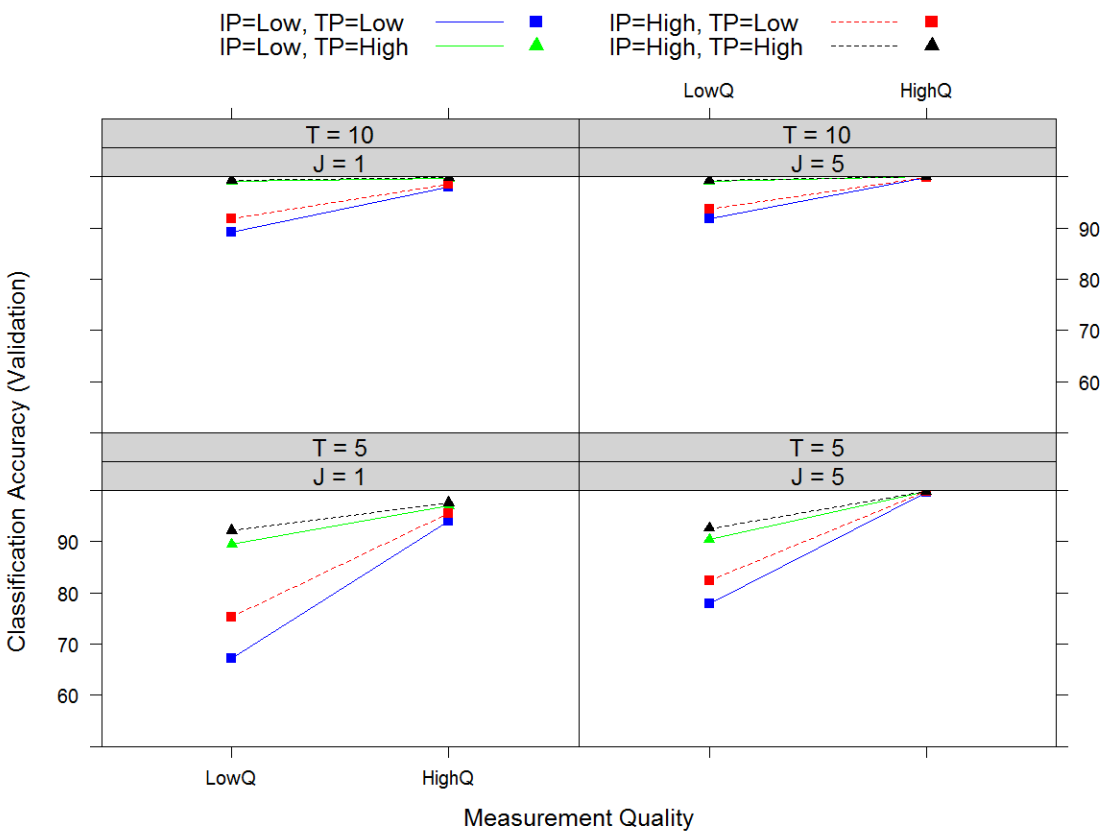

Figure 19. Classification accuracy (validation) when N = 1,000 (Phase 1).

Impact of test length (*J*). Somewhat contrary to expectation, the effect of test length on parameter recovery for the measurement model parameters was somewhat minimal, though parameter recovery for the measurement model parameters was generally quite good across all design facets (Tables 9-14). As with other design facets, effect of increasing J was most notable on the recovery of the initial probability of mastery and transition probability parameters. As previously noted, conditions where J = 1 proved quite problematic when combined with fewer measurement occasions (i.e., T = 5) and/or poor measurement quality (i.e., MQ = Low). As presented in Figure 11 and Figure 12, increasing J from 1 to 5 yielded negligible estimation bias for $P(\theta_{t1} = M)$ and

$P(\theta_{t+1} = M \mid \theta_t = NM)$, respectively even when *N* and *MQ* were at their lowest levels – suggesting that even a relatively short "test" consisting of only five tasks may be sufficient for overcoming poor measurement quality. Finally, increasing *J* led to much improved estimation efficiency (i.e., more stable estimates) for the aforementioned two parameters. This effect was most notable in larger samples (i.e., *N* = 1,000) and when measurement quality was low (Figure 20 and Figure 21).

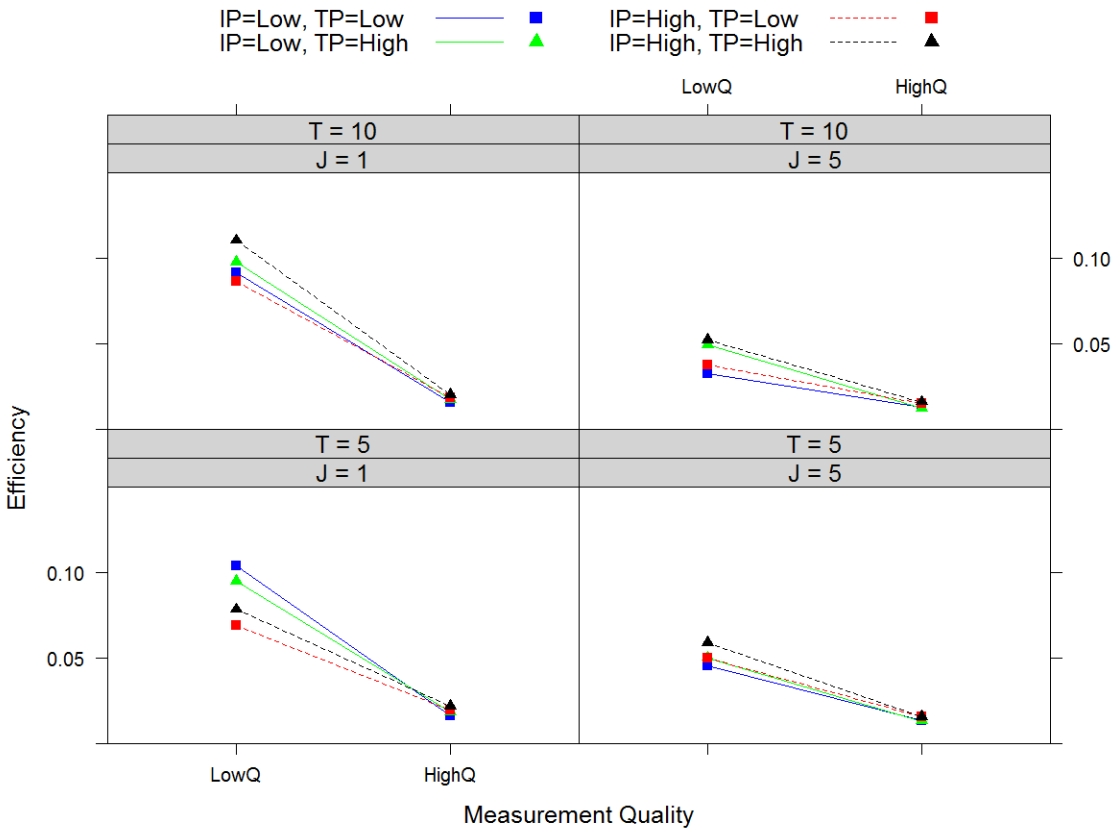

Figure 20. Estimation efficiency for the initial probability of mastery when *N* = 1,000 (Phase 1).

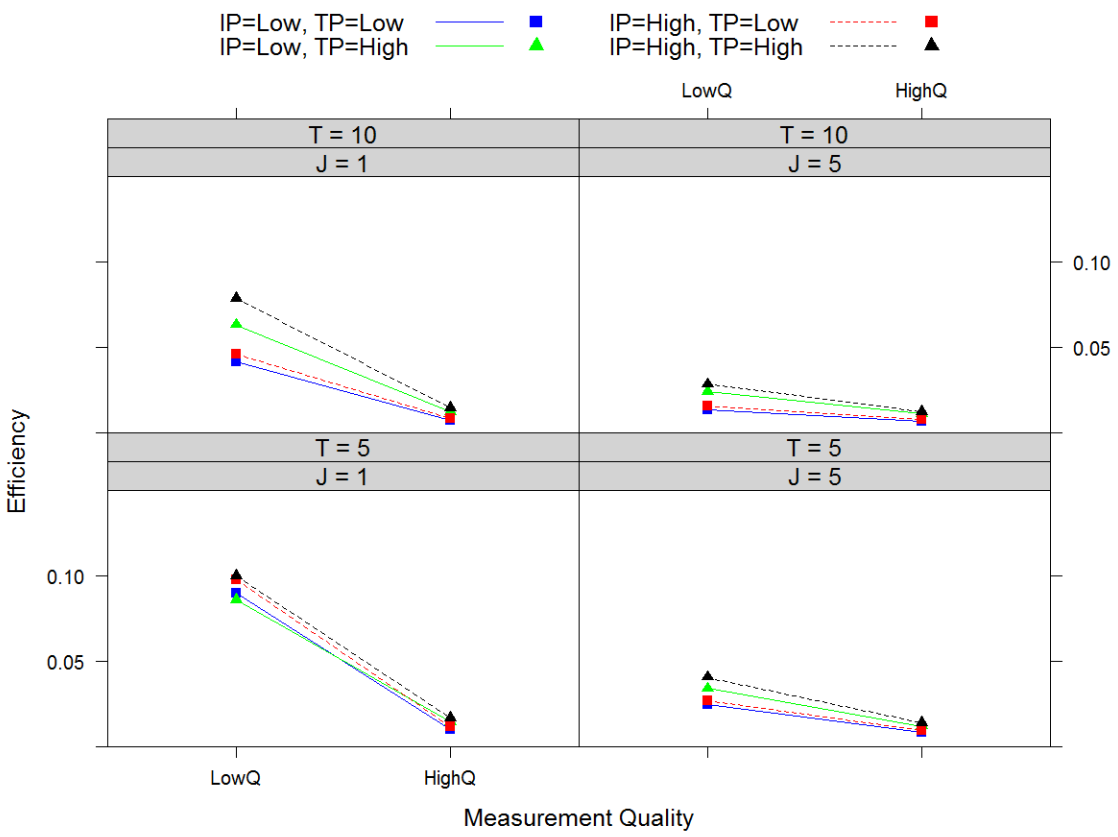

Figure 21. Estimation efficiency for the transition probability when N = 1,000 (Phase 1).

Impact of number of measurement occasions (*T*); true values for transition (*TP*)/initial mastery (*IP*) probabilities. As can be seen in Table 14 and Figure 10, the number of measurement occasions and the true value for the transition probability had a large impact on the ability of a model to correctly classify participants. This effect was most likely due, in part to the presence of the "once a master, always a master" assumption that was applied in the current work. Given this assumption, all participants would eventually reach mastery. Increasing the transition probability, then would serve to hasten this state being reached. It stands to reason, then, that there exists a combination of *T* and *TP* for which both the model and the true status of the participants would suggest the state where all participants are classified as having mastered the content. Any

combination of *T* and *TP* beyond this threshold would results in 100% classification accuracy given that participants do not regress from mastery to non-mastery.

Beyond impacts on classification accuracy, the *TP* = High condition also tended to yield less bias (in magnitude) in the estimation of the transition probability than the *TP* = Low condition (Figure 12). This effect was moderated by *T* in that almost no bias existed in the estimation of that parameter when *T* = 10 even when *MQ* = Low.

The true value of the initial probability of mastery (*IP*) had a reasonably large impact on the estimation of the initial probability of mastery in terms of bias, as might be expected. *IP* = Low tended to yield an overestimation of the parameter while *IP* = High tended to yield underestimation (Figure 11). This might suggest that there is some value between 0.20 and 0.40 which the estimation routine tended to favor holding all other design facets static.

Finally, all three of these facets demonstrated an ability to cause a suppression effect in the estimation of the initial probability of mastery. As can be seen in Figure 11, increasing *T*, increasing *TP*, or increasing *IP* tended to shift the estimates of $P(\theta_{t1} = M)$ downward.

Phase 2 Conditions

Based on the results of Phase 1, several conditions were added/removed for Phase 2. Similar to what Table 7 did for Phase 1, Table 18 presents a summary of the condition values for Phase 2. These changes yielded a total of 184 new experimental conditions to be tested in Phase 2. The results presented in the next section will also incorporate the

results from the 32 conditions from Phase 1 which apply to the new condition matrix for Phase 2.

Table 18.

Model condition values for study design (Phase 2).

	Value(s)
N	{200, 400, 1000}
$P(X_{j,t}=1\|\theta_t=M)$	{"Low" = 0.60, "Med" = 0.75}
$P(X_{j,t}=1\|\theta_t=NM)$	{"Low" = 0.40, "Med" = 0.25}
$P(\theta_{t1})$	{"Low" = 0.20, "High" = 0.40}
$P(\theta_{t2}=M\|\theta_{t1}=NM)$	{"Low" = 0.20, “Med” = 0.40, "High" = 0.80}
J (per t)	{1, 3, 5}
T	{5, 10}

Note. NM = non-master; M = master.

Given the generally excellent parameter recovery performance across all other conditions when MQ = High, that condition was dropped in favor of a more moderate level of measurement quality (MQ = Med). Given the fairly large difference between parameter recovery between the $J = 1/J = 5$ and $N = 200/N = 1{,}000$ conditions, intermediate $J = 3$ and $N = 400$ conditions were added. The value for the latter condition was arrived upon by choosing the midpoint in the standard error of the mean between $N = 200$ and $N = 1{,}000$ using the equation

$$N^* = \left(\frac{1}{\left(\frac{1}{\sqrt{200}} + \frac{1}{\sqrt{1000}} \right) / 2} \right)^2 = 381.966, \tag{13}$$

where N^* is the target sample size, then rounding to the hundreds place. Finally, the TP = High condition ($TP = 0.40$) was recoded as TP = Med and a new TP = High condition

where the true value of the transition probability is equal to 0.80 was added. The purpose of this new condition was two-fold. First, it was noted in Phase 1 that higher true values of *TP* yielded less bias. This new condition would allow for the determination of whether that trend would continue (i.e., the largest possible true *TP* would be ideal in practice) or whether some other trend would emerge (e.g., values close to 0.50 are preferred). Second, the author is not aware any existing methodological work that has used very large, perhaps unreasonably large values for the transition probability. Including this new condition allows for a theoretical exploration of considerations when assessing skills that can be acquired at an uncannily fast rate.

Phase 2 Results

Many of the effects noted in Phase 1 held for Phase 2. As such, the presentation of Phase 2 results will be focused on the extent to which the newly introduced conditions (*MQ* = Med, *N* = 400, and *J* = 3) resulted in sufficient parameter recovery as compared to their more favorable (and extreme) counterparts (*MQ* = High, *N* = 1,000, and *J* = 5). There will also be a presentation of the results obtained under conditions consisting of a very large transition probability.

Impact of a large transition probability. As was mentioned above, an extreme transition probability condition (*TP* = 0.80) was added primarily for exploratory purposes. This condition had a drastic impact of the quality of parameter estimation. Figure 22 and Figure 23 show raw bias when *N* = 200 for $P(\theta_{t1} = M)$ and $P(X = 1 | \theta = NM)$, respectively while Figure 24 and Figure 25 show the estimation efficiency for $P(\theta_{t1} = M)$ when *N* = 200 and *N* = 1,000, respectively.

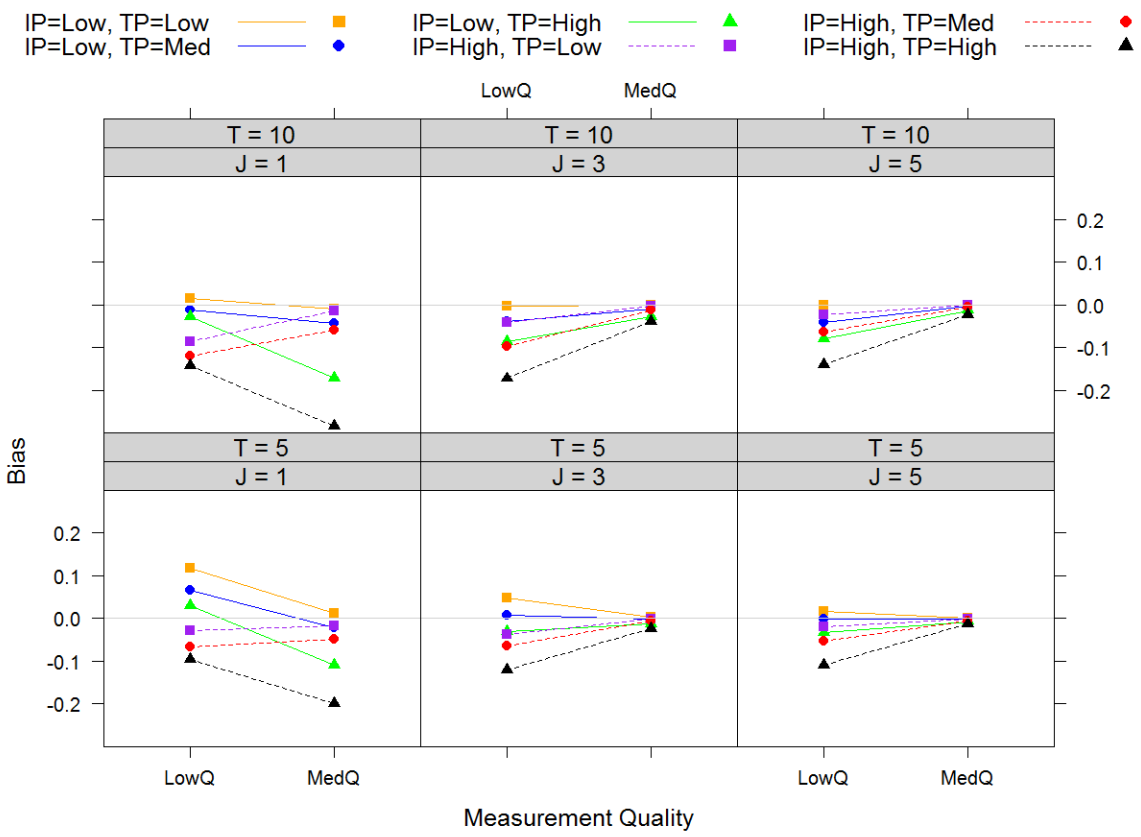

Figure 22. Bias in the initial probability of mastery parameter when N = 200 (Phase 2).

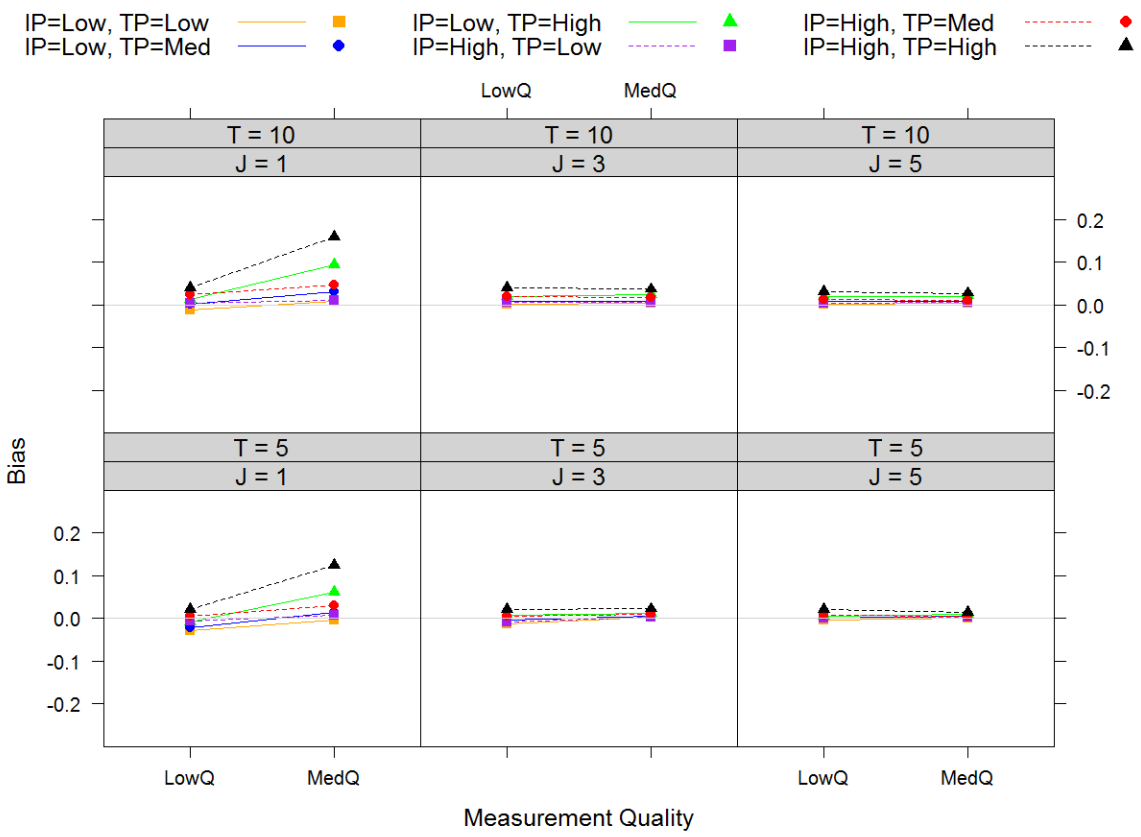

Figure 23. Bias in the probability of a correct response for a non-master when $N = 200$ (Phase 2).

