# Posterior Predictive Model Checking in Bayesian Networks

by

AAron Crawford

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

Approved April 2014 by the
Graduate Supervisory Committee:

Roy Levy, Chair
Samuel Green
Marilyn Thompson

ARIZONA STATE UNIVERSITY

May 2014

## ABSTRACT

This simulation study compared the utility of various discrepancy measures within a posterior predictive model checking (PPMC) framework for detecting different types of data-model misfit in multidimensional Bayesian network (BN) models. The investigated conditions were motivated by an applied research program utilizing an operational complex performance assessment within a digital-simulation educational context grounded in theories of cognition and learning. BN models were manipulated along two factors: latent variable dependency structure and number of latent classes. Distributions of posterior predicted *p*-values (PPP-values) served as the primary outcome measure and were summarized in graphical presentations, by median values across replications, and by proportions of replications in which the PPP-values were extreme. An effect size measure for PPMC was introduced as a supplemental numerical summary to the PPP-value. Consistent with previous PPMC research, all investigated fit functions tended to perform conservatively, but Standardized Generalized Dimensionality Discrepancy Measure (SGDDM), Yen's $Q_3$, and Hierarchy Consistency Index (HCI) only mildly so. Adequate power to detect at least some types of misfit was demonstrated by SGDDM, $Q_3$, HCI, Item Consistency Index (ICI), and to a lesser extent Deviance, while proportion correct (PC), a chi-square-type item-fit measure, Ranked Probability Score (RPS), and Good's Logarithmic Scale (GLS) were powerless across all investigated factors. Bivariate SGDDM and $Q_3$ were found to provide powerful and detailed feedback for all investigated types of misfit.

# DEDICATION

To Linda, for steadfast friendship in trying times, for carrying an excessive burden interminably, for sacrificing more of yourself and subsidizing more of this work than anyone knows.

ACKNOWLEDGMENTS

A heartfelt thanks to everyone who contributed to this work, particularly Dr. Samuel Green and Dr. Marilyn Thompson for constructive criticism of the study design and drafts of the paper, Dr. Dubravka Svetina for the suggestion of using multiple instances of R on a single machine as a time-saving computational strategy, Dr. Hollis Lai for sharing R code for ICI, Katie Kunze, Nedim Yel, and Derek Fay for useful feedback on aspects of presentation and communication in figures and tables, and for responsive suggestions during troubleshooting phases. A disproportionate debt of gratitude is owed to Dr. Roy Levy for his generous guidance on all aspects of the project throughout the entire process, and especially for mentoring with improbable patience and equanimity.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

## Introduction

Psychometric models automate the process of inferring students' cognitive development in a domain of interest. Throughout history, teachers have most commonly assumed the expert role of evaluating the performance of a student and making appropriate interpretations about that student's knowledge, skills, and abilities in relevant content areas. It is this work of evaluation by experts that is mimicked by psychometric models. In turn, the work of the models facilitates the development of theory by helping to coalesce, accumulate, and institutionalize expert knowledge. Expert knowledge is dissipated throughout a field of study, even across time. Psychometric models make it possible to bring together knowledge from various experts and incorporate it into computational components which can be recycled and reused in many different applications.

The end-products of psychometric models are inferences, but to have confidence in those inferences, one must trust that the model appropriately captures the relationships between data and theory. Model checking performs an essential role in the iterative process of validating psychometric models. Model checking serves to characterize the strengths and limitations of a model under various conditions. Model checking provides descriptive evidence about how the model functions with respect to different people in different situations, and with respect to the different levels of all the variables under investigation. How consistent is a model in its predictions? How accurate are the model's predictions compared to observed data? In the world of psychometric inference machines, model checking provides the quality control. Model criticism is a necessary check in the process of producing inferences which hope to carry the "valid" label.

## Purpose of the Study

The goal of this study was to improve the toolkit for assessing the data-model fit of Bayesian networks (BN). The primary focus within this overarching goal was comparing the utility of various discrepancy measures within a posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996) framework. Conditions in the present study were motivated by an applied research program on particular assessments utilized within a particular educational context, but were designed to be applicable to a much broader audience. As will be discussed in greater detail in the method section, some design decisions were intended to maintain close similarities to features of the motivating models, while other decisions were made to eliminate confounds, reduce noise variability, remove unnecessary complexity, or improve generalizability to other psychometric applications.

Model checking for BNs in psychometric applications is still in the early stages of development. There is a need in the literature for simulation studies to guide BN users with recommendations. The performance of fit indices borrowed from more established psychometric modeling paradigms such as item response theory (IRT; e.g. van der Linden & Hambleton, 1997), structural equation modeling (SEM; e.g. Kline, 2005), and latent class analysis (LCA; e.g. Collins & Lanza, 2010) is not well known in the context of BNs. Techniques that have demonstrated usefulness in these and other research contexts were gathered together for comparison within the current simulation study. Specifically, full information fit indices, limited information fit indices, item-fit indices, and person-fit indices were investigated.

On the upside, PPMC is an extremely flexible, and in many ways intuitive, model criticism framework which fits seamlessly with BNs under a common Bayesian modeling umbrella. The downside of PPMC includes its computational requirements and relative newness in the psychometric literature. Consider the limited model checking capacity of Netica (Norsys, 1995-2014), the most widely used commercial BN software package. Similar examples could be constructed using other popular BN software packages such as GeNie (Decision Systems Laboratory, 2012). Building alternative BN models in Netica is relatively time efficient. A base model can be modified quickly to reflect competing theoretical considerations, with each alternative model being estimated separately. Note that Netica does not implement a fully Bayesian approach when estimating BN parameters. Rather, marginal maximum likelihood (MML) is used to obtain an optimized solution which is interpreted from a frequentist perspective. A drawback of Netica is the relative lack of model checking resources to aid users in evaluating the empirical merits of a given model or set of models. The only indicator of global data-model fit provided by Netica is the loglikelihood value. Users can compare the global fit of competing models using this value in isolation, or the loglikelihood value can be used as an ingredient in the computation of a number of fit indices, such as AIC or BIC.

In addition, sensitivity analyses can be conducted. The analyst selects a single variable of interest and Netica will provide information regarding the influence of other variables in the system upon the chosen variable. The purpose is to quantify the sensitivity of the target variable to changes in the other variables. This feature may be sufficiently diagnostic for regression-like models where a single outcome is of primary

concern, but sensitivity analyses are not comprehensive enough to provide all the diagnostics typically desired in psychometric applications.

Another way to evaluate node characteristics in Netica is to individually compare the values within conditional probability tables (see West et al., 2012 for an example of this approach). In such an approach, differences in the conditional probabilities of successfully completing an observable are compared for members of different classes of examinees. Similar comparisons can be made using ratios between successive conditional probabilities. This technique allows researchers to quantify the discriminating power of observed variables, akin to an item discrimination parameter in IRT. In psychometric applications, observables with greater discrimination are generally considered to be of better quality. However, discrimination is not the same thing as fit; observables of varying discrimination can fit a model equally well (due to differing content coverage, for example), and items with the same discrimination can differ in how well they fit (due to construct relevance, for example). Nevertheless, due to the overlap that can exist between discrimination and fit, one way to identify some kinds of poorly fitting models may be to quantify their discriminating power using the above technique.

A more flexible and powerful way to critique BN models estimated in Netica is to conduct resampling analyses by simulating data from the solution network and comparing features of the observed data and simulated data (see section on parametric bootstrapping in the appendix), but these analyses must be conducted in a separate program (e.g. R, R Core Team, 2013; SAS, SAS Institute Inc., 2013) after exporting the simulated data from Netica.

## Literature Review

### Bayesian Networks

**Description of Bayesian Networks.** A BN (Pearl, 1988) is an inference machine for probabilistic reasoning, and its mathematical motor is Bayes' Theorem, also referred to historically as inverse probability (De Morgan, 1837; Fienberg, 2006):

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)} \quad (1),$$

where $P(A|B)$ is the probability distribution of one variable $(A)$ conditional on a second variable $(B)$. This distribution, known as the posterior distribution, is equal to the *un*conditional probability distribution of the first variable$(A)$, also known as the prior distribution, multiplied by the probability distribution of the second variable conditional on the first $P(B|A)$—this inversion of the posterior distribution is also known as the likelihood term—and divided by the *un*conditional probability distribution of the second variable $P(B)$.

Bayes' Theorem provides a method for computing unknown conditional probabilities, a task which the human mind struggles to execute, even under conditions of deliberate concentration by content experts (Kahneman, 2011). A BN extends the bivariate mathematical logic embodied in Bayes' Theorem to a multivariate system of probabilistic reasoning. A BN formalizes a body of evidence represented as distributions of variables and makes the proper (mathematical) inferences human judgment aspires to, based on the principles and assumptions of probability theory.