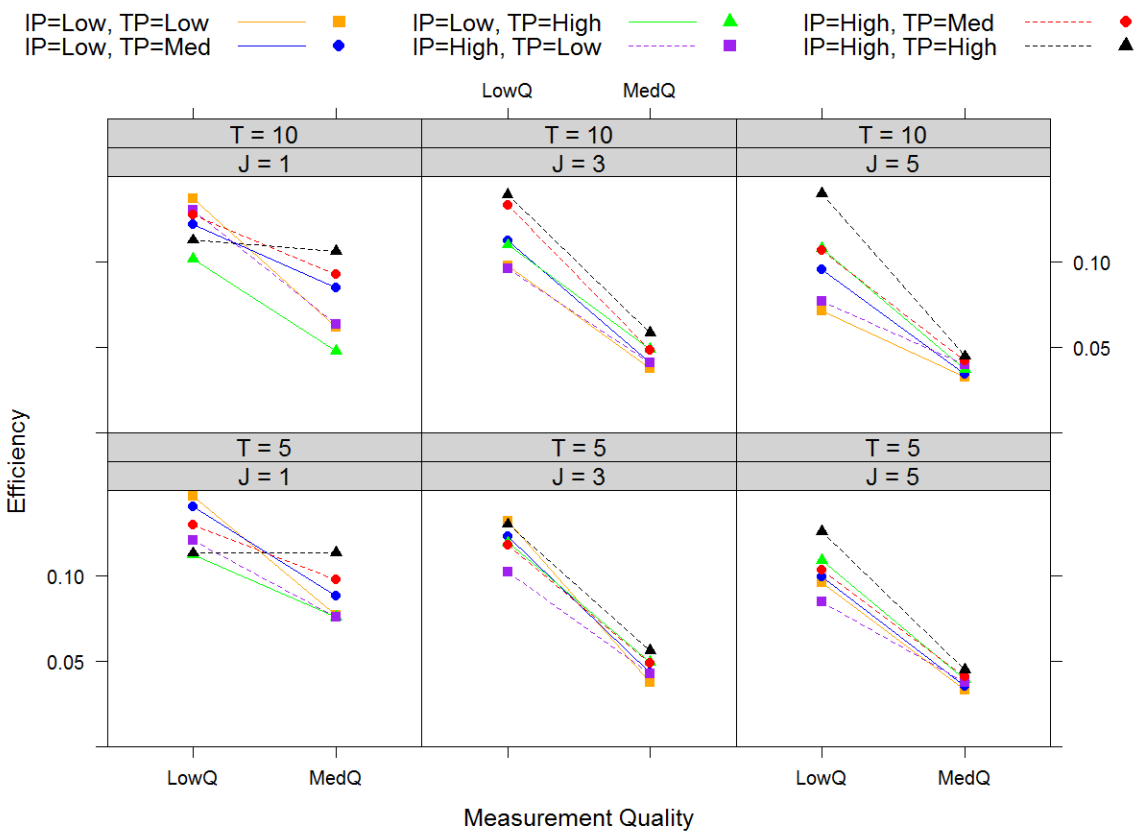

Figure 24. Estimation efficiency for the initial probability of mastery when $N = 200$ (Phase 2).

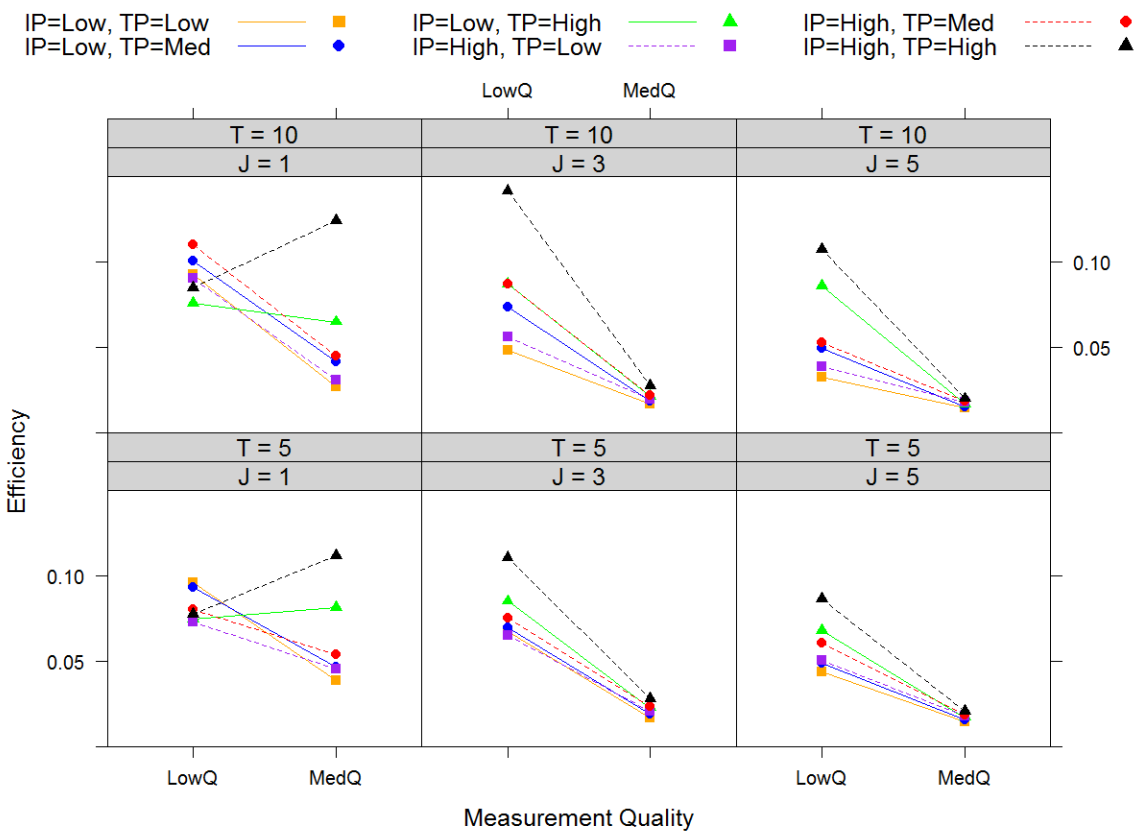

Figure 25. Estimation efficiency for the initial probability of mastery when N = 1,000 (Phase 2).

From these plots we can see that, for a select set of conditions (when J = 1), increasing measurement quality from *MQ* = Low to *MQ* = Med resulted in more biased estimates (in the absolute sense) when *TP* = High (represented by the green and black lines). Though Figure 22 and Figure 23 show only the N = 200 conditions, the same effect was presented when N = 400 or N = 1,000. Furthermore, increasing sample size resulted in poorer estimation efficiency when *TP* = High. Surprisingly, recovery of the transition probability itself was excellent under the same conditions (Figure 26). It would seem, however that some inflation (in the probability of a correct response for non-masters) or deflation (in the initial probability of mastery) was necessary to accommodate such a large transition

probability. This could be due to the fact that a high transition probability would result in all participants achieving mastery relatively quickly – leaving very few non-master cases with which to form an estimate of $P(X=1|\theta=NM)$. There could also be some balancing effect taking place wherein decreasing the initial proportion of mastery and increasing the probability of a correct response for non-masters (thus making them more similar to masters) was necessary to reconcile the very high level of mastery represented in the observed item responses. In either case, the remaining presentation of Phase 2 results will largely ignore the *TP* = High condition given the unexpected nature of the results as well as the low likelihood of such a large transition probability being observed in practice.

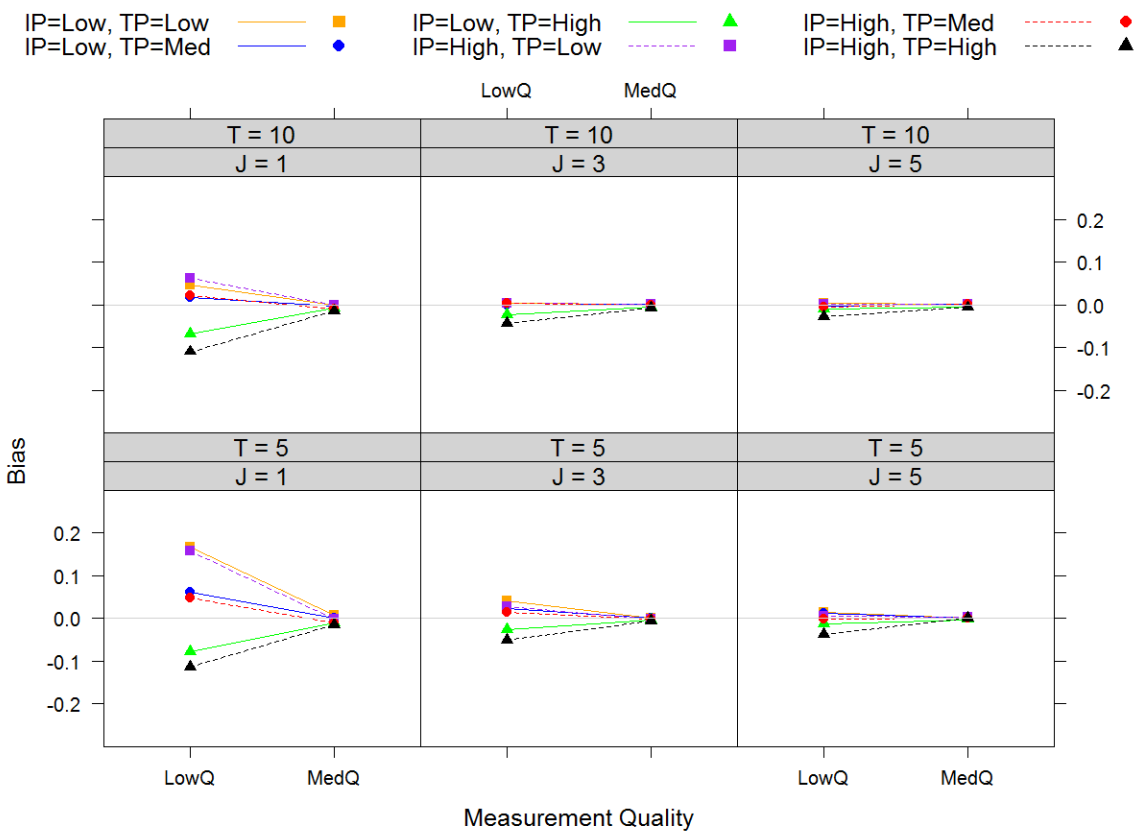

Figure 26. Bias in the transition probability estimates when N = 200 (Phase 2).

Sufficiency of Medium measurement quality. Tables 19-23 present the main effects for bias, relative bias, RMSE, efficiency, and classification accuracy, respectively across the four parameters of interest. These tables do not include the conditions where *TP* = High for the reasons listed in the previous section. The disaggregated results (including the *TP* = High condition) for Phase 2 are presented in Appendix B. From these Tables 19-23 we can see that mean estimation bias, marginalized over all other dimensions, is at or near zero for all four parameters when *MQ* = Med while the marginal mean for classification accuracy is quite high at 95.53%.

Table 19.

Marginal means for bias by experimental factor and parameter (Phase 2; TP = High removed).

		Marginal Mean Bias			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	-0.016	0.015	-0.001	0.005
	400	-0.007	0.010	-0.001	0.002
	1000	-0.005	0.008	0.000	0.000
Measurement Quality (*MQ*)	Low	-0.013	0.022	-0.001	-0.001
	Med	-0.006	0.000	0.000	0.006
Test Length (*J*)	1	-0.008	0.026	-0.002	0.001
	3	-0.012	0.005	0.000	0.003
	5	-0.009	0.002	0.000	0.003
Time Points (*T*)	5	0.003	0.018	-0.002	-0.002
	10	-0.022	0.004	0.001	0.006
Transition Probability (*TP*)	Low	0.003	0.017	-0.003	-0.002
	Med	-0.022	0.005	0.001	0.007
Initial Mastery Probability (*IP*)	Low	0.003	0.009	-0.001	0.001
	High	-0.025	0.010	0.000	0.005

Table 20.

Marginal means for relative bias by experimental factor and parameter (Phase 2; TP = High removed).

		Marginal Mean Relative Bias			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	-3.58%	6.61%	-0.20%	2.07%
	400	-0.87%	4.51%	-0.15%	0.81%
	1000	-0.66%	3.39%	-0.02%	0.14%
Measurement Quality (*MQ*)	Low	-1.42%	9.47%	-0.24%	-0.23%
	Med	-1.99%	0.20%	-2.68E-05	2.24%
Test Length (*J*)	1	0.67%	11.47%	-0.35%	1.07%
	3	-3.19%	2.06%	-0.04%	1.03%
	5	-2.60%	0.98%	0.02%	0.91%
Time Points (*T*)	5	3.34%	7.69%	-0.40%	-0.11%
	10	-6.74%	1.98%	0.15%	2.12%
Transition Probability (*TP*)	Low	2.79%	8.43%	-0.42%	-0.35%
	Med	-6.20%	1.24%	0.18%	2.36%
Initial Mastery Probability (*IP*)	Low	1.40%	4.58%	-0.21%	0.84%
	High	-6.15%	4.40%	0.03%	1.87%

Table 21.

Marginal means for RMSE by experimental factor and parameter (Phase 2; TP = High removed).

		Marginal Mean RMSE			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	0.090	0.072	0.024	0.039
	400	0.071	0.051	0.018	0.030
	1000	0.053	0.033	0.012	0.021
Measurement Quality (*MQ*)	Low	0.102	0.080	0.022	0.035
	Med	0.040	0.024	0.014	0.024
Test Length (*J*)	1	0.099	0.085	0.024	0.041
	3	0.065	0.041	0.016	0.026
	5	0.050	0.030	0.014	0.022
Time Points (*T*)	5	0.073	0.064	0.024	0.033
	10	0.070	0.040	0.012	0.027
Transition Probability (*TP*)	Low	0.065	0.047	0.020	0.027
	Med	0.077	0.057	0.015	0.033
Initial Mastery Probability (*IP*)	Low	0.073	0.054	0.019	0.029
	High	0.073	0.055	0.017	0.033

Table 22.

Marginal means for Efficiency by experimental factor and parameter (Phase 2; TP = High removed).

		Marginal Mean Efficiency			
Factor	Level	$P(\theta_{t1})$	$P(\theta_{t+1} = M \mid \theta_t = NM)$	$P(X = 1 \mid \theta = M)$	$P(X = 1 \mid \theta = NM)$
Sample Size (*N*)	200	0.083	0.068	0.024	0.037
	400	0.066	0.048	0.018	0.029
	1000	0.047	0.031	0.011	0.020
Measurement Quality (*MQ*)	Low	0.092	0.075	0.021	0.034
	Med	0.039	0.024	0.014	0.023
Test Length (*J*)	1	0.087	0.077	0.023	0.038
	3	0.061	0.041	0.016	0.026
	5	0.048	0.030	0.014	0.022
Time Points (*T*)	5	0.067	0.059	0.023	0.031
	10	0.064	0.040	0.012	0.026
Transition Probability (*TP*)	Low	0.061	0.042	0.020	0.026
	Med	0.070	0.056	0.015	0.031
Initial Mastery Probability (*IP*)	Low	0.066	0.050	0.018	0.027
	High	0.066	0.052	0.017	0.031

Table 23.

Marginal means for Classification Accuracy (validation) by experimental factor and parameter (Phase 2; TP = High removed).

Factor	Level	Marginal Mean Classification Accuracy (Validation)
Sample Size (*N*)	200	92.01%
	400	92.30%
	1000	92.52%
Measurement Quality (*MQ*)	Low	89.02%
	Med	95.53%
Test Length (*J*)	1	89.79%
	3	92.71%
	5	94.32%
Time Points (*T*)	5	87.83%
	10	96.72%
Transition Probability (*TP*)	Low	88.36%
	Med	96.19%
Initial Mastery Probability (*IP*)	Low	91.98%
	High	93.23%

Figure 27 (bias in the initial probability of mastery when *N* = 200), Figure 28 (bias in the transition probability when N = 200), and Figure 29 (classification accuracy when *N* = 200) allow for a more detailed investigation of the impact of the *MQ* = Med condition. From these figures we can see that, barring the aforementioned *TP* = High conditions, estimation bias in these two most problematic parameters was excellent under medium measurement quality conditions while classification accuracy was also quite good even when *TP* = Low (a condition which results in lower classification accuracy across the board).

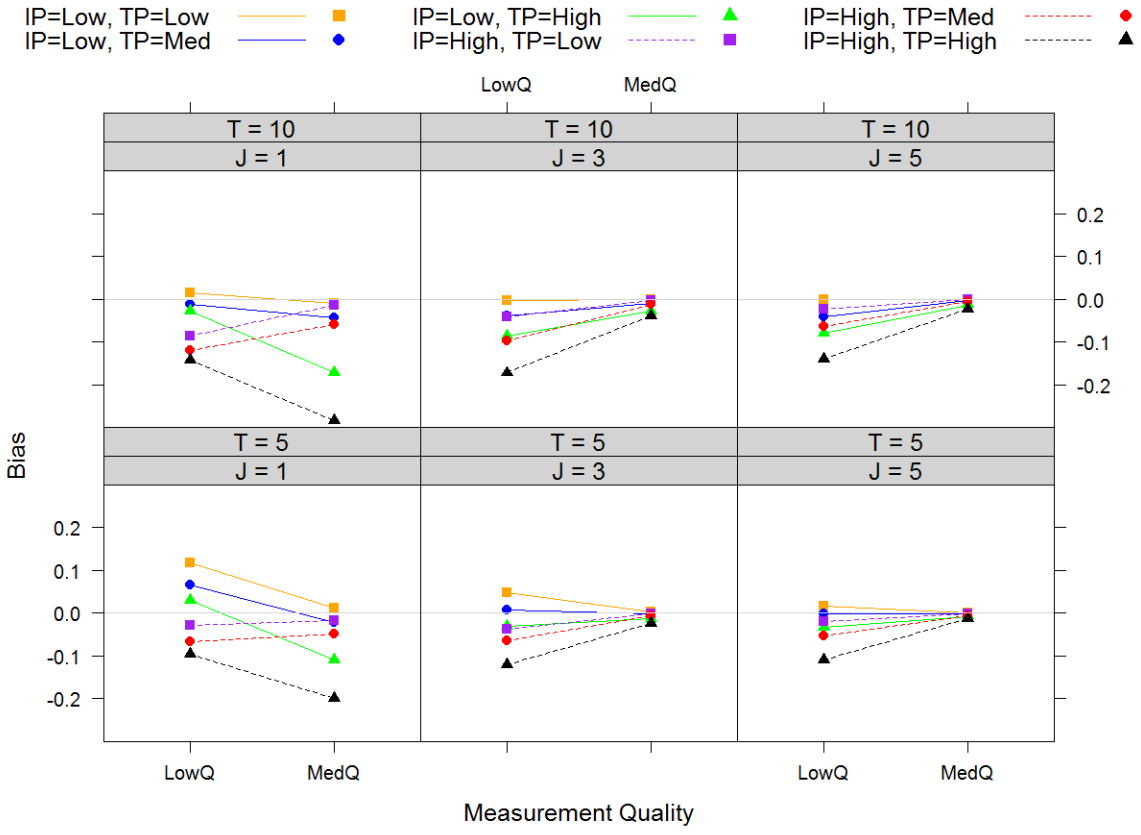

Figure 27. Bias in the initial probability of mastery estimate when *N* = 200 (Phase 2).

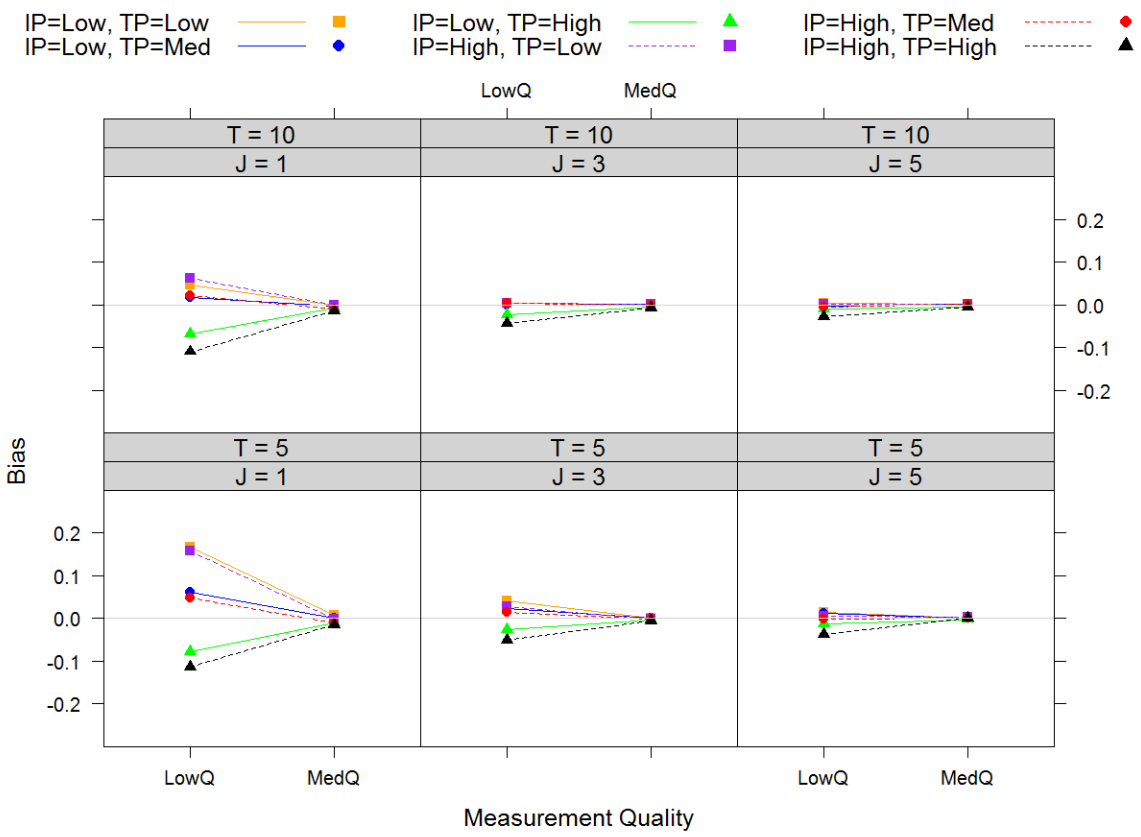

Figure 28. Bias in the transition probability estimate when N = 200 (Phase 2).