A BN is formally defined as the joint probability distribution of a system of interdependent variables; an acyclic directed graph (ADG or DAG) is a useful way to

visually represent the dependencies in a joint probability distribution, which are alternatively represented by equations. The DAG is composed of nodes and edges, which represent the dependencies among the variables of interest. There is a 1:1 correspondence between what is represented in the graph and dependence and conditional independence relationships in the joint distribution; one form is sufficient to generate the other. The DAG makes some features of the BN easier to comprehend, manipulate, and communicate. It is a convenient tool for working with models that can be unwieldy when represented only by equations.

Within a given system of variables, a DAG in which all pairs of variables are connected to each other is a saturated system, and the joint probability distribution is estimable using the general multiplication rule from probability theory. This saturated system can be constrained using expert knowledge about the interdependence among the modeled variables. Human knowledge is thus formalized into the structure of the BN model. BNs are a way to quantify the uncertainty that exists in the realm of human decision making. By expressing knowledge in probabilistic terms, BNs provide a numerically explicit way to test our understanding of the system of variables being investigated.

All variables in a Bayesian network, whether latent or observed, are treated as random variables that take on a discrete number of states. The joint probability distribution is the product of the probability distributions of the variables (nodes) in the network, conditional on the values of each node's parent variables. Parent variables are the immediate antecedents of the target variable in the dependency structure. A variable with no antecedents is modeled as exogenous (a.k.a. orphan). By comparison, other

variables might have a single parent, two parents, or many parents. Specification of variable parentage is how the structure of the joint distribution is established:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} P(X_1 = x_i | pa(X_i)) \quad (2),$$

where $X_i$ is a node in the network, $pa(X_i)$ are the parents of $X_i$, $P(X_i = x_i)$ is the probability distribution of $X_i$, and $P(X_i = x_i \mid pa(X_i))$ is the local probability distribution of variable $X_i$ conditional only on the values of that node's parents.

Bayesian networks have a number of attractive features. They are extremely flexible in the sense that very complicated dependencies can be represented relatively easily using graphical structure. In addition, nodes can vary in their properties (e.g. they do not need to have the same number of parents or states), so models can be customized to a particular situation as opposed to choosing an existing model "off the shelf" and applying it like a cookie cutter to the situation at hand.

After all of the conditional probabilities of a BN have been specified (either via expert knowledge or calibration with data), the model can be applied very quickly to the task for which the BN was designed: making inferences about specific situations based on a given state of knowledge, whether hypothetical or observed. A model in this fully specified, or calibrated, state is also called *ignorant* because it does not reflect specific findings for any particular case. Rather, the network contains marginalized knowledge, akin to what would be believed in aggregate across all cases in the population. If no response data is available for a particular case, predictions can be made using the ignorant (marginal) network. The network will make the same inferences for all cases with completely missing data. When any data for a particular case is available, the

network can be updated to reflect the current state of knowledge. The appropriate propagation of inputted information is applied recursively via Bayes' Theorem. Conditional probabilities are combined accordingly to yield the model-implied joint distribution given the current knowledge state, and the state-specific inferences are outputted.

To date, BNs have been used in a broad array of contexts, including academic, commercial, and governmental sectors, with notable examples from the fields of medicine, engineering, biology, environmental science, psychology, and education. In the following section, the application of BNs to the field of psychometrics is discussed in more detail.

**Bayesian networks in psychometrics.** The application of BNs to the field of educational assessment (Mislevy, 1995; Almond & Mislevy, 1999; Mislevy, Almond, Yan, & Steinberg, 2001; Sinharay, 2006b; Almond, DiBello, Moulder, & Zapata-Rivera, 2007) represents part of a broader, historical shift away from a trait paradigm toward a more cognitive paradigm. There are many different approaches to cognitively based psychometrics. A few examples of modeling paradigms include: Rule space method (Tatsuoka, 2009), attribute hierarchy method (AHM; Leighton, Gierl, & Hunka, 2004), and ordered multiple choice (OMC; Briggs, Alonzo, Schwab, & Wilson, 2006). Examples of design frameworks include evidence centered design (ECD; Mislevy, Steinberg, & Almond, 2003), and cognitive design system approach (Embretson, 1998). They share the common goal of seeking to provide a more detailed account of student learning and performance than has been obtained traditionally within the trait paradigm. The cognitive perspective emphasizes the constituent parts of a learning process which

might formerly have been summarized as a single entity. The cognitive perspective relies on accumulation of finer-grained evidence with which to make inferences about what students have learned. The big payoff is predictive power (greater specificity and accuracy) based on more extensive theoretical understandings of the latent construct(s). Other benefits include increased understanding of examinee behavior, more accurate inferences about students, improved opportunity for remediation, feedback for curriculum and instruction revisions, and improved understanding of the domain such as information about which skills are or are not necessary for successful performance. The term used hereafter to refer to this broad family of psychometric models is cognitive diagnostic models (CDMs).

CDMs are united essentially by their purpose or their applied uses. A methodological subset of this larger group has been referred to by many different names, including diagnostic classification models (DCMs; Rupp & Templin, 2008), the term used hereafter to refer to the subset of CDMs that use discrete latent variables to model cognition and task performance. In other words, DCMs are subsetted from CDMs based on the discrete (categorical) status of the latent variables. Even when it is true that a psychological construct is not distributed categorically in the population, it may still be useful to make evaluations categorically because the human mind is well suited to thinking categorically. Classification is a natural way to simplify complexity, and classification models often fit intuitively with the natural human proclivity to classify. All CDMs make subjective classification decisions in the sense that experts define key structural components of the models, whether setting cut scores or mapping out *Q*-matrices (Gorin, 2009).

A purely exploratory approach to model building in BNs can use intelligent algorithms (e.g. DEAL package in R, see Bottcher and Dethlefsen, 2012) to search for model configurations that optimize model fit for a given data set. By contrast, a more confirmatory approach relies on content expert judgment to construct a theoretically defensible model. For example in BNs, experts might be called upon to draw DAGs, specify parent-child relationships, impose meaningful constraints, specify conditional probabilities, specify the number of latent classes, etc.

Any model, regardless of its relative parsimony or complexity, will require agreement with data to stand the test of time. An advantage of BNs relative to some other types of models, however, is that content expertise (or even theoretical speculations) can take the place of data-estimated parameters in the *initial* stage of model building. In other words, the flexibility of BNs permits users to specify conditional probability distributions based on any source of a priori information. The quality of those specifications will necessarily affect the quality of the model in terms of fitting actual data, but data are not needed to begin the iterative cycle of model building, model criticism, and model refinement.

Another advantage of BNs relative to some other models is the convenient applicability after estimation. Programs like Netica and GeNie provide an intuitive interface that allows users to easily access the inferential power of the completed model. Calibrated models can easily be used by classroom teachers (Shute & Almond, 2008) or researchers to make respondent classifications as new cases become available. Given the assumption that new respondents belong to the population from which the model was constructed and calibrated, additional cases can be evaluated quickly and efficiently,

regardless of whether or not there is data for all observable variables. In fact, the BN can provide an a priori (marginal) classification, which is marginalized across all known model parameters, or any combination of partial observations. Any pattern of missing data is permitted because the uncertainties associated with inferences made by the BN are built into the model explicitly (see West et al, 2012 for examples of BN inferences using incomplete response patterns).

**Assessing Data-Model Fit**

One reason why psychometric models are imperfect is because they oversimplify complexity that exists in the data. Models seek to represent the most important relationships among the variables of interest; they seek to account for the most important sources of variability in the data. The usefulness of a model often rests on its ability to distill key features of the real world into a more manageable form. Fitting models to data often involves tradeoffs between parsimony and fidelity. The attractiveness of a model is inextricably linked to its applied purpose. One way to view validity is whether the model reproduces the inferences a prototypical domain expert would make using the same evidence.

A psychometric inference machine (e.g. BN) must be customized to each particular applied purpose. The goal of model checking is to troubleshoot the performance of the machine in the context of its current application, to find out which parts can be tweaked to improve parsimony or fidelity when representing the real world. Different types of modeling errors suggest different types of adjustments. The goal is to tinker with the functionality of the machine so that the next iteration of production contains fewer and/or less serious inferential errors.