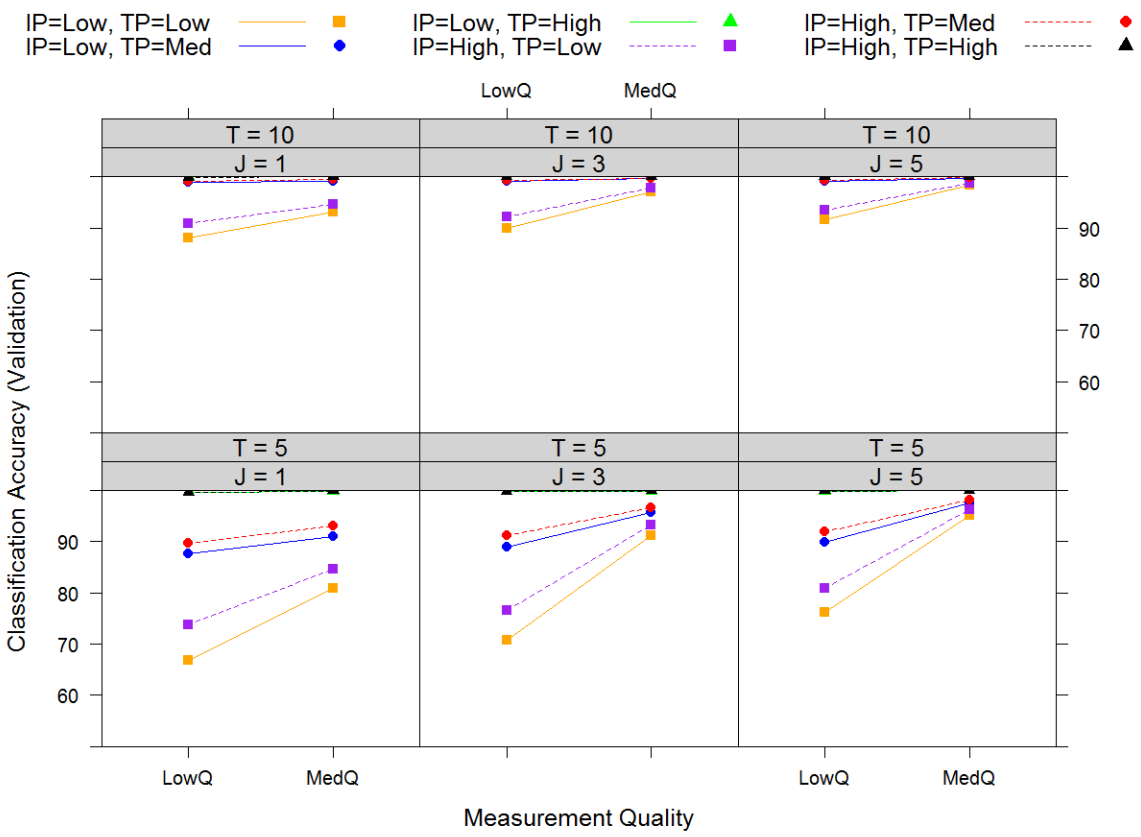

Figure 29. Classification accuracy (validation) when N = 200 (Phase 2).

Sufficiency of *N* = 400. From the Phase 2 main effect tables presented above, it can be seen that relative bias (Table 20) was within the recommended bounds (≤ 5%; Muthén & Muthén, 2002) and that classification accuracy (Table 23) was acceptable when N = 400 suggesting that such a sample size may be sufficient for model calibration under most conditions. Further investigation using Figure 30 (classification accuracy) and Figure 31 (bias for the initial probability of mastery) supports that notion save for when *TP* = High as long as measurement quality is sufficient (at least *MQ* = Med in the current study).

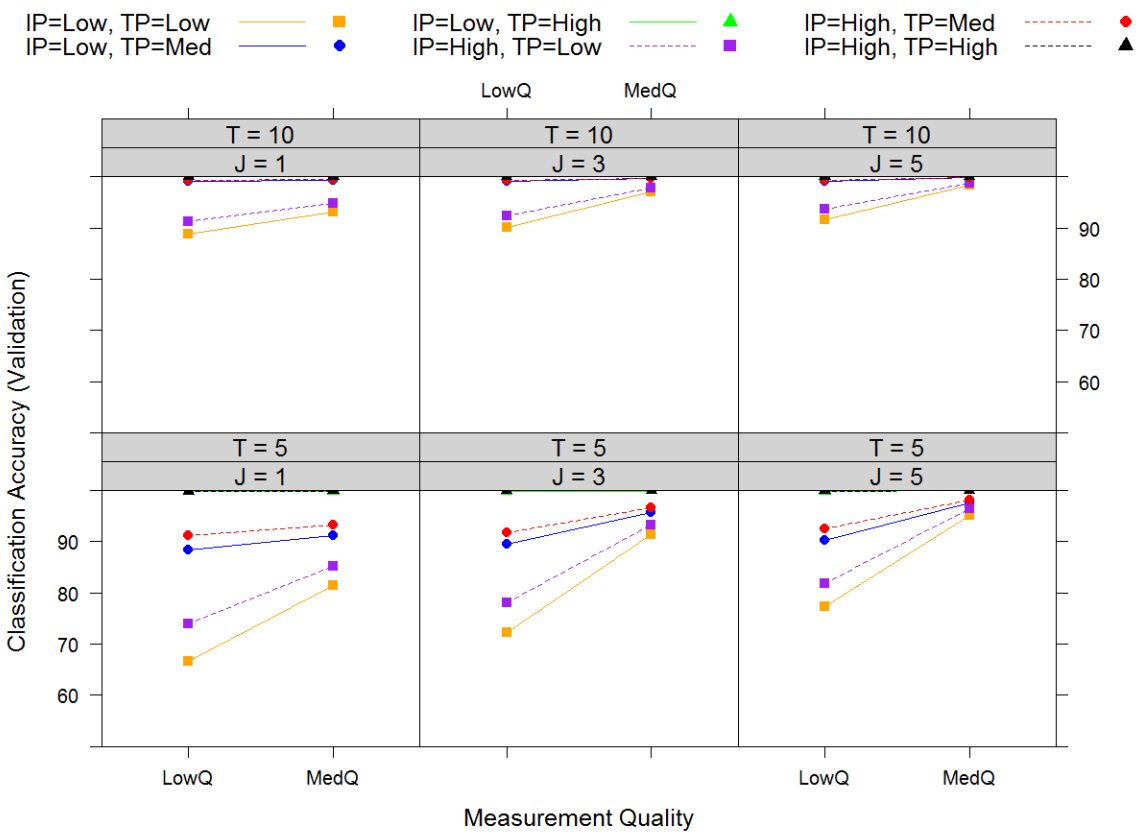

Figure 30. Classification accuracy (validation) when N = 400 (Phase 2).

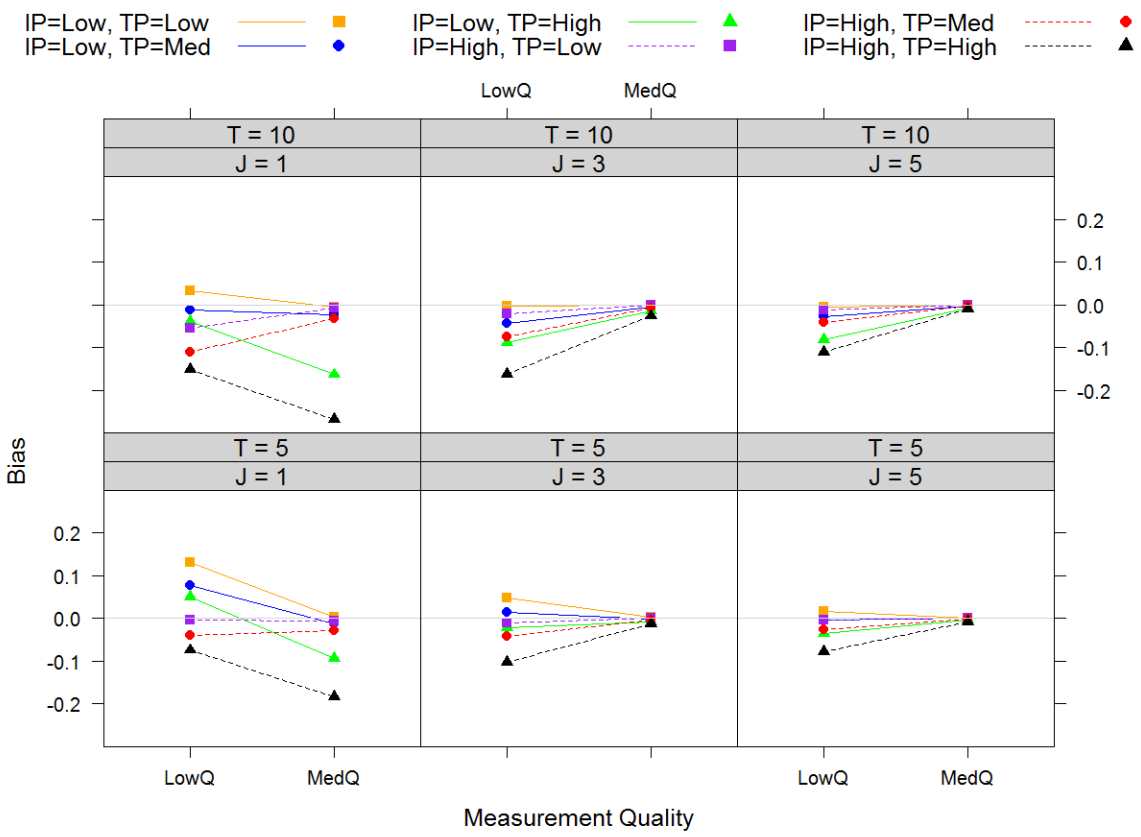

Figure 31. Bias in the estimation of the initial probability of mastery when N = 400 (Phase 2).

Sufficiency of J = 3. Referring again to Table 20 and Table 23, administering a test with at least J = 3 items per time point yields acceptably low absolute relative bias and acceptably high classification accuracy (using an arbitrary 90% cutoff, for example), marginalized across the other design facets. Referring back to Figure 27, we can see that the J = 3 conditions were sufficient in producing estimation bias near zero when measurement quality was at least MQ = Med. Finally, it was noted in the Phase 1 results that there was a significant impact of test length on estimation efficiency for the parameters $P(\theta_{t1} = M)$ and $P(\theta_{t+1} = M \mid \theta_t = NM)$, particularly when N = 1,000 and MQ =

Low. Figure 32 and Figure 33 portray estimation efficiency results under the aforementioned conditions for the two parameters, respectively.

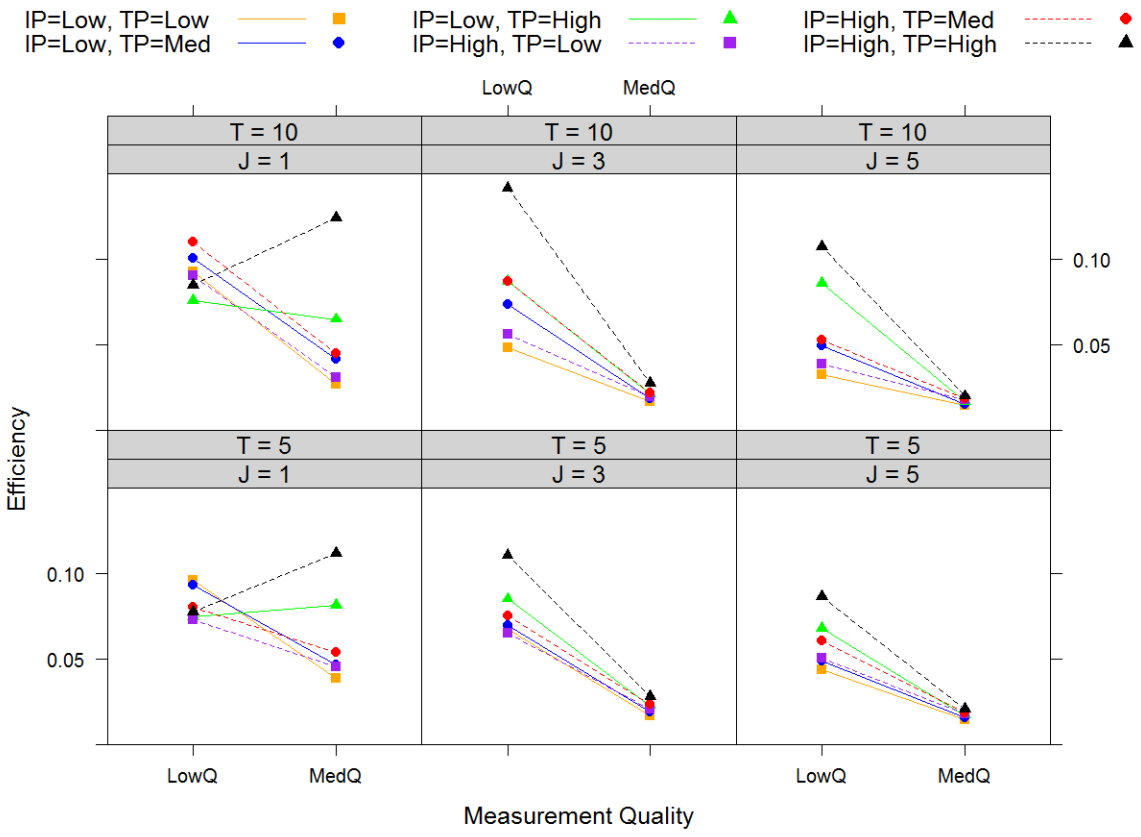

Figure 32. Estimation efficiency for the initial probability of mastery when N = 1,000 (Phase 2).

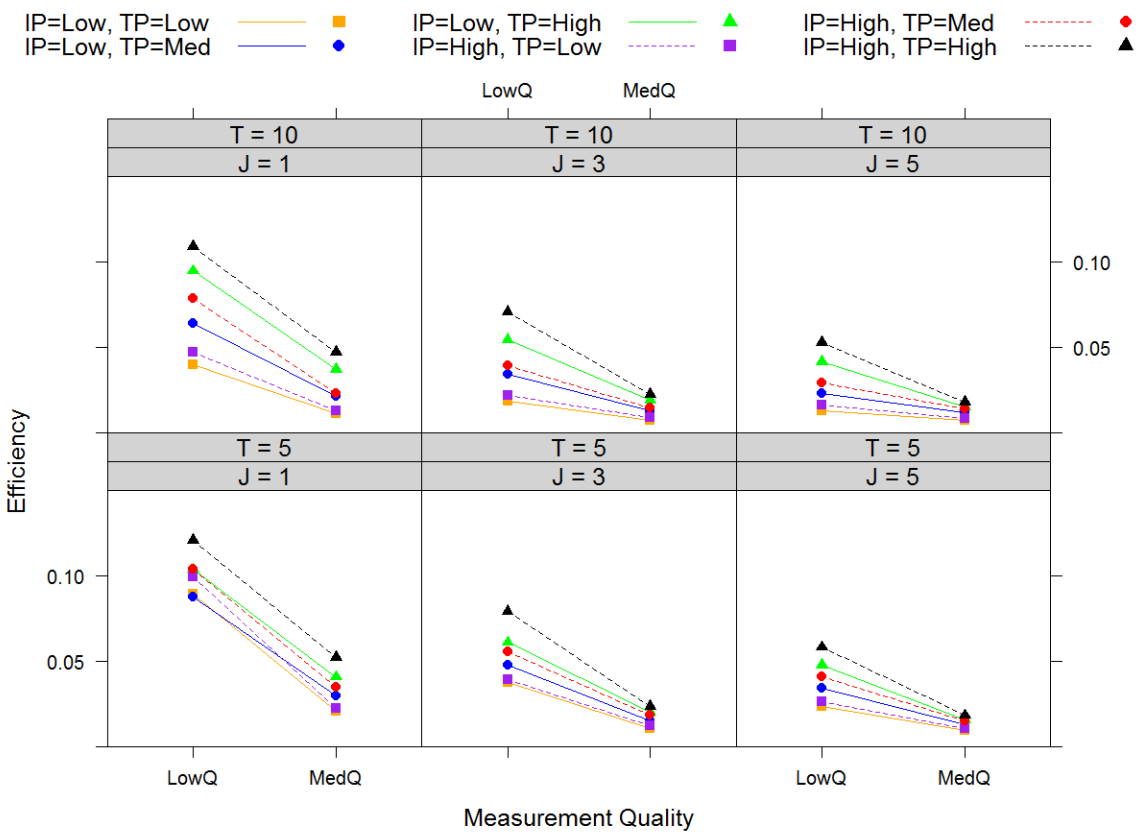

Figure 33. Estimation efficiency for the transition probability when N = 1,000 (Phase 2).

From these plots we can see that the difference in efficiency when N = 1,000 and MQ = Low is negligible when going from J = 3 to J = 5, suggesting that the former may be a sufficient test length for producing stable parameter estimates.

Summary

This chapter has presented results for Phase 1 and Phase 2 as well as a follow-up study investigating the efficacy of various dummy coding strategies for implementing the desired model constraints. Bias, relative bias, RMSE, efficiency, and classification accuracy were the outcomes of interest for the current work. In general, the results

suggested that measurement quality is the primary driver of parameter recovery followed by sample size. Several interactions between other design facets were noted.

Chapter 5

DISCUSSION

Research Hypotheses – Revisited

Several hypotheses were put forth in Chapter 3 with regard to the anticipated results from the current study. Most, though not all of these hypotheses were supported, or at least partially supported by the results. Below is a restatement of these hypotheses accompanied by a brief summary of the evidence for, or against each.

Hypothesis 1. It was expected that parameter recovery index values would improve (i.e., raw and relative bias approach zero, RMSE and efficiency decrease, classification accuracy approaches 100%) as information quantity (sample size) increased. This hypothesis was *partially supported* as a limited number of conditions (*MQ* = Low combined with *J* = 1) yielded parameter recovery (bias, in particular) that was worse as sample size increased. This effect was only noted for the initial probability of mastery parameter and may have been mostly attributed to the dummy coding strategy applied. Furthermore, increasing sample size had little impact on bias for the measurement model parameters or classification accuracy and only impacted bias for the transition probability in the presence of low measurement quality. As one would expect, increasing sample size did result in more stable estimates (i.e., greater efficiency).

Hypothesis 2. It was expected that parameter recovery index values would improve as information quality (measurement quality) increased. Furthermore, it was expected that classification accuracy, in particular, would be extremely poor (relative to other conditions) as item information approached zero. Measurement quality ended up being the primary driver of parameter recovery, thus this hypothesis was *supported.*

Additionally, there was *support* for the notion that poor measurement quality would yield poor classification accuracy. The improvement in parameter recovery as a function of measurement quality held for all parameters and all indices save for the presence of ceiling/floor effects.

Hypothesis 3. It was expected that parameter recovery index values would improve as the number of time slices and the number of items per time slice increased. This hypothesis was only *partially supported*. Increasing test length (*J*) yielded better, or in some cases equivalent, parameter recovery save for when *TP* = High for all parameters/indices. Increasing the number of time points (*T*), however tended to yield more negative bias in the initial probability of mastery parameter, specifically (i.e., the effect was not noted for any other parameter). When $T = 5$ conditions yielded negative bias for this parameter, then, the $T = 10$ conditions often yielded more bias for that parameter in the absolute sense than the $T = 5$ conditions.

Hypothesis 4. It was expected that parameter recovery index values would improve as the true values for the transition probability and initial mastery probability approached either one or zero. This hypothesis was more exploratory in nature and was *not supported* by the results of the current work. Increasing *TP* values to extremely large levels in Phase 2 yielded very problematic results to the extent that those conditions were excluded from many of the results presented in Chapter 4. Both *TP* and *IP* exhibited a significant main effect on classification accuracy, though that effect may be due to the particular assumptions encoded in the models employed in this study. The impact of *TP* and *IP* was negligible in terms of the other parameter recovery indices.

Hypothesis 5. It was expected that there would be noticeable interactions between the effect of the design facets on the values of the parameter recovery indices. This hypothesis was *supported* as many interactions can be noted in the plots presented in Chapter 4. For example, the impact of most every design facet was more prominent when *MQ* = Low as opposed to when *MQ* = Med/High.

Interpretation and Recommendations

The results of the current study uncover a number of phenomena of practical importance. First, increasing item quality would appear to represent the most efficient means of improving parameter recovery for the DBNs tested here. This will, of course, come as no surprise to measurement scientists who know that quality tasks provide the foundation for any psychometric endeavor. Poor items may yield poor measurement regardless of how many poor items one obtains data for (in the practical sense, if not the mathematical sense). In the current work this can be noted via an examination of the plots presented in the previous chapter. In Figure 16 and Figure 17, for example, we see the presence of bias even when $J = 5$ and $N = 1{,}000$ when *MQ* = Low.

Second, there appeared to be diminishing returns in terms of parameter recovery when increasing sample size from $N = 400$ to $N = 1{,}000$. Furthermore, $N = 400$ resulted in sufficient parameter recovery under most conditions, though it should be noted that what constitutes "sufficient" parameter recovery varies by application. This suggests that a sample size of $N = 400$ may be enough for model calibration under most common psychometric settings (e.g., multiple reasonably discriminating items per time point).

Third, it was, in general more difficult to estimate the probability of a correct response for a non-master ($P(X=1|\theta=NM)$) than for a master ($P(X=1|\theta=M)$). This effect increased in prominence as the true value for the transition probability increased. Larger transition probabilities lead to a sharper decrease in the number of non-masters with the passage of time relative to lower transition probabilities. This means that there is less information with which to estimate the $P(X=1|\theta=NM)$ parameter at later time points when transition probabilities are higher. From a practical perspective, one should potentially be wary of estimates for elements of the measurement model related to non-masters when assessing skills with high acquisition rates, particularly in smaller samples. Interestingly, overall model performance in terms of classification accuracy increased as the true value for the transition probability increased.

Finally, though models with a single item per time point ($J = 1$) are common in practice (e.g., the BKT model commonly applied with intelligent tutoring systems), such models may be problematic in terms of estimation and examinee classification under conditions save for those where a very high-quality (i.e., highly discriminating) task is being used. Using at least three items per time point ($J = 3$) worked well as long as measurement quality was not poor. This fact, when combined with the aforementioned impact of measurement quality on parameter recovery speaks to the trade-off between task discrimination and test length when choosing between fine and course-grain skill assessment. Assessment of fine-grain skills often implies very specific, and often very discriminating tasks. Such tasks can be difficult to generate and may become redundant with even short tests. Course-grain skill assessment, on the other hand often implies more general and less discriminating tasks, which then necessitates a larger number of items

per test. The results of the current work offer guidance under both scenarios. For the assessment of very specific skills, such as those often assessed with intelligent tutoring systems, one item per time point must be combined with exceptional measurement quality. When assessing more course-grain skills or broad domains, as few as three items per time point may be sufficient but measurement quality still needs to be moderate or better -- low quality measurement may not be sufficient under any test length condition. These recommendations are, of course predicated on the manner in which the levels of measurement quality were operationalized in this study.

Limitations and Opportunities for Further Research

Several factors were present which may limit the generalizability of the results from the current study. First, as with most all studies, the list of experimental conditions and model structures examined was not comprehensive. Further research could be conducted to fill in the gaps between the levels of the design facets tested in this study (e.g., testing a sample size between $N = 200$ and $N = 400$) as well as to examine parameter recovery performance for more complex models (e.g., multiple latent variables per time point) in an effort to refine the practical recommendations offered above. Second, the study was, in part reliant on several quirks which may be unique to Netica. The dummy coding approach used to encode the "once a master, always a master" assumption, for example, was only necessary due to the inability to fix the value for individual cells within a CPT in Netica. As was mentioned in Chapter 4, this work-around had a non-negligible impact on parameter recovery under certain conditions. To the best of the author's knowledge, however, there is no software available that implements ML estimation for BN/DBN that contain latent (i.e., completely missing)

variables that can also interface with R. Several R packages for fitting BNs exist (the previously mentioned “bnlearn” package, for example) but these packages are not capable of fitting models in the presence of latent variables. There are several commercially-available applications that function in a fashion similar to Netica (e.g., BayesiaLab [www.bayesia.com], Genie [www.bayesfusion.com], Hugin [www.hugin.com]). These applications are not easily accessed from R, however, though most have an API that is compatible with either the Java or C languages. Efforts could be made to either expand on the available R packages such that they would be able to handle the types of data used in the current study or replicate the current study using another software package accessed through a different programming environment.

The current work used ML estimation exclusively. Using an alternate estimation method (e.g., MCMC) could yield advantages over ML particularly in small sample conditions. The software for implementing MCMC (e.g., JAGS [Plummer, 2017], WinBUGS [Lunn, Thomas, Best, & Spiegelhalter, 2000]) would also offer the flexibility to implement any desired model constraints without the need for the type work-around employed in the current work. These advantages would come with a cost, however, in that MCMC estimation can be very computationally expensive and that the software for implementing MCMC often comes with a steep learning curve. There would also be choices that would need to be made (e.g., specifying prior distributions) which some practitioners may lack the training or comfort level necessary to make. Despite this, a logical next step to expand upon the current work would be to offer a comparison of ML and MCMC estimation methods under a shared set of conditions.

Estimation and classification accuracy for the models tested here was best when assessing novel skills (i.e., *IP* = Low) that can be learned relatively quickly, within reason (i.e., *TP* = Med). This outcome carries potentially important practical implications, but was not tested extensively enough in the current work to be considered reliable. This conclusion warrants further testing under an expanded set of realistic *IP*/*TP* combinations. Additionally, the increase in classification accuracy as *TP* increased could be due to the model assumptions, namely the "once a master, always a master" model constraint. Further research should be conducted to evaluate the impact of this constraint as well as other model constraints applied in the current work (e.g., static transition/measurement model CPTs).

Finally, all conditions in the current study used what might be described as "minimally informative" start values in order to alleviate potential issues related to label-switching. It is not clear what effect these start values might have had on the final results relative to more informative start values or no start values at all, though serious estimation problems would be expected under the latter case. More informative start values would almost certainly improve parameter recovery and would be useful under the more problematic conditions noted here (i.e., poor measurement quality and only one item per time point). Based on the results of the current study, improved start values would yield better estimation for the initial probability of mastery and transition probability parameters in particular. While these are often nuisance parameters in practice, they have implications for the estimation of other, perhaps more centrally important parameters. Further research could be conducted to examine strategies for

generating plausible start values based on existing literature, features of extant data, or some other source.

Summary

Parameter recovery performance for DBNs is heavily dependent on the quality of the tasks included in the model. Tasks with exceptionally high quality can mask deficiencies in any of the other facets test in the current study. That is to say that high quality tasks may be sufficient for overcoming the combination of small sample size, single-item tests, and relatively small number of measurement occasions with respect to reducing absolute bias and increasing classification accuracy, particularly for parameters that may be more difficult to estimate such as the prior probability of mastery and the transition probability. On the other hand, poor measurement quality may act as an impediment to satisfactory parameter recovery for DBNs even with longer tests, more measurement occasions, and large samples. In some cases, increasing sample size in the presence of poor measurement quality may even degrade parameter recovery, though that conclusion requires further testing using more flexible software. Further research is also required to compare the performance of different estimation methods under a wider-variety of experimental conditions. In light of the results presented here, it is recommended that practitioners exercise due diligence in evaluating the ability of their tasks/items to differentiate between learners in different performance categories before undertaking any intensive data collection efforts. Assuming the items are of sufficient quality, then a sample size of at least $N = 400$ with at least $J = 3$ items per time point may be sufficient to calibrate models similar in specification to those tested in the current work. Care should be taken when using only a single item per measurement occasion as

exceptionally discriminating items may be required in order to adequately fit these models.

REFERENCES

Agresti, A., Mehta, C. R., & Patel, N. R. (1990). Exact inference for contingency tables with ordered categories. *Journal of the American Statistical Association*, *85*(410), 453–458.

Almond, R. G. (2007a). An illustration of the use of Markov decision processes to represent student growth (learning). *ETS Research Report Series*, *2007*(2). Retrieved from http://onlinelibrary.wiley.com.ezproxy1.lib.asu.edu/doi/10.1002/j.2333-8504.2007.tb02082.x/full

Almond, R. G. (2007b). Cognitive modeling to represent growth (learning) using markov decision processes. *Technology, Instruction, Cognition and Learning (TICL)*, *5*, 313–324.

Almond, R. G. (2017). RNetica release information. Retrieved from: https://pluto.coe.fsu.edu/RNetica/RNetica.html

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223–237.

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian Networks in Educational Assessment*. Springer.

Almond, R. G., Mulder, J., Hemat, L. A., & Yan, D. (2009). Bayesian network models for local dependence among observable outcome variables. *Journal of Educational and Behavioral Statistics*, *34*(4), 491–521.

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J.-D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning*, *50*(3), 450–460.

Almond, R., Yan, D., & Hemat, L. (2007). Parameter recovery studies with a diagnostic Bayesian network model. *Behaviormetrika*, *35*(2), 159–185.

Anderson, E.B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, pp. 3-16.

Baldwin, E. (2015). *A Monte Carlo Simulation Study Examining Statistical Power in Latent Transition Analysis*. UC Santa Barbara. Retrieved from https://alexandria.ucsb.edu/downloads/q524jn927

Balov, N., & Salzman, P. (2017). How to use the catnet package. Retrieved from https://cran.r-project.org/web/packages/catnet/vignettes/catnet.pdf

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling.

Bandalos, D. L., & Leite, W. (2013). The role of simulation in structural equation modeling. *Structural equation modeling: A second course (2nd ed., pp. 625-666). Greenwich, CT: Information Age*.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Retrieved from http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/090310.pdf

Beck, J., & Chang, K. M. (2007). Identifiability: A fundamental problem of student modeling. *User Modeling 2007*, 137-146.

Boutilier, C., Dean, T., & Hanks, S. (1999). Decision-theoretic planning: Structural assumptions and computational leverage. *Journal of Artificial Intelligence Research*, *11*(1), 94.

Carmona, C., Castillo, G., & Millán, E. (2008). Designing a dynamic Bayesian network for modeling students' learning styles. In *Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on* (pp. 346–350). IEEE. Retrieved from http://ieeexplore.ieee.org.ezproxy1.lib.asu.edu/xpls/abs_all.jsp?arnumber=4561705

Chen, Z., & Brown, E. N. (2013). State space model. *Scholarpedia*, *8*(3), 30868. https://doi.org/10.4249/scholarpedia.30868

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*(4), 327–335.

Choi, Y. (2012). Dynamic Bayesian Inference Networks and Hidden Markov Models for Modeling Learning Progressions over Multiple Time Points. Retrieved from http://drum.lib.umd.edu/handle/1903/12739

Chung, G., Baker, E. L., Vendlinski, T. P., Buschang, R. E., Delacruz, G. C., Michiuye, J. K., & Bittick, S. J. (2010). Testing instructional design variations in a prototype math game. In *R. Atkinson (Chair), Current perspectives from three national R&D centers focused on game-based learning: Issues in learning, instruction, assessment, and game design. Structured poster session at the annual meeting of the American Educational Research Association, Denver, CO*.

Coetzee, D. (2014). Choosing Sample Size for Knowledge Tracing Models. In *EDM (Workshops)*.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Collins, L. M., & Tracy, A. J. (1997). Estimation in complex latent transition models with extreme data sparseness. *Kwantitatieve Methoden*, *18*, 57–71.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*(3), 375–389.

Collins, L. M., & Lanza, S. T. (2013). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). John Wiley & Sons.

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

Collins, L. M., Wugalter, S. E., & Fidler, P. L. (1996). Some Practical Issues Related to the Estimation of Latent Class and Latent Transition Parameters-6.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence*, *16*(7–8), 555–575.

Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction*, *19*(3), 267–303.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*(4), 253–278.

Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). Intelligent tutoring systems. *Handbook of Human-Computer Interaction*, *5*, 849–874.

Crawford, A. (2014). *Posterior predictive model checking in Bayesian networks*. Arizona State University. Retried from https://repository.asu.edu/items/24820.

Culbertson, M. J. (2014). *Graphical models for student knowledge: Networks, parameters, and item selection*. University of Illinois at Urbana-Champaign. Retrieved from https://www.ideals.illinois.edu/handle/2142/49372

Culbertson, M. J. (2015). Bayesian Networks in Educational Assessment: The State of the Field. *Applied Psychological Measurement*, 0146621615590401.

Culpepper, S.A. (2014). If at first you don't succeed try, try again: Applications of sequential IRT models to cognitive assessments. *Applied Psychological Measurement, 38(8),* pp. 632-644.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice* (Vol. 96). Sage Publications.

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

Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*(6), 985–987.

Eseryel, D., Ge, X., Ifenthaler, D., & Law, V. (2011). Dynamic modeling as a cognitive regulation scaffold for developing complex problem-solving skills in an educational massively multiplayer online game environment. *Journal of Educational Computing Research*, *45*(3), 265–286.

Fisher, C. R., Walsh, M. M., Blaha, L. M., Gunzelmann, G., & Veksler, B. (2016). Efficient Parameter Estimation of Cognitive Models for Real-Time Performance Monitoring and Adaptive Interfaces. Retrieved from http://acs.ist.psu.edu/iccm2016/proceedings/fisher2016iccm.pdf

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.

Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *15*(01), 9–42.

Ghahramani, Z., & Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural computation*, *12*(4), 831-864.

Gilks, W.R., Best, N. G., & Tan, K. K. C. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics*, 455–472.

Gilks, W.R., Richardson, S., & Spiegelhalter, D. (1996). *Markov chain Monte Carlo in practice*. CRC press.

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (Vol. 20). CRC press.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*(1), 129–187.

González-Brenes, J. P., Behrens, J. T., Mislevy, R. J., Levy, R., & DiCerbo, K. E. (2016). Bayesian Networks. *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, 328.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 107–114.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, *49*(268), 732–764.

Guo, Z., Gao, X., Di, R., & Yang, Y. (2015). Learning Bayesian Network Parameters from Small Data Set: A Spatially Maximum a Posteriori Method. In *Workshop on Advanced Methodologies for Bayesian Networks* (pp. 32–45). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-28379-1_3

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Hoffman, M. D., & Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Iseli, M. R., Koenig, A. D., Lee, J. J., & Wainess, R. (2010). Automated assessment of complex task performance in games and simulations. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference*. Retrieved from http://cresst.org/wp-content/uploads/R775.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.

Jensen, F. V. (1996). *An introduction to Bayesian networks* (Vol. 210). UCL press London.

Johns, J., & Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 21, No. 1, p. 163). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

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.

Klingler, S., Käser, T., Solenthaler, B., & Gross, M. (2015). On the Performance Characteristics of Latent-Factor and Knowledge Tracing Models. *International Educational Data Mining Society*. Retrieved from http://eric.ed.gov/?id=ED560586

LaMar, M. (2018). Markov decision process measurement model. *Psychometrika, 83*(1), 67-88.

Levy, R. (2014). Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments. CRESST Report 837. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. Retrieved from http://eric.ed.gov/?id=ED555714

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*(7), 519–537.

Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *64*(2), 208–232.

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2015). A Standardized Generalized Dimensionality Discrepancy Measure and a Standardized Model-Based Covariance for Dimensionality Assessment for Multidimensional Models. *Journal of Educational Measurement*, *52*(2), 144–158.

Li, F., Cohen, A., Bottge, B., & Templin, J. (2016). A latent transition analysis model for assessing change in cognitive skills. *Educational and Psychological Measurement*, *76*(2), 181-204.

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modeling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337.

Madison, M.J. & Bradshaw, L. (in press). Evaluating intervention effects in a diagnostic classification model framework. *Journal of Educational Measurement.*

Mislevy, R. J., Corrigan, S., Oranje, A., DiCerbo, K., Bauer, M. I., von Davier, A., & John, M. (2015). Psychometrics and game-based assessment. *Technology and Testing: Improving Educational and Psychological Measurement*, 23.

Mislevy, R. J., & Gitomer, D. H. (1995). The role of probability-based inference in an intelligent tutoring system. *ETS Research Report Series*, *1995*(2).

Mood, A. M., Graybill, F. A., & Boes, D. (1974). Introduction to the Theory of. *Statistics*, *3*.

Murphy, K. P. (2002). *Dynamic Bayesian networks: representation, inference and learning*. University of California, Berkeley.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100.

Neapolitan, R. E.. (2004). *Learning Bayesian networks* (Vol. 38). Pearson Prentice Hall Upper Saddle River, NJ.

Nielsen, T. D., & Jensen, F. V. (2009). *Bayesian networks and decision graphs*. Springer Science & Business Media.

Nooraei, B., Pardos, Z., Heffernan, N., & Baker, R. (2010). Less is more: Improving the speed and prediction power of knowledge tracing by using less data. In *Educational Data Mining 2011*. Retrieved from http://www.educationaldatamining.org/conferences/index.php/EDM/2011/paper/download/893/859

Norsys Sofware Corp. (1995-2017). Netica. www.norsys.com.

Ortony, A., Clore, G. L., & Collins, A. (1988). The cognitive structure of emotions. 10.1017. *CBO9780511571299*.

Pardos, Z. A., & Heffernan, N. T. (2010). Modeling individualization in a Bayesian networks implementation of knowledge tracing. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 255–266). Springer. Retrieved from http://link.springer.com.ezproxy1.lib.asu.edu/10.1007%2F978-3-642-13470-8_24

Pardos, Z., Bergner, Y., Seaton, D., & Pritchard, D. (2013). Adapting Bayesian knowledge tracing to a massive open online course in edx. In *Educational Data Mining 2013*. Retrieved from http://www.educationaldatamining.org/conferences/index.php/EDM/2013/paper/download/1030/996

Pardos, Z., Dailey, M., & Heffernan, N. (2010). Learning what works in ITS from non-traditional randomized controlled trial data. In *Intelligent Tutoring Systems* (pp. 41–50). Springer. Retrieved from http://www.springerlink.com.ezproxy1.lib.asu.edu/index/Q4W0078606301W00.pdf

Pavlik, P. I., Cen, H., & Koedinger, K. R. KR: Performance Factors Analysis--A New Alternative to Knowledge Tracing. In *The 14th International Conference on Artificial Intelligence in Education, 2009*.

Pearl, J. (1988). *Probabilistic inference in intelligent systems*. Morgan Kaufmann San Mateo, CA.

Pitchforth, J., & Mengersen, K. (2013). A proposed validation framework for expert elicited Bayesian Networks. *Expert Systems with Applications*, *40*(1), 162–167.

Plummer, M. (2017). JAGS version 4.3.0 user manual [computer software manual]. Retrieved from sourceforge.net/projects/mcmc-jags/files/manuals/4.x.

Qiu, Y., Qi, Y., Lu, H., Pardos, Z., & Heffernan, N. (2010). Does time matter? Modeling the effect of time with Bayesian knowledge tracing. In *Educational Data Mining 2011*. Retrieved from http://www.educationaldatamining.org/conferences/index.php/EDM/2011/paper/download/897/863.

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

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2011). Faster teaching by POMDP planning. In *International Conference on Artificial Intelligence in Education* (pp. 280–287). Springer. Retrieved from http://link.springer.com.ezproxy1.lib.asu.edu/chapter/10.1007/978-3-642-21869-9_37

Read, T. R., & Cressie, N. A. (1988). Introduction to the Power-Divergence Statistic. In *Goodness-of-Fit Statistics for Discrete Multivariate Data* (pp. 1–4). Springer. Retrieved from http://link.springer.com.ezproxy1.lib.asu.edu/chapter/10.1007/978-1-4612-4578-0_1

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In *Handbook of modern item response theory* (pp. 271–286). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4757-2691-6_16

Reichenberg, R.E. (in press). Dynamic Bayesian networks in educational measurement: Reviewing and advancing the state of the field. *Applied Measurement in Education.*

Renooij, S. (2001). Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review*, *16*(03), 255–269.

Reye, J. (2004). Student modelling based on belief networks. *International Journal of Artificial Intelligence in Education*, *14*(1), 63–96.

Rodríguez, C. E., & Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, *23*(1), 25-45.

Rowe, J. P., & Lester, J. C. (2010). Modeling User Knowledge with Dynamic Bayesian Networks in Interactive Narrative Environments. In *AIIDE*. Retrieved from https://pdfs.semanticscholar.org/08a8/8dc4db85164de01f8fa1cfe3013066c49f7d.pdf

Rubin, D. B., & others. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151–1172.

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

Sabourin, J., Mott, B., & Lester, J. (2013). Utilizing dynamic bayes nets to improve early prediction models of self-regulated learning. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 228–241). Springer. Retrieved from http://link.springer.com.ezproxy1.lib.asu.edu/chapter/10.1007/978-3-642-38844-6_19

Sao Pedro, M. A., de Baker, R. S., Gobert, J. D., Montalvo, O., & Nakama, A. (2013). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, 1–39.

Scutari, M. Learning Bayesian networks with the bnlearn R package. In *Journal of Statistical Software*.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., Leighton, J. P., Jang, E. E., & Chu, M.-W. (2016). Advances in the science of assessment. *Educational Assessment*, *21*(1), 34–59.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*(1), 1–33.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, *67*(2), 239–257.

Sinharay, S., Almond, R., & Yan, D. (2004). Assessing fit of models with discrete proficiency variables in educational assessment. *ETS Research Report Series*, *2004*(1). Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2004.tb01934.x/full

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.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 795-809.

Tversky, A., & Kahneman, D. (1975). Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making* (pp. 141–162). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-94-010-1834-0_8

van de Sande, B. (2013). Applying Three Models of Learning to Individual Student Log Data. In *D'Mello, S. K., Calvo, R. A., and Olney, A. (eds.) Proceedings of the 6th International Conference on Educational Data Mining.* (pp. 193–199). International Educational Data Mining Society. Retrieved from http://www.educationaldatamining.org/conferences/index.php/EDM/2013/paper/download/1037/1003

VanLehn, K. (2008). Intelligent tutoring systems for continuous, embedded assessment. *The Future of Assessment: Shaping Teaching and Learning*, 113–138.

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),* pp. 318-336.

Weaver, W. (1948). Probability, rarity, interest and surprise. *Scientific Monthly*, *67*(6), 390–392.

Williamson, D. M., Almond, R. G., & Mislevy, R. J. (2000). Model criticism of Bayesian networks with latent variables. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence* (pp. 634-643). Morgan Kaufmann Publishers Inc..

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing. New Jersey: Lawrence Erlbaum Associates*. Inc.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized Bayesian knowledge tracing models. In *International Conference on Artificial Intelligence in Education* (pp. 171–180). Springer. Retrieved from http://link.springer.com/10.1007/978-3-642-39112-5_18

Zapata-Rivera, D., & Bauer, M. (2012). Exploring the role of games in educational assessment. *Technology-Based Assessments for Twenty-First-Century Skills: Theoretical and Practical Implications from Modern Research*, 147–169.

APPENDIX A

RAW RESULTS TABLES FOR PHASE 1

Table A1.

Raw results for estimation bias in the initial probability of mastery parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.114	0.164	0.002	-0.001
	J = 1, *T* = 10	0.024	0.016	-0.004	-0.001
	J = 5, *T* = 5	0.020	0.005	-0.001	0.000
	J = 5, *T* = 10	-0.007	-0.003	0.002	0.000
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.066	0.078	-0.009	-0.001
	J = 1, *T* = 10	-0.016	-0.038	-0.016	-0.003
	J = 5, *T* = 5	-0.008	-0.006	-0.001	0.000
	J = 5, *T* = 10	-0.035	-0.024	0.000	-0.001
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.030	0.016	0.000	0.000
	J = 1, *T* = 10	-0.076	-0.027	0.000	-0.001
	J = 5, *T* = 5	-0.018	-0.005	0.000	0.000
	J = 5, *T* = 10	-0.021	-0.007	0.000	0.000
TP = High, *IP* = High	*J* = 1, *T* = 5	-0.065	-0.023	-0.014	-0.003
	J = 1, *T* = 10	-0.128	-0.124	-0.021	-0.004
	J = 5, *T* = 5	-0.046	-0.018	-0.002	0.001
	J = 5, *T* = 10	-0.059	-0.033	0.000	0.000

Table A2.