According to Rupp, Templin, and Henson, authors of *Diagnostic Measurement* (2010), "many DCM applications are plagued by model-data misfit" (p. 166). This statement admits much about the state of cognitively based modeling in general. There are few prototypes in the literature that have developed sufficiently in the theoretical sense to withstand rigorous model criticism. The most frequently cited data source is Tatsuoka's mixed-number subtraction data (Tatsuoka, 1984). Ironically, it may be the improvement of model criticism tools that helps most to spur theoretical development because content experts often benefit from the feedback that model criticism brings. By providing a framework where specific features of theory and data coexist (e.g. in BNs dependency relationships and estimates of uncertainty must be made explicit), experts are pushed to formalize and explicate their theoretical understanding in new ways, and sometimes to consider new ideas or reconsider formerly discounted ideas. The feedback to content experts that comes from analyses of data is an exciting part of the iterative validation process. CDMs are currently being used to help build the cognitive theories that will be needed to justify their usefulness (compared to more conventional models) in applied settings. The process of building, troubleshooting, and validating models is necessarily iterative; it is a process of accumulated productivity (see Box, 1976). One of the greatest benefits of CDMs to the psychometric field is that they provide a way to test the theoretical knowledge provided by content experts. Models are built to help evaluate what students know. The models are themselves evaluated to see what the model-builders (domain experts) know. In this sense, model criticism serves to teach the experts about the weaknesses in their models. Model criticism is thus an integral part of theoretical

validation (see Gelman & Shalizi, 2013), and model criticism is essential for building evidentiary arguments about human learning.

**Posterior predictive model checking.** Posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Guttman, 1967; Levy, Mislevy, & Sinharay, 2009; Meng, 1994; Rubin, 1984; Sinharay & Johnson, 2003) has been used for evaluating the fit of many types of psychometric models, including BNs.

***Description of PPMC.*** PPMC circumvents the problem of calculating test statistic reference distributions by empirically building the reference distributions of interest using replicated data $\boldsymbol{y}^{\text{rep}}$generated from numerous draws of the model parameters $\Theta^1, \Theta^2, \ldots, \Theta^N$ from the posterior distribution

$$p(\boldsymbol{\Theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{\int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})d\boldsymbol{\Theta}} \quad (3).$$

A number *N* of replicated datasets are generated from the posterior distribution, with each generated dataset $\boldsymbol{y}^{\text{rep},n}$corresponding to a different draw of model parameters $\Theta^n$. Features of the replicated data are then compared to features of the observed (realized) data $\boldsymbol{y}$, using a range of techniques including graphical displays and summary statistics. Any desirable feature of the data can be compared in this way (see Gelman, Carlin, Stern, & Rubin, 2003 for examples and discussion of this topic).

When a statistic is calculated from model parameters for use as a comparison of replicated and realized data in PPMC, it is referred to as a *discrepancy measure* $D(\boldsymbol{y}, \boldsymbol{\Theta})$. Discrepancy measures should be selected according to the type of model and aspect of fit that are of interest. Particular discrepancy measures will be of use in some situations but not others. For example, Sinharay & Johnson (2003) found odds ratios to be a powerful

discrepancy measure for detecting inadequacy of a Rasch model for data from 2PL/3PL models, a 3PL model for 2-dimensional data, a 3PL model for data from a testlet model, and a 3PL model for speededness data, but ineffective for detecting inadequacy of a 2PL model for data from a 3PL model.

For each of the (*N*) draws from the posterior distribution, two values of a discrepancy measure are calculated: one using the observed data and one using the replicated data. The discrepancy measures that result from using the replicated data $D(\boldsymbol{y}^{\boldsymbol{rep}}, \boldsymbol{\Theta})$are compared to the values of the discrepancy measures using the observed data $D(\boldsymbol{y}, \boldsymbol{\Theta})$. In this way, the replicated data serves as an empirical reference distribution for evaluating the observed values of the model-fit statistics. PPMC does not require re-estimation of the model, but does require generating replicated data sets and computing discrepancy measures from the generated data.

One way to summarize discrepancy measures is with empirical *p*-values (also called *posterior predictive p-values*, or PPP-values). In a simulation environment such as Markov chain Monte Carlo (MCMC; e.g. Gelman et al., 2003), PPP-values are the proportion of draws in which the replicated values are greater than the values using observed data. The expectation is that PPP-values will be at or near .5 when the model fits the data. More extreme values in either direction are indicative of data-model misfit, because they suggest that the model is systematically under- or over-producing the discrepancies. Direction of misfit is not necessarily important out of context, but patterns of directionality may be informative within the context of a particular discrepancy measure and/or model of interest. Importantly, the PPP-values are not statistical tests, so they should not be interpreted in the same way as traditional frequentist *p*-values. PPP-

values are simply one way of summarizing the relative values of the discrepancy measures, and should be used as part of a larger evidence argument when assessing data-model fit using PPMC (Levy, Mislevy, & Sinharay, 2009; Sinharay 2006b).

A graphical way to compare discrepancy measures is with a scatterplot of predicted discrepancy values (based on replicated data) vs. the realized discrepancy values (based on observed data). Figure 1 shows an example of such a plot taken from Sinharay (2006b).