Raw results for estimation bias in the transition probability parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.169	0.098	0.002	0.000
	J = 1, *T* = 10	0.045	0.014	0.001	0.000
	J = 5, *T* = 5	0.017	0.005	0.000	0.000
	J = 5, *T* = 10	0.003	0.005	0.000	0.000
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.067	0.036	0.000	0.001
	J = 1, *T* = 10	0.018	-0.001	-0.002	0.000
	J = 5, *T* = 5	0.012	0.004	0.002	0.001
	J = 5, *T* = 10	-0.001	0.004	0.002	0.000
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.167	0.078	-0.001	0.001
	J = 1, *T* = 10	0.054	0.016	-0.001	0.000
	J = 5, *T* = 5	0.006	0.005	0.001	0.000
	J = 5, *T* = 10	0.003	0.007	0.002	0.000
TP = High, *IP* = High	*J* = 1, *T* = 5	0.062	0.024	-0.003	0.000
	J = 1, *T* = 10	0.028	-0.007	-0.001	0.000
	J = 5, *T* = 5	0.003	-0.002	0.001	0.001
	J = 5, *T* = 10	-0.004	0.003	0.003	0.000

Table A3.

Raw results for estimation bias in the probability of a correct response (non-master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-0.028	-0.042	0.004	0.001
	J = 1, *T* = 10	-0.010	-0.007	0.008	0.001
	J = 5, *T* = 5	-0.004	-0.001	0.004	0.001
	J = 5, *T* = 10	0.002	-0.001	0.005	0.001
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.022	-0.023	0.013	0.002
	J = 1, *T* = 10	0.001	0.005	0.025	0.005
	J = 5, *T* = 5	0.002	0.001	0.006	0.001
	J = 5, *T* = 10	0.010	0.004	0.011	0.002
TP = High, *IP* = Low	*J* = 1, *T* = 5	-0.001	-0.030	0.008	0.001
	J = 1, *T* = 10	0.002	-0.005	0.011	0.002
	J = 5, *T* = 5	0.000	-0.001	0.005	0.001
	J = 5, *T* = 10	0.004	0.000	0.007	0.002
TP = High, *IP* = High	*J* = 1, *T* = 5	0.003	-0.004	0.024	0.003
	J = 1, *T* = 10	0.025	0.024	0.038	0.008
	J = 5, *T* = 5	0.005	0.004	0.007	0.001
	J = 5, *T* = 10	0.013	0.007	0.014	0.003

Table A4.

Raw results for estimation bias in the probability of a correct response (master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-0.029	-0.031	-0.008	0.000
	J = 1, *T* = 10	-0.006	0.000	-0.002	0.000
	J = 5, *T* = 5	-0.004	0.000	-0.005	-0.001
	J = 5, *T* = 10	0.000	0.002	-0.003	0.000
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.001	-0.004	-0.003	0.000
	J = 1, *T* = 10	0.003	0.003	-0.001	0.001
	J = 5, *T* = 5	-0.002	0.001	-0.003	-0.001
	J = 5, *T* = 10	0.001	0.003	-0.002	0.000
TP = High, *IP* = Low	*J* = 1, *T* = 5	-0.015	-0.017	-0.003	0.001
	J = 1, *T* = 10	0.000	0.002	-0.002	0.000
	J = 5, *T* = 5	0.000	0.002	-0.004	-0.001
	J = 5, *T* = 10	0.001	0.002	-0.003	0.000
TP = High, *IP* = High	*J* = 1, *T* = 5	0.002	0.000	-0.001	0.001
	J = 1, *T* = 10	0.003	0.004	0.000	0.001
	J = 5, *T* = 5	0.002	0.002	-0.003	0.000
	J = 5, *T* = 10	0.002	0.003	-0.002	0.000

Table A5.

Raw results for relative estimation bias in the initial probability of mastery parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	56.93%	81.92%	1.04%	-0.38%
	J = 1, *T* = 10	11.79%	7.92%	-1.86%	-0.47%
	J = 5, *T* = 5	10.22%	2.31%	-0.37%	0.07%
	J = 5, *T* = 10	-3.51%	-1.68%	0.86%	0.05%
TP = High, *IP* = Low	*J* = 1, *T* = 5	32.92%	39.15%	-4.30%	-0.43%
	J = 1, *T* = 10	-8.19%	-18.78%	-8.10%	-1.37%
	J = 5, *T* = 5	-3.91%	-2.86%	-0.50%	-0.19%
	J = 5, *T* = 10	-17.33%	-11.84%	0.05%	-0.43%
TP = Low, *IP* = High	*J* = 1, *T* = 5	-7.41%	3.96%	0.08%	0.01%
	J = 1, *T* = 10	-18.89%	-6.75%	-0.09%	-0.16%
	J = 5, *T* = 5	-4.58%	-1.25%	-0.03%	-0.07%
	J = 5, *T* = 10	-5.21%	-1.69%	0.03%	0.09%
TP = High, *IP* = High	*J* = 1, *T* = 5	-16.23%	-5.63%	-3.60%	-0.69%
	J = 1, *T* = 10	-32.12%	-30.98%	-5.31%	-0.98%
	J = 5, *T* = 5	-11.55%	-4.41%	-0.60%	0.23%
	J = 5, *T* = 10	-14.66%	-8.31%	-0.07%	0.05%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red).

Table A6.

Raw results for relative estimation bias in the transition probability parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	84.74%	48.82%	0.94%	0.17%
	J = 1, *T* = 10	22.49%	6.98%	0.30%	0.15%
	J = 5, *T* = 5	8.57%	2.48%	-0.04%	0.10%
	J = 5, *T* = 10	1.66%	2.73%	0.25%	0.00%
TP = High, *IP* = Low	*J* = 1, *T* = 5	16.67%	9.00%	-0.11%	0.27%
	J = 1, *T* = 10	4.42%	-0.15%	-0.58%	-0.08%
	J = 5, *T* = 5	2.89%	0.99%	0.45%	0.19%
	J = 5, *T* = 10	-0.19%	0.99%	0.56%	0.07%
TP = Low, *IP* = High	*J* = 1, *T* = 5	83.30%	39.00%	-0.66%	0.48%
	J = 1, *T* = 10	26.79%	7.89%	-0.27%	-0.01%
	J = 5, *T* = 5	2.77%	2.29%	0.62%	0.24%
	J = 5, *T* = 10	1.35%	3.56%	0.79%	0.11%
TP = High, *IP* = High	*J* = 1, *T* = 5	15.39%	5.93%	-0.87%	-0.02%
	J = 1, *T* = 10	7.02%	-1.85%	-0.33%	-0.08%
	J = 5, *T* = 5	0.71%	-0.43%	0.27%	0.36%
	J = 5, *T* = 10	-1.04%	0.86%	0.70%	0.04%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red).

Table A7.

Raw results for relative estimation bias in the probability of a correct response (non-master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-7.00%	-10.62%	3.52%	1.32%
	J = 1, *T* = 10	-2.59%	-1.77%	8.50%	1.41%
	J = 5, *T* = 5	-0.90%	-0.18%	4.10%	0.62%
	J = 5, *T* = 10	0.51%	-0.18%	5.18%	0.99%
TP = Low, *IP* = High	*J* = 1, *T* = 5	-5.46%	-5.81%	13.14%	1.55%
	J = 1, *T* = 10	0.24%	1.34%	25.17%	5.19%
	J = 5, *T* = 5	0.40%	0.15%	5.55%	0.98%
	J = 5, *T* = 10	2.41%	0.88%	11.17%	2.32%
TP = High, *IP* = Low	*J* = 1, *T* = 5	-0.37%	-7.44%	8.29%	0.54%
	J = 1, *T* = 10	0.59%	-1.19%	11.45%	2.25%
	J = 5, *T* = 5	-0.06%	-0.23%	5.31%	0.69%
	J = 5, *T* = 10	1.09%	0.01%	7.39%	1.88%
TP = High, *IP* = High	*J* = 1, *T* = 5	0.79%	-1.12%	23.72%	3.50%
	J = 1, *T* = 10	6.15%	6.11%	37.61%	7.90%
	J = 5, *T* = 5	1.33%	1.02%	7.45%	1.27%
	J = 5, *T* = 10	3.17%	1.86%	13.91%	2.95%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red).

Table A8.

Raw results for relative estimation bias in the probability of a correct response (master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-4.85%	-5.22%	-0.91%	0.00%
	J = 1, *T* = 10	-0.97%	0.03%	-0.27%	-0.05%
	J = 5, *T* = 5	-0.61%	-0.06%	-0.57%	-0.07%
	J = 5, *T* = 10	-0.04%	0.29%	-0.31%	-0.04%
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.21%	-0.65%	-0.31%	0.05%
	J = 1, *T* = 10	0.49%	0.57%	-0.08%	0.07%
	J = 5, *T* = 5	-0.27%	0.18%	-0.35%	-0.07%
	J = 5, *T* = 10	0.13%	0.51%	-0.26%	-0.04%
TP = High, *IP* = Low	*J* = 1, *T* = 5	-2.46%	-2.80%	-0.36%	0.06%
	J = 1, *T* = 10	0.01%	0.34%	-0.23%	0.01%
	J = 5, *T* = 5	-0.06%	0.35%	-0.43%	-0.09%
	J = 5, *T* = 10	0.12%	0.34%	-0.31%	-0.05%
TP = High, *IP* = High	*J* = 1, *T* = 5	0.40%	-0.05%	-0.07%	0.08%
	J = 1, *T* = 10	0.51%	0.74%	0.00%	0.06%
	J = 5, *T* = 5	0.41%	0.39%	-0.28%	-0.05%
	J = 5, *T* = 10	0.33%	0.49%	-0.24%	-0.03%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red).

Table A9.

Raw results for RMSE in the estimation of the initial probability of mastery parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.184	0.194	0.036	0.016
	J = 1, *T* = 10	0.138	0.093	0.034	0.016
	J = 5, *T* = 5	0.100	0.046	0.028	0.013
	J = 5, *T* = 10	0.072	0.033	0.029	0.013
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.152	0.123	0.040	0.018
	J = 1, *T* = 10	0.125	0.105	0.042	0.018
	J = 5, *T* = 5	0.096	0.050	0.027	0.013
	J = 5, *T* = 10	0.101	0.055	0.028	0.013
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.124	0.071	0.042	0.020
	J = 1, *T* = 10	0.151	0.091	0.040	0.019
	J = 5, *T* = 5	0.089	0.050	0.034	0.016
	J = 5, *T* = 10	0.076	0.038	0.035	0.015
TP = High, *IP* = High	*J* = 1, *T* = 5	0.145	0.081	0.049	0.022
	J = 1, *T* = 10	0.181	0.166	0.050	0.021
	J = 5, *T* = 5	0.113	0.062	0.034	0.016
	J = 5, *T* = 10	0.119	0.062	0.035	0.016

Table A10.

Raw results for RMSE in the in the estimation of the transition probability parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.251	0.133	0.022	0.010
	J = 1, *T* = 10	0.117	0.044	0.017	0.007
	J = 5, *T* = 5	0.063	0.025	0.018	0.008
	J = 5, *T* = 10	0.029	0.014	0.016	0.007
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.185	0.093	0.031	0.014
	J = 1, *T* = 10	0.134	0.063	0.028	0.012
	J = 5, *T* = 5	0.075	0.034	0.027	0.012
	J = 5, *T* = 10	0.054	0.025	0.024	0.011
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.258	0.125	0.026	0.012
	J = 1, *T* = 10	0.148	0.049	0.019	0.008
	J = 5, *T* = 5	0.063	0.027	0.022	0.009
	J = 5, *T* = 10	0.035	0.017	0.018	0.008
TP = High, *IP* = High	*J* = 1, *T* = 5	0.197	0.103	0.037	0.017
	J = 1, *T* = 10	0.169	0.079	0.032	0.015
	J = 5, *T* = 5	0.091	0.041	0.031	0.014
	J = 5, *T* = 10	0.065	0.029	0.028	0.013

Table A11.

Raw results for RMSE in the estimation of the probability of a correct response (non-master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.059	0.053	0.020	0.009
	J = 1, *T* = 10	0.040	0.025	0.017	0.007
	J = 5, *T* = 5	0.034	0.015	0.013	0.006
	J = 5, *T* = 10	0.025	0.012	0.012	0.005
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.053	0.040	0.030	0.014
	J = 1, *T* = 10	0.038	0.029	0.035	0.013
	J = 5, *T* = 5	0.040	0.018	0.017	0.007
	J = 5, *T* = 10	0.036	0.017	0.018	0.007
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.061	0.050	0.023	0.011
	J = 1, *T* = 10	0.047	0.032	0.021	0.008
	J = 5, *T* = 5	0.040	0.018	0.016	0.007
	J = 5, *T* = 10	0.032	0.014	0.015	0.006
TP = High, *IP* = High	*J* = 1, *T* = 5	0.056	0.040	0.041	0.016
	J = 1, *T* = 10	0.049	0.045	0.047	0.016
	J = 5, *T* = 5	0.049	0.023	0.019	0.008
	J = 5, *T* = 10	0.046	0.023	0.023	0.008

Table A12.

Raw results for RMSE in the estimation of the probability of a correct response (master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.065	0.045	0.023	0.010
	J = 1, *T* = 10	0.029	0.014	0.010	0.004
	J = 5, *T* = 5	0.035	0.017	0.015	0.006
	J = 5, *T* = 10	0.017	0.008	0.009	0.004
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.042	0.020	0.017	0.008
	J = 1, *T* = 10	0.020	0.010	0.009	0.004
	J = 5, *T* = 5	0.027	0.012	0.013	0.005
	J = 5, *T* = 10	0.014	0.007	0.008	0.004
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.060	0.032	0.017	0.008
	J = 1, *T* = 10	0.029	0.013	0.009	0.004
	J = 5, *T* = 5	0.030	0.014	0.013	0.006
	J = 5, *T* = 10	0.015	0.007	0.008	0.003
TP = High, *IP* = High	*J* = 1, *T* = 5	0.040	0.017	0.015	0.007
	J = 1, *T* = 10	0.020	0.010	0.008	0.004
	J = 5, *T* = 5	0.025	0.012	0.012	0.005
	J = 5, *T* = 10	0.014	0.007	0.007	0.003

Table A13.

Raw results for efficiency in the estimation of the initial probability of mastery parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.145	0.104	0.036	0.016
	J = 1, *T* = 10	0.136	0.092	0.034	0.016
	J = 5, *T* = 5	0.098	0.046	0.028	0.013
	J = 5, *T* = 10	0.072	0.033	0.029	0.013
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.137	0.095	0.039	0.018
	J = 1, *T* = 10	0.124	0.098	0.039	0.018
	J = 5, *T* = 5	0.096	0.050	0.027	0.013
	J = 5, *T* = 10	0.095	0.050	0.028	0.012
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.120	0.069	0.042	0.020
	J = 1, *T* = 10	0.131	0.086	0.040	0.019
	J = 5, *T* = 5	0.087	0.050	0.034	0.016
	J = 5, *T* = 10	0.073	0.038	0.035	0.015
TP = High, *IP* = High	*J* = 1, *T* = 5	0.130	0.078	0.047	0.022
	J = 1, *T* = 10	0.128	0.111	0.046	0.020
	J = 5, *T* = 5	0.103	0.059	0.034	0.016
	J = 5, *T* = 10	0.103	0.052	0.035	0.016

Table A14.

Raw results for efficiency in the estimation of the transition probability parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.185	0.090	0.022	0.010
	J = 1, *T* = 10	0.108	0.042	0.017	0.007
	J = 5, *T* = 5	0.060	0.024	0.018	0.008
	J = 5, *T* = 10	0.029	0.013	0.016	0.007
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.173	0.086	0.031	0.014
	J = 1, *T* = 10	0.133	0.063	0.028	0.012
	J = 5, *T* = 5	0.074	0.034	0.026	0.012
	J = 5, *T* = 10	0.054	0.024	0.024	0.011
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.197	0.098	0.026	0.012
	J = 1, *T* = 10	0.138	0.046	0.019	0.008
	J = 5, *T* = 5	0.063	0.027	0.022	0.009
	J = 5, *T* = 10	0.035	0.016	0.018	0.008
TP = High, *IP* = High	*J* = 1, *T* = 5	0.187	0.100	0.037	0.017
	J = 1, *T* = 10	0.167	0.079	0.032	0.015
	J = 5, *T* = 5	0.091	0.041	0.031	0.014
	J = 5, *T* = 10	0.064	0.028	0.028	0.013

Table A15.

Raw results for efficiency in the estimation of the probability of a correct response (non-master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.052	0.032	0.019	0.009
	J = 1, *T* = 10	0.038	0.024	0.015	0.007
	J = 5, *T* = 5	0.034	0.015	0.013	0.006
	J = 5, *T* = 10	0.025	0.012	0.011	0.005
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.049	0.032	0.027	0.014
	J = 1, *T* = 10	0.038	0.028	0.024	0.011
	J = 5, *T* = 5	0.039	0.018	0.016	0.007
	J = 5, *T* = 10	0.035	0.016	0.015	0.007
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.061	0.041	0.022	0.011
	J = 1, *T* = 10	0.047	0.032	0.017	0.008
	J = 5, *T* = 5	0.040	0.018	0.015	0.007
	J = 5, *T* = 10	0.031	0.014	0.013	0.006
TP = High, *IP* = High	*J* = 1, *T* = 5	0.056	0.039	0.033	0.016
	J = 1, *T* = 10	0.042	0.037	0.029	0.014
	J = 5, *T* = 5	0.049	0.023	0.018	0.008
	J = 5, *T* = 10	0.044	0.021	0.018	0.008

Table A16.

Raw results for efficiency in the estimation of the probability of a correct response (master) parameter.

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.059	0.033	0.022	0.010
	J = 1, *T* = 10	0.029	0.014	0.010	0.004
	J = 5, *T* = 5	0.035	0.017	0.014	0.006
	J = 5, *T* = 10	0.017	0.008	0.008	0.004
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.042	0.020	0.017	0.008
	J = 1, *T* = 10	0.020	0.009	0.009	0.004
	J = 5, *T* = 5	0.027	0.012	0.012	0.005
	J = 5, *T* = 10	0.014	0.006	0.008	0.003
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.058	0.028	0.017	0.008
	J = 1, *T* = 10	0.029	0.013	0.009	0.004
	J = 5, *T* = 5	0.030	0.014	0.012	0.006
	J = 5, *T* = 10	0.015	0.007	0.007	0.003
TP = High, *IP* = High	*J* = 1, *T* = 5	0.040	0.017	0.015	0.007
	J = 1, *T* = 10	0.020	0.009	0.008	0.004
	J = 5, *T* = 5	0.025	0.012	0.012	0.005
	J = 5, *T* = 10	0.014	0.006	0.007	0.003

Table A17.

Raw results for classification accuracy (validation set).

		MQ = Low		*MQ* = High	
		N = 200	*N* = 1000	*N* = 200	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	66.49%	67.24%	94.10%	94.03%
	J = 1, *T* = 10	88.06%	89.12%	97.97%	98.00%
	J = 5, *T* = 5	76.16%	77.89%	99.62%	99.64%
	J = 5, *T* = 10	91.54%	91.80%	99.88%	99.88%
TP = High, *IP* = Low	*J* = 1, *T* = 5	87.99%	89.43%	96.81%	96.92%
	J = 1, *T* = 10	99.02%	99.19%	99.73%	99.74%
	J = 5, *T* = 5	89.88%	90.37%	99.84%	99.85%
	J = 5, *T* = 10	99.22%	99.20%	99.99%	99.99%
TP = Low, *IP* = High	*J* = 1, *T* = 5	73.60%	75.19%	95.35%	95.44%
	J = 1, *T* = 10	90.53%	91.86%	98.50%	98.50%
	J = 5, *T* = 5	80.74%	82.32%	99.72%	99.73%
	J = 5, *T* = 10	93.55%	93.71%	99.91%	99.91%
TP = High, *IP* = High	*J* = 1, *T* = 5	90.46%	92.15%	97.54%	97.60%
	J = 1, *T* = 10	99.16%	99.39%	99.80%	99.81%
	J = 5, *T* = 5	92.00%	92.59%	99.89%	99.88%
	J = 5, *T* = 10	99.39%	99.38%	99.99%	99.99%

Note. Red: ≤ 70%, Orange: 70% - 79.99%, Yellow: 80% - 89.99%, Green: ≥ 90%.

APPENDIX B

RAW RESULTS TABLES FOR PHASE 2

Table B1.

Raw results for estimation bias in the initial probability of mastery parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.117	0.132	0.168	0.013	0.003	0.002
	J = 1, *T* = 10	0.016	0.032	0.008	-0.009	-0.004	0.000
	J = 3, *T* = 5	0.049	0.047	0.026	0.003	0.002	0.000
	J = 3, *T* = 10	-0.004	-0.004	-0.007	-0.002	0.000	0.001
	J = 5, *T* = 5	0.017	0.015	0.005	0.001	0.001	0.000
	J = 5, *T* = 10	-0.001	-0.005	-0.005	0.000	-0.001	0.000
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.065	0.078	0.075	-0.021	-0.012	-0.005
	J = 1, *T* = 10	-0.011	-0.011	-0.032	-0.044	-0.024	-0.013
	J = 3, *T* = 5	0.007	0.013	-0.007	-0.003	-0.001	-0.001
	J = 3, *T* = 10	-0.040	-0.043	-0.059	-0.009	-0.005	-0.002
	J = 5, *T* = 5	-0.001	-0.004	-0.008	-0.002	0.000	-0.001
	J = 5, *T* = 10	-0.041	-0.028	-0.025	-0.002	-0.002	-0.001
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.029	0.051	0.033	-0.109	-0.094	-0.086
	J = 1, *T* = 10	-0.027	-0.039	-0.062	-0.171	-0.163	-0.159
	J = 3, *T* = 5	-0.031	-0.023	-0.048	-0.013	-0.008	-0.004
	J = 3, *T* = 10	-0.087	-0.089	-0.127	-0.027	-0.013	-0.007
	J = 5, *T* = 5	-0.034	-0.036	-0.034	-0.007	-0.003	-0.002
	J = 5, *T* = 10	-0.079	-0.081	-0.087	-0.013	-0.008	-0.004
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.029	-0.004	0.017	-0.017	-0.007	0.000
	J = 1, *T* = 10	-0.087	-0.054	-0.029	-0.014	-0.008	-0.001
	J = 3, *T* = 5	-0.038	-0.010	0.004	-0.001	0.002	0.001
	J = 3, *T* = 10	-0.042	-0.020	-0.018	-0.002	-0.001	-0.001
	J = 5, *T* = 5	-0.020	-0.003	-0.002	-0.001	0.000	0.001
	J = 5, *T* = 10	-0.023	-0.011	-0.006	-0.001	0.000	0.000
TP = Med, *IP* = High	*J* = 1, *T* = 5	-0.068	-0.040	-0.029	-0.049	-0.028	-0.011
	J = 1, *T* = 10	-0.121	-0.111	-0.119	-0.060	-0.033	-0.017
	J = 3, *T* = 5	-0.064	-0.042	-0.030	-0.006	-0.003	-0.001
	J = 3, *T* = 10	-0.098	-0.074	-0.069	-0.012	-0.007	-0.004
	J = 5, *T* = 5	-0.053	-0.026	-0.020	-0.005	-0.002	0.000
	J = 5, *T* = 10	-0.064	-0.042	-0.029	-0.006	-0.003	-0.002
TP = High, *IP* = High	*J* = 1, *T* = 5	-0.095	-0.074	-0.095	-0.199	-0.184	-0.163
	J = 1, *T* = 10	-0.143	-0.152	-0.180	-0.284	-0.269	-0.254
	J = 3, *T* = 5	-0.121	-0.103	-0.098	-0.024	-0.012	-0.006
	J = 3, *T* = 10	-0.173	-0.163	-0.176	-0.039	-0.026	-0.011
	J = 5, *T* = 5	-0.110	-0.077	-0.053	-0.013	-0.008	-0.002
	J = 5, *T* = 10	-0.140	-0.110	-0.099	-0.023	-0.011	-0.005

Note. Black cells are duplicate conditions from Phase 1.