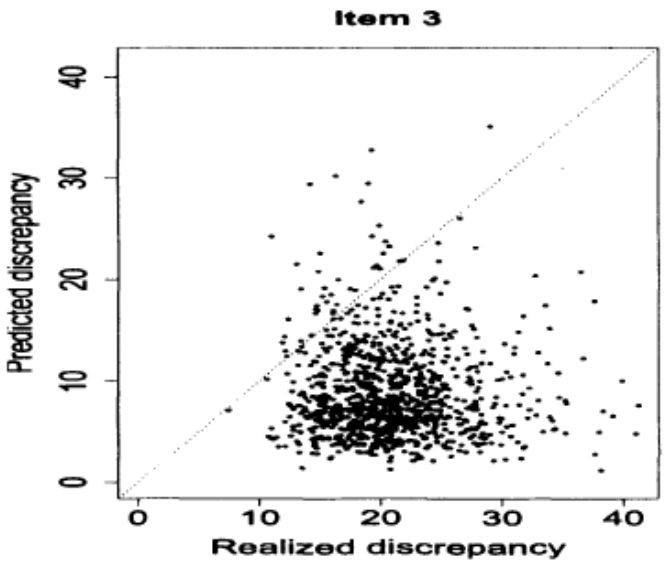



*Figure 1*. Scatterplot of predicted versus realized discrepancies. The associated PPP-value was .04, which was suggestive of misfit. Taken from Sinharay (2006b).

Each point in this plot represents a draw of model parameters from the posterior distribution. A 45°-reference line shows where the points would fall if the values of the discrepancy measure from the replicated and original data sets were equal to each other. Points that fall away from the reference line indicate draws where there is a difference between replicated and original data on the statistic of interest. The reference line therefore serves to separate the points into two categories: one where the replicated

values are larger than the realized and one where the realized values are larger than the replicated. The graph serves as a holistic device for detecting systematic differences between the realized and replicated discrepancies. To the extent that points fall evenly on both sides of the line, the model is said to adequately fit the data with respect to the discrepancy of interest. Conversely, to the extent that most points fall on one side of the line, evidence of model misfit is indicated. Note that distance from the line is not necessarily of principal interest in a graph like Figure 1; however, distance from the line does represent the magnitude of the difference between observed and predicted data for any given draw. An open area of research is how to characterize and summarize systematic differences in the observed patterns of these types of PPMC results.

A potential disadvantage associated with PPMC is that it may require the researcher to use multiple software packages. This is not a disadvantage of PPMC per se, in that the process of implementing the technique will likely improve with software developments, but the current software options do impact practical considerations. Mplus 6.0 is capable of doing PPMC, but is not suited for BNs (Muthén & Muthén, 2010). WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007) is more flexible than Mplus in terms of the types of models it can estimate, but it tends to be relatively slow, and output from WinBUGS most often needs to be passed to another package, such as R (R Development Core Team, 2013), to compute the discrepancy measures of interest, which increases computational time as well.

In summary, PPMC is often more computationally intensive than alternative frameworks. In addition, current software options may necessitate an investment of time to customize programming code. PPMC is also remarkably flexible, and is potentially

more informative in the sense that a greater variety of statistics can be used as model-checking tools because the reference distributions are generated empirically.

***Example using $Q_3$.*** To aid in describing the PPMC framework, I will draw upon an example using Yen's $Q_3$ (Yen, 1993), described next. $Q_3$ has been used in a variety of modeling contexts as a check of the local independence (LI) assumption. The LI assumption asserts that responses are conditionally independent, meaning that after accounting for the parameters within the model, responses are independent of each other (Levy & Svetina, 2011). Yen (1993) provided a description of the following sources of LI violations: external assistance or interference, speededness, fatigue, practice, item or response format, passage dependence, item chaining, explanation of previous answer, scoring rubrics or raters, and content knowledge. All LI violations can be framed in terms of multidimensionality (Ip, 2001), but a violation of LI is not necessarily evidence of dimensionality misspecification. Under-specifying dimensionality will result in local dependence, but if dimensionality is over-specified local independence will hold. Yen explained that constant effects do not produce dependence. To produce dependence, these sources must have differential effects on items or respondents. $Q_3$ is a statistic for evaluating the degree of dependency between pairs of observed variables, conditional on an assumed model. $Q_3$ is defined as the correlation between a pair of residuals from observables $j$ and $j'$ :