Table B2.

Raw results for estimation bias in the transition probability parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.166	0.129	0.102	0.008	0.001	0.002
	J = 1, *T* = 10	0.047	0.025	0.015	0.000	0.001	0.003
	J = 3, *T* = 5	0.041	0.023	0.015	0.001	0.000	0.001
	J = 3, *T* = 10	0.005	0.005	0.007	0.001	0.001	0.001
	J = 5, *T* = 5	0.014	0.008	0.005	0.000	0.001	0.001
	J = 5, *T* = 10	0.004	0.005	0.005	0.002	0.000	0.001
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.061	0.045	0.037	0.000	0.001	-0.001
	J = 1, *T* = 10	0.018	0.007	-0.003	-0.003	-0.001	0.000
	J = 3, *T* = 5	0.024	0.011	0.004	0.002	0.001	0.001
	J = 3, *T* = 10	0.000	0.003	-0.003	0.002	0.001	0.002
	J = 5, *T* = 5	0.011	0.004	0.003	0.002	0.001	0.001
	J = 5, *T* = 10	-0.003	0.001	0.002	0.001	0.001	0.001
TP = High, *IP* = Low	*J* = 1, *T* = 5	-0.079	-0.054	-0.031	-0.010	-0.001	-0.007
	J = 1, *T* = 10	-0.068	-0.050	-0.030	-0.007	-0.003	-0.011
	J = 3, *T* = 5	-0.026	-0.017	-0.011	-0.004	0.000	0.000
	J = 3, *T* = 10	-0.024	-0.011	-0.021	-0.005	-0.001	-0.001
	J = 5, *T* = 5	-0.013	-0.006	-0.009	-0.003	0.001	0.000
	J = 5, *T* = 10	-0.010	-0.011	-0.014	-0.005	0.000	-0.001
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.158	0.118	0.080	-0.002	0.000	0.002
	J = 1, *T* = 10	0.061	0.028	0.018	0.000	0.002	0.003
	J = 3, *T* = 5	0.027	0.015	0.009	-0.001	0.000	0.001
	J = 3, *T* = 10	0.004	0.005	0.008	0.002	0.001	0.000
	J = 5, *T* = 5	0.005	0.004	0.005	0.002	0.001	0.001
	J = 5, *T* = 10	0.003	0.005	0.007	0.003	0.000	0.000
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.047	0.049	0.029	-0.011	-0.008	-0.002
	J = 1, *T* = 10	0.021	0.005	-0.007	-0.010	-0.004	0.000
	J = 3, *T* = 5	0.014	0.007	-0.001	-0.001	0.001	0.002
	J = 3, *T* = 10	0.003	-0.008	-0.006	-0.001	-0.001	0.001
	J = 5, *T* = 5	-0.002	-0.002	0.003	0.001	0.000	0.001
	J = 5, *T* = 10	-0.004	0.000	0.005	0.002	0.001	0.001
TP = High, *IP* = High	*J* = 1, *T* = 5	-0.114	-0.095	-0.074	-0.015	-0.018	-0.016
	J = 1, *T* = 10	-0.110	-0.083	-0.050	-0.014	-0.014	-0.015
	J = 3, *T* = 5	-0.050	-0.038	-0.022	-0.006	-0.002	0.000
	J = 3, *T* = 10	-0.043	-0.034	-0.033	-0.008	-0.003	-0.001
	J = 5, *T* = 5	-0.038	-0.016	-0.015	0.001	0.001	0.000
	J = 5, *T* = 10	-0.028	-0.012	-0.019	-0.006	0.000	0.001

Note. Black cells are duplicate conditions from Phase 1.

Table B3.

Raw results for estimation bias in the probability of a correct response (non-master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-0.029	-0.037	-0.044	-0.004	0.000	-0.001
	J = 1, *T* = 10	-0.013	-0.012	-0.006	0.007	0.004	-0.001
	J = 3, *T* = 5	-0.012	-0.009	-0.006	0.002	0.001	0.000
	J = 3, *T* = 10	0.002	0.000	0.000	0.003	0.002	0.001
	J = 5, *T* = 5	-0.003	-0.002	-0.001	0.002	0.002	0.000
	J = 5, *T* = 10	0.001	0.000	0.000	0.004	0.002	0.000
TP = Med, *IP* = Low	*J* = 1, *T* = 5	-0.021	-0.024	-0.023	0.013	0.007	0.002
	J = 1, *T* = 10	0.002	0.001	0.005	0.030	0.015	0.007
	J = 3, *T* = 5	-0.003	-0.004	0.001	0.005	0.003	0.001
	J = 3, *T* = 10	0.009	0.007	0.011	0.009	0.005	0.002
	J = 5, *T* = 5	0.001	0.000	0.001	0.004	0.001	0.001
	J = 5, *T* = 10	0.009	0.005	0.005	0.008	0.004	0.002
TP = High, *IP* = Low	*J* = 1, *T* = 5	-0.009	-0.015	-0.010	0.062	0.052	0.046
	J = 1, *T* = 10	0.013	0.011	0.014	0.093	0.087	0.083
	J = 3, *T* = 5	0.006	0.004	0.009	0.012	0.004	0.003
	J = 3, *T* = 10	0.019	0.020	0.025	0.025	0.012	0.007
	J = 5, *T* = 5	0.005	0.007	0.008	0.010	0.004	0.002
	J = 5, *T* = 10	0.019	0.017	0.018	0.021	0.010	0.004
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.006	-0.015	-0.027	0.007	0.003	-0.001
	J = 1, *T* = 10	0.003	-0.002	-0.004	0.011	0.005	0.000
	J = 3, *T* = 5	-0.008	-0.005	-0.004	0.004	0.001	0.000
	J = 3, *T* = 10	0.006	0.003	0.001	0.005	0.003	0.001
	J = 5, *T* = 5	0.000	-0.003	-0.001	0.003	0.001	0.001
	J = 5, *T* = 10	0.004	0.001	0.001	0.006	0.003	0.001
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.005	0.001	0.000	0.031	0.016	0.007
	J = 1, *T* = 10	0.025	0.020	0.024	0.046	0.023	0.010
	J = 3, *T* = 5	0.006	0.005	0.005	0.009	0.004	0.001
	J = 3, *T* = 10	0.019	0.016	0.015	0.017	0.008	0.003
	J = 5, *T* = 5	0.006	0.004	0.003	0.006	0.003	0.000
	J = 5, *T* = 10	0.013	0.008	0.005	0.011	0.006	0.003
TP = High, *IP* = High	*J* = 1, *T* = 5	0.021	0.019	0.021	0.124	0.111	0.098
	J = 1, *T* = 10	0.040	0.042	0.045	0.160	0.150	0.141
	J = 3, *T* = 5	0.021	0.020	0.020	0.022	0.011	0.004
	J = 3, *T* = 10	0.040	0.036	0.037	0.037	0.021	0.010
	J = 5, *T* = 5	0.021	0.015	0.011	0.015	0.007	0.002
	J = 5, *T* = 10	0.031	0.024	0.024	0.027	0.015	0.006

Note. Black cells are duplicate conditions from Phase 1.

Table B4.

Raw results for estimation bias in the probability of a correct response (master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-0.027	-0.028	-0.033	-0.008	-0.002	-0.001
	J = 1, *T* = 10	-0.003	-0.004	0.000	-0.001	0.001	0.002
	J = 3, *T* = 5	-0.011	-0.009	-0.005	-0.003	-0.002	0.000
	J = 3, *T* = 10	0.000	0.000	0.003	0.000	0.000	0.000
	J = 5, *T* = 5	-0.004	-0.003	0.000	-0.004	-0.002	0.000
	J = 5, *T* = 10	0.000	0.001	0.002	-0.001	-0.001	0.000
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.001	-0.002	-0.004	0.001	0.000	0.002
	J = 1, *T* = 10	0.001	0.002	0.003	0.002	0.002	0.003
	J = 3, *T* = 5	-0.002	-0.001	0.002	-0.002	0.000	0.000
	J = 3, *T* = 10	0.002	0.001	0.004	-0.001	0.000	0.000
	J = 5, *T* = 5	0.000	0.000	0.002	-0.002	-0.001	0.000
	J = 5, *T* = 10	0.001	0.002	0.002	-0.001	0.000	0.000
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.010	0.006	0.004	0.001	0.001	0.001
	J = 1, *T* = 10	0.002	0.002	0.002	-0.001	0.000	0.002
	J = 3, *T* = 5	0.002	0.002	0.001	0.000	0.001	0.001
	J = 3, *T* = 10	0.000	0.000	0.004	0.000	0.000	0.001
	J = 5, *T* = 5	0.001	0.001	0.002	-0.001	0.000	0.000
	J = 5, *T* = 10	0.000	0.001	0.003	-0.001	0.000	0.000
TP = Low, *IP* = High	*J* = 1, *T* = 5	-0.014	-0.015	-0.017	0.000	0.001	0.001
	J = 1, *T* = 10	-0.001	0.000	0.001	0.001	0.002	0.002
	J = 3, *T* = 5	-0.002	-0.003	-0.001	-0.002	-0.001	0.000
	J = 3, *T* = 10	0.002	0.001	0.003	-0.001	0.000	0.000
	J = 5, *T* = 5	0.000	0.000	0.001	-0.002	0.000	0.000
	J = 5, *T* = 10	0.000	0.002	0.002	-0.001	-0.001	0.000
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.006	0.001	0.000	0.005	0.004	0.003
	J = 1, *T* = 10	0.003	0.003	0.004	0.002	0.003	0.003
	J = 3, *T* = 5	0.001	0.001	0.003	0.000	-0.001	0.001
	J = 3, *T* = 10	0.002	0.003	0.005	0.000	0.001	0.000
	J = 5, *T* = 5	0.003	0.001	0.002	-0.001	0.000	0.000
	J = 5, *T* = 10	0.002	0.003	0.003	-0.001	0.000	0.000
TP = High, *IP* = High	*J* = 1, *T* = 5	0.010	0.007	0.005	0.002	0.002	0.003
	J = 1, *T* = 10	0.003	0.002	0.003	0.000	0.001	0.002
	J = 3, *T* = 5	0.004	0.002	0.003	0.001	0.001	0.001
	J = 3, *T* = 10	0.001	0.001	0.003	0.000	0.001	0.001
	J = 5, *T* = 5	0.003	0.002	0.002	0.000	0.000	0.000
	J = 5, *T* = 10	0.001	0.001	0.003	0.000	0.001	0.000

Note. Black cells are duplicate conditions from Phase 1.

Table B5.

Raw results for relative estimation bias in the initial probability of mastery parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	58.38%	66.03%	84.10%	6.34%	1.70%	0.89%
	J = 1, *T* = 10	7.92%	16.11%	4.16%	-4.43%	-2.11%	0.13%
	J = 3, *T* = 5	24.37%	23.66%	12.84%	1.36%	0.98%	0.03%
	J = 3, *T* = 10	-1.89%	-1.88%	-3.70%	-0.78%	-0.17%	0.27%
	J = 5, *T* = 5	8.62%	7.64%	2.55%	0.62%	0.52%	-0.21%
	J = 5, *T* = 10	-0.72%	-2.60%	-2.62%	0.11%	-0.52%	-0.20%
TP = Med, *IP* = Low	*J* = 1, *T* = 5	32.58%	38.84%	37.31%	-10.60%	-6.03%	-2.31%
	J = 1, *T* = 10	-5.55%	-5.59%	-15.87%	-22.12%	-12.05%	-6.26%
	J = 3, *T* = 5	3.36%	6.72%	-3.68%	-1.36%	-0.40%	-0.41%
	J = 3, *T* = 10	-19.99%	-21.38%	-29.39%	-4.44%	-2.38%	-0.93%
	J = 5, *T* = 5	-0.74%	-2.13%	-4.04%	-1.24%	-0.10%	-0.40%
	J = 5, *T* = 10	-20.40%	-14.02%	-12.26%	-1.11%	-1.04%	-0.52%
TP = High, *IP* = Low	*J* = 1, *T* = 5	14.73%	25.48%	16.29%	-54.71%	-47.11%	-42.78%
	J = 1, *T* = 10	-13.46%	-19.36%	-30.81%	-85.47%	-81.74%	-79.59%
	J = 3, *T* = 5	-15.30%	-11.46%	-23.93%	-6.33%	-3.76%	-1.81%
	J = 3, *T* = 10	-43.33%	-44.73%	-63.29%	-13.39%	-6.67%	-3.32%
	J = 5, *T* = 5	-17.12%	-18.11%	-16.84%	-3.64%	-1.61%	-0.88%
	J = 5, *T* = 10	-39.75%	-40.58%	-43.32%	-6.62%	-3.84%	-1.87%
TP = Low, *IP* = High	*J* = 1, *T* = 5	-7.30%	-0.94%	4.36%	-4.14%	-1.67%	-0.02%
	J = 1, *T* = 10	-21.75%	-13.44%	-7.36%	-3.58%	-1.99%	-0.25%
	J = 3, *T* = 5	-9.51%	-2.56%	1.10%	-0.13%	0.39%	0.33%
	J = 3, *T* = 10	-10.42%	-5.12%	-4.56%	-0.47%	-0.17%	-0.22%
	J = 5, *T* = 5	-5.08%	-0.83%	-0.54%	-0.19%	0.09%	0.13%
	J = 5, *T* = 10	-5.82%	-2.73%	-1.56%	-0.32%	-0.01%	0.03%
TP = Med, *IP* = High	*J* = 1, *T* = 5	-16.98%	-9.91%	-7.21%	-12.17%	-7.05%	-2.80%
	J = 1, *T* = 10	-30.15%	-27.79%	-29.70%	-14.93%	-8.16%	-4.27%
	J = 3, *T* = 5	-16.00%	-10.46%	-7.42%	-1.52%	-0.70%	-0.15%
	J = 3, *T* = 10	-24.62%	-18.52%	-17.20%	-3.04%	-1.78%	-0.93%
	J = 5, *T* = 5	-13.37%	-6.58%	-4.92%	-1.28%	-0.53%	-0.07%
	J = 5, *T* = 10	-15.92%	-10.48%	-7.30%	-1.48%	-0.82%	-0.53%
TP = High, *IP* = High	*J* = 1, *T* = 5	-23.86%	-18.60%	-23.63%	-49.80%	-45.89%	-40.63%
	J = 1, *T* = 10	-35.69%	-37.94%	-45.05%	-70.88%	-67.13%	-63.52%
	J = 3, *T* = 5	-30.31%	-25.85%	-24.45%	-5.97%	-3.10%	-1.56%
	J = 3, *T* = 10	-43.20%	-40.69%	-44.10%	-9.87%	-6.52%	-2.76%
	J = 5, *T* = 5	-27.42%	-19.32%	-13.26%	-3.16%	-1.90%	-0.47%
	J = 5, *T* = 10	-35.02%	-27.60%	-24.84%	-5.78%	-2.63%	-1.13%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red). Black cells are duplicate conditions from Phase 1.

Table B6.

Raw results for relative estimation bias in the transition probability parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	83.11%	64.57%	50.87%	3.85%	0.39%	1.06%
	J = 1, *T* = 10	23.45%	12.65%	7.32%	0.16%	0.58%	1.51%
	J = 3, *T* = 5	20.32%	11.36%	7.31%	0.63%	0.09%	0.58%
	J = 3, *T* = 10	2.40%	2.31%	3.35%	0.38%	0.31%	0.44%
	J = 5, *T* = 5	6.91%	4.11%	2.69%	0.07%	0.33%	0.51%
	J = 5, *T* = 10	1.77%	2.28%	2.64%	0.83%	0.19%	0.31%
TP = Med, *IP* = Low	*J* = 1, *T* = 5	15.23%	11.34%	9.29%	-0.06%	0.37%	-0.29%
	J = 1, *T* = 10	4.47%	1.85%	-0.78%	-0.82%	-0.26%	-0.10%
	J = 3, *T* = 5	6.02%	2.77%	1.07%	0.39%	0.33%	0.33%
	J = 3, *T* = 10	-0.02%	0.67%	-0.65%	0.50%	0.14%	0.38%
	J = 5, *T* = 5	2.70%	1.11%	0.70%	0.43%	0.26%	0.31%
	J = 5, *T* = 10	-0.74%	0.21%	0.48%	0.29%	0.18%	0.36%
TP = High, *IP* = Low	*J* = 1, *T* = 5	-9.84%	-6.77%	-3.82%	-1.27%	-0.13%	-0.89%
	J = 1, *T* = 10	-8.48%	-6.28%	-3.71%	-0.83%	-0.33%	-1.38%
	J = 3, *T* = 5	-3.28%	-2.08%	-1.40%	-0.49%	-0.04%	0.01%
	J = 3, *T* = 10	-2.97%	-1.33%	-2.61%	-0.64%	-0.15%	-0.15%
	J = 5, *T* = 5	-1.67%	-0.76%	-1.13%	-0.37%	0.10%	0.03%
	J = 5, *T* = 10	-1.31%	-1.39%	-1.76%	-0.60%	-0.04%	-0.08%
TP = Low, *IP* = High	*J* = 1, *T* = 5	78.93%	58.91%	40.08%	-0.87%	-0.22%	0.97%
	J = 1, *T* = 10	30.74%	14.23%	9.17%	-0.21%	0.91%	1.29%
	J = 3, *T* = 5	13.30%	7.26%	4.26%	-0.56%	-0.11%	0.52%
	J = 3, *T* = 10	1.97%	2.26%	3.98%	1.08%	0.42%	0.15%
	J = 5, *T* = 5	2.37%	1.83%	2.42%	1.03%	0.47%	0.45%
	J = 5, *T* = 10	1.26%	2.34%	3.61%	1.30%	0.16%	-0.03%
TP = Med, *IP* = High	*J* = 1, *T* = 5	11.72%	12.23%	7.23%	-2.81%	-1.88%	-0.60%
	J = 1, *T* = 10	5.30%	1.19%	-1.72%	-2.61%	-1.09%	-0.05%
	J = 3, *T* = 5	3.56%	1.84%	-0.16%	-0.22%	0.28%	0.39%
	J = 3, *T* = 10	0.70%	-2.07%	-1.41%	-0.15%	-0.22%	0.18%
	J = 5, *T* = 5	-0.62%	-0.46%	0.67%	0.20%	0.11%	0.18%
	J = 5, *T* = 10	-1.02%	-0.11%	1.25%	0.42%	0.31%	0.23%
TP = High, *IP* = High	*J* = 1, *T* = 5	-14.24%	-11.92%	-9.22%	-1.88%	-2.26%	-1.98%
	J = 1, *T* = 10	-13.73%	-10.38%	-6.22%	-1.75%	-1.78%	-1.82%
	J = 3, *T* = 5	-6.28%	-4.79%	-2.77%	-0.77%	-0.26%	0.03%
	J = 3, *T* = 10	-5.39%	-4.23%	-4.18%	-0.97%	-0.43%	-0.12%
	J = 5, *T* = 5	-4.73%	-2.04%	-1.92%	0.09%	0.13%	-0.02%
	J = 5, *T* = 10	-3.47%	-1.56%	-2.39%	-0.74%	-0.01%	0.07%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red). Black cells are duplicate conditions from Phase 1.

Table B7.

Raw results for relative estimation bias in the probability of a correct response (non-master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-7.19%	-9.21%	-10.90%	-1.51%	0.06%	-0.36%
	J = 1, *T* = 10	-3.19%	-3.12%	-1.46%	2.98%	1.47%	-0.38%
	J = 3, *T* = 5	-3.09%	-2.31%	-1.43%	0.94%	0.56%	-0.07%
	J = 3, *T* = 10	0.59%	-0.04%	0.02%	1.14%	0.74%	0.33%
	J = 5, *T* = 5	-0.87%	-0.49%	-0.36%	0.66%	0.76%	-0.07%
	J = 5, *T* = 10	0.22%	-0.04%	0.05%	1.43%	0.96%	0.14%
TP = Med, *IP* = Low	*J* = 1, *T* = 5	-5.24%	-6.09%	-5.70%	5.35%	2.77%	0.86%
	J = 1, *T* = 10	0.41%	0.14%	1.35%	11.96%	5.83%	2.90%
	J = 3, *T* = 5	-0.87%	-0.90%	0.22%	1.84%	1.10%	0.24%
	J = 3, *T* = 10	2.14%	1.80%	2.76%	3.74%	2.12%	0.64%
	J = 5, *T* = 5	0.26%	-0.07%	0.28%	1.69%	0.59%	0.37%
	J = 5, *T* = 10	2.29%	1.26%	1.31%	3.33%	1.53%	0.61%
TP = High, *IP* = Low	*J* = 1, *T* = 5	-2.23%	-3.82%	-2.41%	24.73%	20.91%	18.27%
	J = 1, *T* = 10	3.18%	2.84%	3.46%	37.37%	34.92%	33.18%
	J = 3, *T* = 5	1.61%	1.06%	2.29%	4.99%	1.78%	1.37%
	J = 3, *T* = 10	4.82%	4.93%	6.25%	9.83%	4.82%	2.69%
	J = 5, *T* = 5	1.36%	1.78%	1.94%	4.15%	1.53%	0.73%
	J = 5, *T* = 10	4.72%	4.35%	4.51%	8.24%	3.93%	1.77%
TP = Low, *IP* = High	*J* = 1, *T* = 5	-1.38%	-3.77%	-6.75%	2.77%	1.36%	-0.36%
	J = 1, *T* = 10	0.80%	-0.47%	-0.95%	4.26%	1.86%	0.11%
	J = 3, *T* = 5	-2.03%	-1.25%	-1.00%	1.40%	0.46%	-0.01%
	J = 3, *T* = 10	1.40%	0.63%	0.30%	2.12%	1.12%	0.54%
	J = 5, *T* = 5	-0.06%	-0.72%	-0.16%	1.33%	0.46%	0.48%
	J = 5, *T* = 10	1.08%	0.28%	0.14%	2.22%	1.28%	0.36%
TP = Med, *IP* = High	*J* = 1, *T* = 5	1.17%	0.31%	-0.10%	12.29%	6.43%	2.93%
	J = 1, *T* = 10	6.21%	5.10%	6.06%	18.53%	9.37%	4.07%
	J = 3, *T* = 5	1.55%	1.25%	1.26%	3.72%	1.50%	0.38%
	J = 3, *T* = 10	4.69%	4.11%	3.75%	6.95%	3.36%	1.01%
	J = 5, *T* = 5	1.59%	0.92%	0.76%	2.35%	1.30%	0.13%
	J = 5, *T* = 10	3.32%	1.98%	1.20%	4.25%	2.42%	1.02%
TP = High, *IP* = High	*J* = 1, *T* = 5	5.32%	4.66%	5.36%	49.56%	44.58%	39.13%
	J = 1, *T* = 10	10.06%	10.55%	11.35%	63.80%	59.98%	56.41%
	J = 3, *T* = 5	5.33%	5.05%	4.93%	8.87%	4.45%	1.65%
	J = 3, *T* = 10	10.01%	9.04%	9.34%	14.64%	8.52%	3.93%
	J = 5, *T* = 5	5.21%	3.79%	2.69%	5.85%	2.77%	0.91%
	J = 5, *T* = 10	7.80%	6.09%	6.02%	10.99%	5.90%	2.56%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red). Black cells are duplicate conditions from Phase 1.

Table B8.

Raw results for relative estimation bias in the probability of a correct response (master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	-4.52%	-4.59%	-5.49%	-1.11%	-0.28%	-0.14%
	J = 1, *T* = 10	-0.45%	-0.65%	-0.01%	-0.10%	0.10%	0.22%
	J = 3, *T* = 5	-1.87%	-1.45%	-0.75%	-0.39%	-0.28%	-0.04%
	J = 3, *T* = 10	0.00%	0.07%	0.46%	-0.06%	-0.03%	0.02%
	J = 5, *T* = 5	-0.71%	-0.45%	0.07%	-0.49%	-0.24%	-0.01%
	J = 5, *T* = 10	-0.03%	0.16%	0.28%	-0.16%	-0.12%	-0.02%
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.10%	-0.29%	-0.66%	0.13%	0.04%	0.22%
	J = 1, *T* = 10	0.24%	0.33%	0.56%	0.24%	0.26%	0.44%
	J = 3, *T* = 5	-0.31%	-0.18%	0.25%	-0.22%	-0.02%	0.05%
	J = 3, *T* = 10	0.28%	0.22%	0.68%	-0.07%	0.04%	0.02%
	J = 5, *T* = 5	-0.01%	-0.04%	0.27%	-0.22%	-0.08%	0.04%
	J = 5, *T* = 10	0.24%	0.29%	0.38%	-0.17%	-0.06%	0.01%
TP = High, *IP* = Low	*J* = 1, *T* = 5	1.62%	0.92%	0.58%	0.11%	0.07%	0.14%
	J = 1, *T* = 10	0.36%	0.32%	0.36%	-0.12%	0.07%	0.33%
	J = 3, *T* = 5	0.32%	0.33%	0.24%	-0.01%	0.07%	0.12%
	J = 3, *T* = 10	0.06%	0.08%	0.63%	0.01%	0.03%	0.13%
	J = 5, *T* = 5	0.21%	0.17%	0.32%	-0.16%	0.02%	0.00%
	J = 5, *T* = 10	0.06%	0.20%	0.54%	-0.07%	0.06%	0.06%
TP = Low, *IP* = High	*J* = 1, *T* = 5	-2.28%	-2.50%	-2.88%	-0.03%	0.14%	0.14%
	J = 1, *T* = 10	-0.24%	-0.03%	0.22%	0.12%	0.24%	0.23%
	J = 3, *T* = 5	-0.41%	-0.52%	-0.15%	-0.27%	-0.08%	-0.06%
	J = 3, *T* = 10	0.28%	0.23%	0.58%	-0.17%	-0.04%	0.04%
	J = 5, *T* = 5	-0.04%	-0.08%	0.23%	-0.24%	-0.06%	0.04%
	J = 5, *T* = 10	0.02%	0.26%	0.29%	-0.20%	-0.12%	-0.02%
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.99%	0.09%	-0.06%	0.71%	0.57%	0.37%
	J = 1, *T* = 10	0.50%	0.48%	0.71%	0.33%	0.35%	0.42%
	J = 3, *T* = 5	0.19%	0.17%	0.42%	-0.05%	-0.07%	0.07%
	J = 3, *T* = 10	0.28%	0.47%	0.77%	0.04%	0.10%	0.02%
	J = 5, *T* = 5	0.45%	0.14%	0.31%	-0.10%	0.00%	0.00%
	J = 5, *T* = 10	0.30%	0.43%	0.47%	-0.08%	-0.05%	-0.01%
TP = High, *IP* = High	*J* = 1, *T* = 5	1.70%	1.13%	0.88%	0.26%	0.31%	0.34%
	J = 1, *T* = 10	0.54%	0.42%	0.43%	-0.02%	0.17%	0.30%
	J = 3, *T* = 5	0.72%	0.34%	0.45%	0.17%	0.07%	0.13%
	J = 3, *T* = 10	0.22%	0.22%	0.58%	0.05%	0.08%	0.14%
	J = 5, *T* = 5	0.54%	0.29%	0.37%	-0.02%	-0.01%	0.03%
	J = 5, *T* = 10	0.23%	0.25%	0.53%	0.04%	0.07%	0.04%

Note. Absolute relative bias < 5% (green), 5.01% - 10% (yellow), > 10% (red). Black cells are duplicate conditions from Phase 1.

Table B9.

Raw results for RMSE in the estimation of the initial probability of mastery parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.187	0.187	0.194	0.078	0.059	0.038
	J = 1, *T* = 10	0.138	0.125	0.093	0.062	0.043	0.027
	J = 3, *T* = 5	0.141	0.116	0.072	0.038	0.028	0.017
	J = 3, *T* = 10	0.098	0.075	0.049	0.038	0.026	0.017
	J = 5, *T* = 5	0.097	0.074	0.044	0.033	0.023	0.015
	J = 5, *T* = 10	0.071	0.054	0.033	0.033	0.023	0.014
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.155	0.145	0.119	0.091	0.068	0.047
	J = 1, *T* = 10	0.123	0.115	0.105	0.096	0.068	0.044
	J = 3, *T* = 5	0.123	0.101	0.070	0.043	0.029	0.019
	J = 3, *T* = 10	0.119	0.106	0.094	0.042	0.029	0.019
	J = 5, *T* = 5	0.100	0.078	0.050	0.035	0.025	0.015
	J = 5, *T* = 10	0.104	0.079	0.055	0.035	0.024	0.015
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.116	0.102	0.081	0.133	0.122	0.118
	J = 1, *T* = 10	0.105	0.095	0.098	0.178	0.174	0.172
	J = 3, *T* = 5	0.123	0.111	0.098	0.051	0.038	0.023
	J = 3, *T* = 10	0.140	0.138	0.154	0.056	0.037	0.023
	J = 5, *T* = 5	0.114	0.098	0.076	0.040	0.027	0.017
	J = 5, *T* = 10	0.134	0.126	0.122	0.039	0.028	0.017
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.124	0.101	0.075	0.078	0.059	0.046
	J = 1, *T* = 10	0.157	0.124	0.095	0.065	0.048	0.031
	J = 3, *T* = 5	0.109	0.080	0.065	0.042	0.033	0.020
	J = 3, *T* = 10	0.105	0.077	0.059	0.041	0.031	0.019
	J = 5, *T* = 5	0.087	0.069	0.050	0.037	0.028	0.018
	J = 5, *T* = 10	0.081	0.058	0.039	0.040	0.028	0.018
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.147	0.113	0.085	0.109	0.082	0.055
	J = 1, *T* = 10	0.176	0.162	0.162	0.110	0.074	0.048
	J = 3, *T* = 5	0.134	0.108	0.081	0.049	0.035	0.023
	J = 3, *T* = 10	0.166	0.133	0.111	0.050	0.035	0.022
	J = 5, *T* = 5	0.116	0.084	0.064	0.041	0.031	0.019
	J = 5, *T* = 10	0.125	0.091	0.060	0.043	0.029	0.019
TP = High, *IP* = High	*J* = 1, *T* = 5	0.148	0.115	0.122	0.229	0.215	0.197
	J = 1, *T* = 10	0.182	0.182	0.199	0.303	0.292	0.283
	J = 3, *T* = 5	0.178	0.158	0.148	0.061	0.044	0.029
	J = 3, *T* = 10	0.222	0.215	0.226	0.071	0.050	0.030
	J = 5, *T* = 5	0.167	0.133	0.102	0.047	0.034	0.021
	J = 5, *T* = 10	0.198	0.171	0.147	0.050	0.033	0.021

Note. Black cells are duplicate conditions from Phase 1.

Table B10.

Raw results for RMSE in the in the estimation of the transition probability parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.254	0.195	0.135	0.046	0.031	0.020
	J = 1, *T* = 10	0.122	0.073	0.043	0.024	0.017	0.012
	J = 3, *T* = 5	0.104	0.065	0.040	0.024	0.016	0.011
	J = 3, *T* = 10	0.043	0.029	0.020	0.017	0.011	0.008
	J = 5, *T* = 5	0.061	0.040	0.024	0.020	0.015	0.010
	J = 5, *T* = 10	0.030	0.021	0.014	0.016	0.011	0.007
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.191	0.143	0.095	0.060	0.046	0.030
	J = 1, *T* = 10	0.142	0.104	0.064	0.043	0.031	0.021
	J = 3, *T* = 5	0.115	0.076	0.048	0.033	0.024	0.015
	J = 3, *T* = 10	0.076	0.052	0.034	0.029	0.020	0.013
	J = 5, *T* = 5	0.075	0.055	0.034	0.028	0.021	0.013
	J = 5, *T* = 10	0.055	0.037	0.023	0.026	0.020	0.012
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.204	0.150	0.108	0.086	0.065	0.041
	J = 1, *T* = 10	0.177	0.140	0.099	0.080	0.060	0.039
	J = 3, *T* = 5	0.121	0.096	0.062	0.044	0.030	0.020
	J = 3, *T* = 10	0.120	0.086	0.059	0.044	0.030	0.019
	J = 5, *T* = 5	0.099	0.073	0.048	0.036	0.025	0.016
	J = 5, *T* = 10	0.091	0.067	0.044	0.036	0.025	0.015
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.243	0.193	0.128	0.048	0.033	0.023
	J = 1, *T* = 10	0.154	0.092	0.051	0.029	0.020	0.013
	J = 3, *T* = 5	0.118	0.064	0.040	0.028	0.020	0.012
	J = 3, *T* = 10	0.051	0.035	0.024	0.020	0.013	0.009
	J = 5, *T* = 5	0.060	0.043	0.027	0.024	0.018	0.011
	J = 5, *T* = 10	0.034	0.025	0.018	0.020	0.014	0.008
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.194	0.161	0.108	0.067	0.050	0.035
	J = 1, *T* = 10	0.161	0.120	0.079	0.053	0.037	0.023
	J = 3, *T* = 5	0.127	0.091	0.055	0.038	0.027	0.018
	J = 3, *T* = 10	0.100	0.063	0.040	0.033	0.023	0.014
	J = 5, *T* = 5	0.090	0.062	0.041	0.035	0.024	0.015
	J = 5, *T* = 10	0.069	0.046	0.030	0.030	0.021	0.014
TP = High, *IP* = High	*J* = 1, *T* = 5	0.231	0.189	0.141	0.108	0.086	0.055
	J = 1, *T* = 10	0.220	0.179	0.120	0.104	0.080	0.049
	J = 3, *T* = 5	0.159	0.122	0.082	0.054	0.038	0.023
	J = 3, *T* = 10	0.148	0.111	0.078	0.053	0.035	0.022
	J = 5, *T* = 5	0.126	0.091	0.060	0.040	0.029	0.018
	J = 5, *T* = 10	0.120	0.086	0.056	0.042	0.029	0.018

Note. Black cells are duplicate conditions from Phase 1.

Table B11.

Raw results for RMSE in the estimation of the probability of a correct response (non-master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.060	0.059	0.055	0.040	0.030	0.020
	J = 1, *T* = 10	0.041	0.035	0.025	0.030	0.022	0.014
	J = 3, *T* = 5	0.045	0.033	0.020	0.022	0.015	0.010
	J = 3, *T* = 10	0.029	0.022	0.014	0.018	0.014	0.008
	J = 5, *T* = 5	0.034	0.023	0.015	0.021	0.014	0.009
	J = 5, *T* = 10	0.026	0.019	0.012	0.018	0.013	0.008
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.054	0.049	0.039	0.053	0.039	0.028
	J = 1, *T* = 10	0.038	0.034	0.030	0.053	0.038	0.026
	J = 3, *T* = 5	0.046	0.035	0.022	0.028	0.020	0.012
	J = 3, *T* = 10	0.037	0.031	0.024	0.027	0.019	0.012
	J = 5, *T* = 5	0.038	0.029	0.018	0.026	0.018	0.012
	J = 5, *T* = 10	0.036	0.025	0.017	0.024	0.017	0.011
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.043	0.038	0.030	0.078	0.069	0.066
	J = 1, *T* = 10	0.035	0.031	0.028	0.100	0.094	0.090
	J = 3, *T* = 5	0.046	0.038	0.027	0.040	0.029	0.019
	J = 3, *T* = 10	0.047	0.040	0.035	0.043	0.030	0.019
	J = 5, *T* = 5	0.044	0.035	0.024	0.036	0.025	0.016
	J = 5, *T* = 10	0.047	0.037	0.031	0.039	0.025	0.016
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.063	0.057	0.049	0.046	0.035	0.026
	J = 1, *T* = 10	0.047	0.042	0.032	0.036	0.026	0.016
	J = 3, *T* = 5	0.054	0.037	0.025	0.026	0.019	0.012
	J = 3, *T* = 10	0.039	0.027	0.018	0.022	0.015	0.010
	J = 5, *T* = 5	0.040	0.029	0.018	0.024	0.017	0.011
	J = 5, *T* = 10	0.032	0.022	0.014	0.021	0.015	0.009
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.056	0.047	0.035	0.067	0.055	0.038
	J = 1, *T* = 10	0.048	0.045	0.044	0.071	0.049	0.032
	J = 3, *T* = 5	0.057	0.043	0.029	0.036	0.025	0.016
	J = 3, *T* = 10	0.053	0.040	0.031	0.035	0.024	0.015
	J = 5, *T* = 5	0.052	0.037	0.025	0.030	0.021	0.013
	J = 5, *T* = 10	0.043	0.033	0.022	0.029	0.020	0.013
TP = High, *IP* = High	*J* = 1, *T* = 5	0.051	0.043	0.041	0.137	0.129	0.118
	J = 1, *T* = 10	0.053	0.053	0.052	0.167	0.159	0.154
	J = 3, *T* = 5	0.064	0.054	0.043	0.051	0.037	0.023
	J = 3, *T* = 10	0.063	0.057	0.053	0.057	0.040	0.024
	J = 5, *T* = 5	0.060	0.048	0.034	0.045	0.030	0.019
	J = 5, *T* = 10	0.060	0.050	0.040	0.047	0.032	0.019

Note. Black cells are duplicate conditions from Phase 1.

Table B12.

Raw results for RMSE in the estimation of the probability of a correct response (master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.067	0.060	0.044	0.044	0.034	0.022
	J = 1, *T* = 10	0.030	0.022	0.014	0.017	0.013	0.008
	J = 3, *T* = 5	0.047	0.035	0.022	0.024	0.017	0.011
	J = 3, *T* = 10	0.019	0.014	0.009	0.013	0.009	0.006
	J = 5, *T* = 5	0.037	0.026	0.016	0.022	0.015	0.010
	J = 5, *T* = 10	0.017	0.012	0.008	0.012	0.009	0.006
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.044	0.034	0.020	0.031	0.024	0.015
	J = 1, *T* = 10	0.020	0.014	0.009	0.014	0.011	0.008
	J = 3, *T* = 5	0.031	0.021	0.014	0.021	0.014	0.009
	J = 3, *T* = 10	0.015	0.011	0.008	0.011	0.008	0.005
	J = 5, *T* = 5	0.027	0.018	0.012	0.018	0.013	0.008
	J = 5, *T* = 10	0.014	0.010	0.007	0.011	0.008	0.005
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.029	0.019	0.012	0.020	0.015	0.009
	J = 1, *T* = 10	0.013	0.009	0.007	0.011	0.008	0.006
	J = 3, *T* = 5	0.020	0.014	0.009	0.017	0.011	0.007
	J = 3, *T* = 10	0.012	0.009	0.007	0.010	0.008	0.005
	J = 5, *T* = 5	0.019	0.014	0.009	0.016	0.011	0.007
	J = 5, *T* = 10	0.012	0.009	0.007	0.010	0.007	0.005
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.055	0.048	0.033	0.034	0.026	0.018
	J = 1, *T* = 10	0.030	0.021	0.012	0.016	0.013	0.008
	J = 3, *T* = 5	0.039	0.027	0.018	0.021	0.015	0.009
	J = 3, *T* = 10	0.018	0.013	0.009	0.012	0.008	0.005
	J = 5, *T* = 5	0.029	0.022	0.014	0.019	0.013	0.008
	J = 5, *T* = 10	0.016	0.011	0.007	0.011	0.009	0.005
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.044	0.032	0.018	0.027	0.021	0.014
	J = 1, *T* = 10	0.020	0.014	0.010	0.014	0.010	0.007
	J = 3, *T* = 5	0.031	0.021	0.013	0.019	0.013	0.008
	J = 3, *T* = 10	0.015	0.011	0.009	0.011	0.008	0.005
	J = 5, *T* = 5	0.027	0.018	0.012	0.018	0.013	0.008
	J = 5, *T* = 10	0.014	0.010	0.007	0.011	0.008	0.005
TP = High, *IP* = High	*J* = 1, *T* = 5	0.028	0.019	0.012	0.021	0.016	0.010
	J = 1, *T* = 10	0.015	0.010	0.007	0.011	0.009	0.006
	J = 3, *T* = 5	0.021	0.014	0.010	0.016	0.012	0.008
	J = 3, *T* = 10	0.012	0.009	0.007	0.011	0.007	0.005
	J = 5, *T* = 5	0.020	0.013	0.009	0.015	0.011	0.007
	J = 5, *T* = 10	0.012	0.009	0.007	0.010	0.008	0.005

Note. Black cells are duplicate conditions from Phase 1.

Table B13.

Raw results for efficiency in the estimation of the initial probability of mastery parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.147	0.133	0.096	0.077	0.058	0.038
	J = 1, *T* = 10	0.137	0.121	0.093	0.062	0.043	0.027
	J = 3, *T* = 5	0.132	0.106	0.067	0.038	0.027	0.017
	J = 3, *T* = 10	0.098	0.075	0.048	0.038	0.026	0.017
	J = 5, *T* = 5	0.096	0.072	0.044	0.033	0.023	0.015
	J = 5, *T* = 10	0.071	0.054	0.033	0.033	0.023	0.014
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.141	0.123	0.093	0.088	0.067	0.047
	J = 1, *T* = 10	0.122	0.115	0.101	0.085	0.064	0.042
	J = 3, *T* = 5	0.123	0.100	0.069	0.043	0.029	0.019
	J = 3, *T* = 10	0.113	0.097	0.074	0.041	0.029	0.019
	J = 5, *T* = 5	0.100	0.078	0.049	0.035	0.025	0.015
	J = 5, *T* = 10	0.096	0.074	0.050	0.034	0.024	0.015
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.112	0.088	0.075	0.076	0.077	0.081
	J = 1, *T* = 10	0.102	0.087	0.076	0.048	0.058	0.065
	J = 3, *T* = 5	0.119	0.109	0.085	0.049	0.037	0.023
	J = 3, *T* = 10	0.110	0.105	0.087	0.049	0.034	0.022
	J = 5, *T* = 5	0.109	0.091	0.068	0.039	0.027	0.017
	J = 5, *T* = 10	0.108	0.096	0.086	0.037	0.027	0.017
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.121	0.101	0.073	0.076	0.058	0.046
	J = 1, *T* = 10	0.130	0.112	0.090	0.064	0.048	0.031
	J = 3, *T* = 5	0.102	0.080	0.065	0.042	0.033	0.020
	J = 3, *T* = 10	0.096	0.074	0.056	0.041	0.031	0.019
	J = 5, *T* = 5	0.085	0.069	0.050	0.037	0.028	0.018
	J = 5, *T* = 10	0.077	0.057	0.039	0.040	0.028	0.018
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.130	0.106	0.080	0.097	0.077	0.054
	J = 1, *T* = 10	0.128	0.117	0.111	0.093	0.067	0.045
	J = 3, *T* = 5	0.118	0.099	0.075	0.049	0.035	0.023
	J = 3, *T* = 10	0.134	0.111	0.087	0.048	0.034	0.022
	J = 5, *T* = 5	0.103	0.080	0.060	0.041	0.031	0.019
	J = 5, *T* = 10	0.107	0.080	0.053	0.042	0.029	0.019
TP = High, *IP* = High	*J* = 1, *T* = 5	0.113	0.088	0.078	0.113	0.111	0.112
	J = 1, *T* = 10	0.113	0.100	0.085	0.106	0.113	0.124
	J = 3, *T* = 5	0.130	0.119	0.111	0.056	0.042	0.028
	J = 3, *T* = 10	0.140	0.141	0.142	0.059	0.042	0.027
	J = 5, *T* = 5	0.126	0.108	0.087	0.045	0.034	0.021
	J = 5, *T* = 10	0.140	0.131	0.108	0.045	0.031	0.020

Note. Black cells are duplicate conditions from Phase 1.