$$Q_{3_{ij}} = r_{e_{ij}\, e_{ij'}} \qquad (4),$$

where $e_{ij}$ is the difference between the observed response $X_{ij}$ and expected (model-implied) response $E(X_{ij})$ for person $i$ on observable $j$. Values of $Q_3$ indicate the extent to which there are dependencies between pairs of observed variables that are not accounted

for by the model. As values approach 1 (or -1), the function indicates that the association between a particular pair of variables in the data is still strongly positive (negative) even after accounting for the influence of the modeled relationships. Positive associations indicate that as the value of one variable changes, the second variable changes in the same direction. Negative associations indicate that as the value of one variable changes, the second variable changes in the opposite direction. The local independence assumption is undermined when $Q_3$ values are sufficiently large (e.g. > .2, see Yen, 1993; Chen & Thissen, 1997) because the pairs of variables in question exhibit positive (negative) dependence above and beyond what is accounted for by the model. Stated differently, large $Q_3$ values indicate that there are positive (negative) residual dependencies between data values that cannot be explained by the model structure alone. Conversely, as values approach 0, the $Q_3$ function indicates that the association between a particular pair of variables in the data is weak after accounting for the influence of the modeled relationships. In other words, the local independence assumption appears warranted because the pairs of variables in question exhibit limited residual dependence above and beyond what is accounted for by the model.

In the context of PPMC, observed $Q_3$ values are measured against empirically generated reference values. For each draw of model parameters from the posterior distribution, a $Q_3$ value for each pair of observed variables is calculated using the realized data and another $Q_3$ value is calculated using predicted data. If 500 draws are taken from the posterior distribution, then there are 500 sets of $Q_3$ values, and each set would contain two $Q_3$ values for each pair of variables: one $Q_3$ value using realized data and one $Q_3$ value using predicted data (these are the two values that would constitute the

coordinates for each point in a scatterplot such as Figure 1). The proportion of draws in which the posterior predicted value exceeded the observed value provides the researcher with a summary of the pattern of values that can be used to judge the degree of misfit.

***Applications of PPMC.*** Sandip Sinharay and his coauthors have demonstrated several techniques for assessing model fit within a PPMC framework. This line of research has included unidimensional IRT models (Sinharay, 2003; Sinharay & Johnson, 2003; Sinharay 2005; Sinharay, 2006a; Sinharay, Johnson, & Stern, 2006) as well as BNs (Sinharay, 2004; Sinharay, 2006b; Sinharay & Almond, 2007). A variety of statistics and graphical displays have been proposed by these authors for use with PPMC, including: direct data displays for overall fit (first demonstrated in Gelman, Carlin, Stern, and Rubin, 2003), $X^2$ and $G^2$-type measures (based on equivalence class membership and then on raw score) to assess item fit, point-biserial correlations and odds ratios as measures of inter-item associations; a variant of the Mantel-Haenszel statistic (Holland, 1985) for assessing differential item functioning (DIF), and checks of parameter identifiability.

**Fit functions.** In this section, specific fit functions are presented in greater detail to inform the method section which follows. Many of the fit functions could theoretically be implemented using any of the four model criticism frameworks discussed in the appendix, but in practice some fit functions do not lend themselves conveniently to all of the frameworks. For example, within a hypothesis testing (HT) framework the analytical derivation of reference distributions is often prohibitive, so researchers often avoid using a fit function for which the reference distribution has not been established.

***Global fit.*** Global fit refers to the fit of the model as a whole, summarized as a single number. Several techniques have been developed to compare fit across different models based on the estimated value of the likelihood function. This maximum likelihood value is often reported as a deviance (*d*):

$$d = 2\ln\left(L - L_{sat}\right) \quad (5),$$

where $L$ is the value of the likelihood function of the model under investigation and $L_{sat}$ is the likelihood estimator of the saturated model. Deviance was first proposed as a model-checking tool by Nelder and Wedderburn (1972).

Among the most common global fit statistics are the Pearson $X^2$ test and the $G^2$ test of overall model fit:

$$X^2 = \sum_{r=1}^{2^j} \frac{O_r - E_r}{E_r} \quad (6),$$

and

$$G^2 = 2\sum_{r=1}^{2^j} O^r ln \frac{O_r}{E_r} \quad (7),$$

where *r* is a particular response pattern, and *J* is the number of items. When implemented within an HT framework, the final summation for both equations is evaluated as a $\chi^2$ statistic with $df = (2^J$ – number of estimated parameters – 1).