Table B14.

Raw results for efficiency in the estimation of the transition probability parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.192	0.146	0.089	0.045	0.031	0.020
	J = 1, *T* = 10	0.113	0.069	0.040	0.024	0.017	0.011
	J = 3, *T* = 5	0.096	0.061	0.037	0.024	0.016	0.011
	J = 3, *T* = 10	0.043	0.029	0.019	0.017	0.011	0.007
	J = 5, *T* = 5	0.059	0.039	0.023	0.020	0.015	0.010
	J = 5, *T* = 10	0.030	0.020	0.013	0.016	0.011	0.007
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.181	0.136	0.087	0.060	0.046	0.029
	J = 1, *T* = 10	0.141	0.104	0.064	0.043	0.031	0.021
	J = 3, *T* = 5	0.113	0.075	0.047	0.033	0.024	0.015
	J = 3, *T* = 10	0.076	0.052	0.034	0.029	0.020	0.013
	J = 5, *T* = 5	0.074	0.055	0.034	0.028	0.021	0.013
	J = 5, *T* = 10	0.055	0.037	0.023	0.026	0.020	0.012
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.188	0.140	0.104	0.085	0.065	0.041
	J = 1, *T* = 10	0.164	0.131	0.095	0.080	0.060	0.037
	J = 3, *T* = 5	0.118	0.095	0.061	0.044	0.030	0.020
	J = 3, *T* = 10	0.117	0.085	0.055	0.044	0.030	0.019
	J = 5, *T* = 5	0.098	0.073	0.048	0.036	0.025	0.016
	J = 5, *T* = 10	0.091	0.066	0.042	0.035	0.025	0.015
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.185	0.152	0.100	0.048	0.033	0.022
	J = 1, *T* = 10	0.141	0.088	0.047	0.029	0.020	0.013
	J = 3, *T* = 5	0.115	0.062	0.039	0.028	0.020	0.012
	J = 3, *T* = 10	0.051	0.035	0.022	0.020	0.013	0.009
	J = 5, *T* = 5	0.060	0.043	0.026	0.024	0.018	0.011
	J = 5, *T* = 10	0.033	0.024	0.016	0.019	0.014	0.008
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.188	0.153	0.104	0.066	0.049	0.035
	J = 1, *T* = 10	0.160	0.119	0.079	0.052	0.037	0.023
	J = 3, *T* = 5	0.126	0.091	0.055	0.038	0.027	0.018
	J = 3, *T* = 10	0.100	0.063	0.039	0.033	0.023	0.014
	J = 5, *T* = 5	0.090	0.062	0.041	0.035	0.024	0.015
	J = 5, *T* = 10	0.069	0.046	0.029	0.030	0.021	0.014
TP = High, *IP* = High	*J* = 1, *T* = 5	0.201	0.163	0.121	0.107	0.084	0.052
	J = 1, *T* = 10	0.191	0.159	0.109	0.103	0.078	0.047
	J = 3, *T* = 5	0.151	0.116	0.079	0.054	0.038	0.023
	J = 3, *T* = 10	0.142	0.106	0.071	0.052	0.035	0.022
	J = 5, *T* = 5	0.121	0.089	0.058	0.040	0.029	0.018
	J = 5, *T* = 10	0.116	0.085	0.053	0.042	0.029	0.018

Note. Black cells are duplicate conditions from Phase 1.

Table B15.

Raw results for efficiency in the estimation of the probability of a correct response (non-master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.052	0.046	0.033	0.040	0.030	0.020
	J = 1, *T* = 10	0.039	0.033	0.024	0.029	0.021	0.014
	J = 3, *T* = 5	0.043	0.031	0.019	0.022	0.015	0.010
	J = 3, *T* = 10	0.029	0.022	0.014	0.018	0.013	0.008
	J = 5, *T* = 5	0.034	0.023	0.015	0.020	0.014	0.009
	J = 5, *T* = 10	0.026	0.019	0.012	0.018	0.012	0.008
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.050	0.042	0.031	0.051	0.038	0.027
	J = 1, *T* = 10	0.038	0.034	0.029	0.044	0.036	0.025
	J = 3, *T* = 5	0.046	0.034	0.022	0.028	0.020	0.012
	J = 3, *T* = 10	0.036	0.031	0.021	0.026	0.019	0.012
	J = 5, *T* = 5	0.038	0.029	0.018	0.025	0.017	0.011
	J = 5, *T* = 10	0.035	0.025	0.016	0.023	0.017	0.011
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.042	0.034	0.028	0.047	0.045	0.047
	J = 1, *T* = 10	0.033	0.029	0.024	0.036	0.035	0.035
	J = 3, *T* = 5	0.045	0.038	0.026	0.038	0.028	0.019
	J = 3, *T* = 10	0.043	0.035	0.025	0.035	0.027	0.018
	J = 5, *T* = 5	0.044	0.034	0.023	0.034	0.024	0.016
	J = 5, *T* = 10	0.043	0.033	0.025	0.033	0.023	0.015
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.063	0.055	0.041	0.046	0.034	0.026
	J = 1, *T* = 10	0.047	0.042	0.032	0.035	0.026	0.016
	J = 3, *T* = 5	0.053	0.037	0.024	0.026	0.018	0.012
	J = 3, *T* = 10	0.038	0.027	0.018	0.021	0.015	0.009
	J = 5, *T* = 5	0.040	0.029	0.018	0.024	0.017	0.011
	J = 5, *T* = 10	0.031	0.022	0.014	0.020	0.014	0.009
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.056	0.047	0.035	0.059	0.052	0.037
	J = 1, *T* = 10	0.041	0.040	0.037	0.054	0.043	0.030
	J = 3, *T* = 5	0.056	0.042	0.029	0.034	0.025	0.016
	J = 3, *T* = 10	0.050	0.037	0.028	0.031	0.023	0.015
	J = 5, *T* = 5	0.051	0.037	0.025	0.029	0.021	0.013
	J = 5, *T* = 10	0.041	0.032	0.021	0.027	0.020	0.012
TP = High, *IP* = High	*J* = 1, *T* = 5	0.046	0.039	0.035	0.058	0.065	0.066
	J = 1, *T* = 10	0.035	0.032	0.026	0.048	0.053	0.061
	J = 3, *T* = 5	0.060	0.050	0.038	0.046	0.035	0.023
	J = 3, *T* = 10	0.049	0.044	0.038	0.043	0.033	0.022
	J = 5, *T* = 5	0.056	0.046	0.033	0.042	0.029	0.019
	J = 5, *T* = 10	0.052	0.044	0.032	0.038	0.028	0.018

Note. Black cells are duplicate conditions from Phase 1.

Table B16.

Raw results for efficiency in the estimation of the probability of a correct response (master) parameter.

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	0.062	0.053	0.029	0.043	0.034	0.022
	J = 1, *T* = 10	0.030	0.021	0.014	0.017	0.013	0.008
	J = 3, *T* = 5	0.046	0.034	0.021	0.023	0.017	0.011
	J = 3, *T* = 10	0.019	0.014	0.009	0.013	0.009	0.006
	J = 5, *T* = 5	0.037	0.026	0.016	0.022	0.015	0.010
	J = 5, *T* = 10	0.017	0.012	0.008	0.012	0.009	0.006
TP = Med, *IP* = Low	*J* = 1, *T* = 5	0.044	0.034	0.019	0.031	0.024	0.015
	J = 1, *T* = 10	0.020	0.014	0.009	0.014	0.011	0.007
	J = 3, *T* = 5	0.031	0.021	0.014	0.021	0.014	0.009
	J = 3, *T* = 10	0.015	0.011	0.007	0.011	0.008	0.005
	J = 5, *T* = 5	0.027	0.018	0.012	0.018	0.013	0.008
	J = 5, *T* = 10	0.014	0.010	0.006	0.011	0.008	0.005
TP = High, *IP* = Low	*J* = 1, *T* = 5	0.027	0.018	0.011	0.020	0.015	0.009
	J = 1, *T* = 10	0.013	0.009	0.006	0.011	0.008	0.005
	J = 3, *T* = 5	0.020	0.014	0.009	0.017	0.011	0.007
	J = 3, *T* = 10	0.012	0.009	0.006	0.010	0.008	0.005
	J = 5, *T* = 5	0.019	0.014	0.009	0.016	0.011	0.007
	J = 5, *T* = 10	0.012	0.009	0.006	0.010	0.007	0.005
TP = Low, *IP* = High	*J* = 1, *T* = 5	0.053	0.046	0.028	0.034	0.026	0.018
	J = 1, *T* = 10	0.029	0.021	0.012	0.016	0.012	0.007
	J = 3, *T* = 5	0.039	0.027	0.018	0.021	0.015	0.009
	J = 3, *T* = 10	0.018	0.013	0.009	0.012	0.008	0.005
	J = 5, *T* = 5	0.029	0.022	0.014	0.018	0.013	0.008
	J = 5, *T* = 10	0.016	0.011	0.007	0.011	0.008	0.005
TP = Med, *IP* = High	*J* = 1, *T* = 5	0.043	0.032	0.018	0.027	0.021	0.014
	J = 1, *T* = 10	0.019	0.014	0.009	0.014	0.010	0.006
	J = 3, *T* = 5	0.031	0.021	0.013	0.019	0.013	0.008
	J = 3, *T* = 10	0.015	0.011	0.007	0.011	0.008	0.005
	J = 5, *T* = 5	0.027	0.018	0.012	0.018	0.013	0.008
	J = 5, *T* = 10	0.014	0.010	0.006	0.011	0.008	0.005
TP = High, *IP* = High	*J* = 1, *T* = 5	0.026	0.018	0.011	0.020	0.016	0.009
	J = 1, *T* = 10	0.014	0.010	0.006	0.011	0.009	0.005
	J = 3, *T* = 5	0.021	0.014	0.009	0.016	0.012	0.007
	J = 3, *T* = 10	0.012	0.009	0.006	0.011	0.007	0.005
	J = 5, *T* = 5	0.020	0.013	0.009	0.015	0.011	0.007
	J = 5, *T* = 10	0.012	0.009	0.006	0.010	0.008	0.005

Note. Black cells are duplicate conditions from Phase 1.

Table B17.

Raw results for classification accuracy (validation set).

		MQ = Low			*MQ* = Med		
		N = 200	*N* = 400	*N* = 1000	*N* = 200	*N* = 400	*N* = 1000
TP = Low, *IP* = Low	*J* = 1, *T* = 5	66.77%	66.67%	67.35%	80.94%	81.42%	81.84%
	J = 1, *T* = 10	88.06%	88.84%	89.21%	93.19%	93.20%	93.25%
	J = 3, *T* = 5	70.76%	72.23%	73.47%	91.19%	91.42%	91.46%
	J = 3, *T* = 10	89.91%	90.20%	90.33%	97.13%	97.11%	97.09%
	J = 5, *T* = 5	76.17%	77.26%	77.93%	95.08%	95.18%	95.12%
	J = 5, *T* = 10	91.59%	91.63%	91.84%	98.40%	98.37%	98.39%
TP = Med, *IP* = Low	*J* = 1, *T* = 5	87.58%	88.39%	89.48%	91.09%	91.22%	91.38%
	J = 1, *T* = 10	99.00%	99.18%	99.20%	99.22%	99.27%	99.26%
	J = 3, *T* = 5	88.86%	89.41%	89.73%	95.63%	95.64%	95.65%
	J = 3, *T* = 10	99.15%	99.19%	99.19%	99.64%	99.63%	99.65%
	J = 5, *T* = 5	89.92%	90.18%	90.31%	97.48%	97.49%	97.52%
	J = 5, *T* = 10	99.17%	99.20%	99.21%	99.80%	99.81%	99.80%
TP = High, *IP* = Low	*J* = 1, *T* = 5	99.63%	99.83%	99.87%	99.87%	99.86%	99.87%
	J = 1, *T* = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	J = 3, *T* = 5	99.87%	99.87%	99.88%	99.88%	99.90%	99.89%
	J = 3, *T* = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	J = 5, *T* = 5	99.87%	99.86%	99.87%	99.92%	99.92%	99.92%
	J = 5, *T* = 10	100.00%	100.00%	100.00%	99.99%	100.00%	100.00%
TP = Low, *IP* = High	*J* = 1, *T* = 5	73.75%	73.97%	75.06%	84.66%	85.14%	85.38%
	J = 1, *T* = 10	90.93%	91.30%	91.84%	94.59%	94.81%	94.91%
	J = 3, *T* = 5	76.57%	78.02%	78.84%	93.20%	93.34%	93.38%
	J = 3, *T* = 10	92.19%	92.48%	92.65%	97.82%	97.83%	97.83%
	J = 5, *T* = 5	80.92%	81.82%	82.35%	96.22%	96.34%	96.33%
	J = 5, *T* = 10	93.47%	93.68%	93.73%	98.83%	98.81%	98.79%
TP = Med, *IP* = High	*J* = 1, *T* = 5	89.70%	91.12%	92.07%	93.02%	93.19%	93.35%
	J = 1, *T* = 10	99.20%	99.37%	99.39%	99.43%	99.44%	99.45%
	J = 3, *T* = 5	91.18%	91.80%	92.18%	96.61%	96.68%	96.70%
	J = 3, *T* = 10	99.33%	99.37%	99.40%	99.73%	99.72%	99.73%
	J = 5, *T* = 5	91.96%	92.50%	92.67%	98.14%	98.12%	98.14%
	J = 5, *T* = 10	99.39%	99.41%	99.41%	99.85%	99.86%	99.86%
TP = High, *IP* = High	*J* = 1, *T* = 5	99.60%	99.86%	99.90%	99.90%	99.90%	99.91%
	J = 1, *T* = 10	99.95%	100.00%	100.00%	100.00%	100.00%	100.00%
	J = 3, *T* = 5	99.80%	99.91%	99.91%	99.90%	99.91%	99.91%
	J = 3, *T* = 10	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	J = 5, *T* = 5	99.89%	99.90%	99.91%	99.93%	99.93%	99.94%
	J = 5, *T* = 10	100.00%	100.00%	100.00%	99.99%	100.00%	100.00%

Note. Red: ≤ 70%, Orange: 70% - 79.99%, Yellow: 80% - 89.99%, Green: ≥ 90%. Black cells are duplicate conditions from Phase 1.

APPENDIX C

SAMPLE R CODE

```
#####Load required libraries
library(RNetica)
library(bnlearn)

#####Define simulation parameters
N = 200         #Sample size
C = 2           #Number of latent classes
J = 1           #Number of items per time slice
T = 5           #Number of time slices
R = 1000        #Number of Monte Carlo replications
DummyN = 1  #Number of dummy cases to include

#####Establish empty matrix for storing CPT estimates
CPTest <- matrix(NA, nrow=R, ncol=(2+(2*J)))
colnames(CPTest) <- c("P(Mt1)", "P(trans)", "P(X1=1|NM)", "P(X1=1|M)")

#####Establish empty array for storing category membership estimates from the
#validation set
CatEstTRN <- array(NA, c(N, T, R))
dimnames(CatEstTRN) <- list(NULL, c("ThetaT1", "ThetaT2", "ThetaT3", "ThetaT4",
"ThetaT5"), NULL)
CatEstVAL <- array(NA, c(N, T, R))
dimnames(CatEstVAL) <- list(NULL, c("ThetaT1", "ThetaT2", "ThetaT3", "ThetaT4",
"ThetaT5"), NULL)

#####Establish an empty array for the true values of the proficiency nodes from the
#validation set
CatTrueTRN <- array(NA, c(N, T, R))
dimnames(CatTrueTRN) <- list(NULL, c("ThetaT1", "ThetaT2", "ThetaT3", "ThetaT4",
"ThetaT5"), NULL)
CatTrueVAL <- array(NA, c(N, T, R))
dimnames(CatTrueVAL) <- list(NULL, c("ThetaT1", "ThetaT2", "ThetaT3", "ThetaT4",
"ThetaT5"), NULL)

#####Define network for BNlearn
newnet <- model2network("[ThetaT1][X1T1|ThetaT1][ThetaT2|ThetaT1]
        [X1T2|ThetaT2][ThetaT3|ThetaT2][X1T3|ThetaT3][ThetaT4|ThetaT3]
        [X1T4|ThetaT4][ThetaT5|ThetaT4][X1T5|ThetaT5]")

#####Specify CPTs for BNlearn
CPTinit <- matrix(c((1-init.m), init.m), ncol=C, dimnames=list(NULL, c("NM","M")))
CPTmeasJ1 <- matrix(c((1-p.x1.nm), p.x1.nm, (1-p.x1.m), p.x1.m), ncol=C, nrow=C,
dimnames=list(c("Incorrect", "Correct"), c("NM","M")))
CPTmeasJ2 <- CPTmeasJ3 <- CPTmeasJ4 <- CPTmeasJ5 <- CPTmeasJ1
CPTtrans <-matrix(c((1-p.trans), p.trans, 0, 1), ncol=C, nrow=C, dimnames=list(c("NM",
"M"), c("NM","M")))
```

```
#####Add CPTs to BNlearn model
newnet.comp <- custom.fit(newnet, dist=list(ThetaT1=CPTinit, ThetaT2=CPTtrans,
ThetaT3=CPTtrans, ThetaT4=CPTtrans, ThetaT5=CPTtrans,

#####Start Netica Session
startSession(DefaultNeticaSession)

#####Read in network from Netica file
net.newnet <- ReadNetworks(paths=paste(netPath, model.condition, ".dne", sep=""),
        session=DefaultNeticaSession)

#####Generate and save data from model using function from BNlearn
newnet.dat <- rbn(newnet.comp, n=N)

#####Reformat data and add dummy cases
newnet.dat.obs <- newnet.dat[,(T+1):(T*J+T)]
newnet.temp <- matrix("*", nrow=DummyN, ncol=ncol(newnet.dat.obs))
colnames(newnet.temp) <- colnames(newnet.dat.obs)
newnet.dat.obs <- rbind(newnet.dat.obs, newnet.temp)
Dummy <- matrix(c(rep("*", N), rep("M", DummyN), rep("*", N), rep("M", DummyN)),
nrow=(N+DummyN), ncol=2)
colnames(Dummy) <- c("Dummy1", "Dummy2")
newnet.dat.obs <- cbind(newnet.dat.obs, Dummy)
newnet.final <- newnet.dat.obs
newnet.final.full <- newnet.dat

#####Write a temporary case file (memory stream not working in Netica API v5.04)
newnet.file <- tempfile("newnet", fileext=".cas")
write.CaseFile(newnet.final, newnet.file, session=DefaultNeticaSession)

#####Set prior experience for nodes to help with label-switching
item.nodes <- c("X1T1", "X1T2", "X1T3", "X1T4", "X1T5")
for(elem in item.nodes){
        NodeProbs(nodes.newnet[[elem]]) <- matrix(c(.51, .49, .49, .51), nrow=C)
        NodeExperience(nodes.newnet[[elem]]) <- 1
}

#####Learn the CPT values using EM algorithm
LearnCPTs(newnet.file, nodes.newnet, method="EM", maxIters=10000)

#####Check if P(X=1|NM) > P(X=1|M) (evidence of label-switching)
if(NodeProbs(nodes.newnet$X1T1)[1,2] > NodeProbs(nodes.newnet$X1T1)[2,2]) next

#####Extract CPT estimates to R objects with corrections to alleviate labal-switching
#r.complete refers to the replication number
```

```
CPTest[r.complete,1] <- min(NodeProbs(nodes.newnet$ThetaT1))
CPTest[r.complete,2] <- max(c(NodeProbs(nodes.newnet$ThetaT2)[1,2],
NodeProbs(nodes.newnet$ThetaT2)[2,1]))
CPTest[r.complete,3] <- min(c(NodeProbs(nodes.newnet$X1T1)[1,2],
NodeProbs(nodes.newnet$X1T1)[2,2]))
CPTest[r.complete,4] <- max(c(NodeProbs(nodes.newnet$X1T1)[1,2],
NodeProbs(nodes.newnet$X1T1)[2,2]))

#####Conduct inference for each of the N individuals saving results to a master list
#r.complete refers to the replication number
newnet.dat.temp <- rbn(newnet.comp, n=N)
if(T==10){
        newnet.dat.temp <- newnet.dat.temp[,c(1,3:10,2,11:ncol(newnet.dat.temp))]
}
newnet.dat.obs.temp <- newnet.dat.temp[,(T+1):(T*J+T)]
newnet.dat.latent.temp <- as.numeric(unlist(newnet.dat.temp[,1:T]))
newnet.dat.latent <- as.numeric(unlist(newnet.final.full[,1:T]))
CatTrueVAL[,,r.complete] <- newnet.dat.latent.temp
CatTrueTRN[,,r.complete] <- newnet.dat.latent

for(n in 1:N){
        EnterFindings(net.newnet, newnet.dat.obs.temp[n,])
        for(elem in prof.nodes){
                CatEstVAL[n,elem,r.complete] <-
which.max(NodeBeliefs(nodes.newnet[[elem]]))
        }
        RetractNetFindings(net.newnet)
}

for(n in 1:N){
        EnterFindings(net.newnet, newnet.dat.obs[n,])
        for(elem in prof.nodes){
                CatEstTRN[n,elem,r.complete] <-
which.max(NodeBeliefs(nodes.newnet[[elem]]))
        }
        RetractNetFindings(net.newnet)
}

#####Stop Netica Session
DeleteNetwork(net.newnet)
stopSession(DefaultNeticaSession)
```