These statistics will only follow a $\chi^2$distribution when all of the response patterns are adequately represented in the sample; in other words, when the contingency table is sufficiently populated. Because the number of response patterns is $2^J$, these hypothesis tests become problematic for long tests and/or small samples. Stated differently, full information fit indices (e.g. global $X^2$ and $G^2$statistics) are generally not usable within an

HT framework to critique BNs because of the sparsely-filled contingency tables that result from prohibitively large numbers of response patterns in most psychometric applications. The sparsely populated response patterns make these tests impractical for most BNs. The problematic behavior of the sampling distributions in such situations can be circumvented using a framework where the reference distribution is generated empirically, i.e. parametric bootstrapping (PB) or PPMC. In addition, these indices have variants at the item level (see below). Separate from the issue of an appropriate reference distribution for the statistic is the issue of whether these fit statistics would provide useful feedback for different types of data-model misfit. For example, Levy, Mislevy, and Sinharay (2009) and Levy (2011) found these statistics to be useless for the detection of multidimensionality in IRT models. In a rare example of an applied study that used multiple model-checking frameworks simultaneously, Loken (2004) demonstrated that hypothesis tests with inexact reference distributions can still provide heuristic value in applied settings because they do give researchers a sense of the magnitude of misfit even when the *p*-values cannot be trusted at face value.

When models are nested, meaning that one model is a constrained version of the other, a likelihood ratio (LR) test can be performed to compare model fit. The LR statistic is the difference between the deviances of the two models:

$$LR = d' - d \quad (8),$$

where $d'$ is the deviance of the more restrictive model. The more restrictive model will never fit better than the less restrictive model, so the result of Equation 8 will never be a negative number. Within an HT framework, the resulting difference is then evaluated as a $\chi^2$statistic with degrees of freedom equal to the difference in the number of parameters

between the less restrictive and more restrictive models. However, the LR statistic does not follow a $\chi^2$distribution when model parameters take on boundary values. The design of the current study did not emphasize model comparison of nested models, so the LR test was not discussed further.

One common classification system utilized in means and covariance structure modeling puts global fit indices into three broad groups: absolute, parsimonious, and incremental. Absolute indices compare the observed variance-covariance matrix to the model-implied variance-covariance matrix. Examples of absolute indices include the Model *T* statistic, which is the foundation for some of the other fit indices (Yuan, 2005), Standardized Root Mean Squared Residual (SRMR), and Goodness-of-fit Index (GFI).

Parsimonious indices (also called penalized indices) make adjustments based on the number of parameters in the model. Model complexity is considered in conjunction with the deviance statistic. Each variant modifies the deviance statistic in a different way, but they all offer a way to evaluate whether models of increasing complexity (i.e. more parameters) are worth it. Examples of parsimonious indices include Adjusted Goodness-of-Fit Index (AGFI), Root Mean Squared Error of Approximation (RMSEA), AIC, and BIC.

Incremental indices compare the model of interest to a baseline model where all the model parameters are independent of each other. The baseline model is a worst-case scenario which provides no explanatory power whatsoever, which is akin to having no model at all (i.e. associations are products of chance alone). Examples of incremental indices include Comparative Fit Index (CFI), Normed Fit Index (NFI), and Nonnormed Fit Index (NNFI). Residuals between the observed and model-implied variance-

covariance matrices can be inspected manually or graphically to investigate localized data-model fit. In addition, these residuals are incorporated into many fit indices.

Two of the most commonly used relative indices are Akaike's information criterion (AIC; Akaike, 1974), and Bayesian information criterion (BIC; Schwarz, 1978). When used in a framework with no reference distribution (NRD), the model with the lowest index value is taken as the best-fitting model. Relative fit indices should not be used as the sole justification for a model because a model that fits better than its competitors may still fit poorly by absolute criteria. These indices are useful for ranking a set of models that all fit adequately in an absolute sense (Rupp, Templin, & Henson, 2010, p. 279). AIC is given by

$$AIC = -2\ln(L) + 2p \quad (9),$$

where $p$ is the number of estimated parameters. BIC is given by

$$BIC = -2\ln(L) + \ln(n)\,p \quad (10),$$

where $n$ is the sample size.

Tests of incremental fit, including the LR test and information criteria like the AIC and BIC, may be used with BNs, but not much research has been done to guide interpretations of these statistics in this context (Rupp, Templin, & Henson, 2010). Within a PPMC framework, the deviance term (-2ln$L$) varies across replications within a given model, and therefore can be utilized as a discrepancy measure (Gelman et al., 2003; see Steedle, 2008, for an application). By contrast, when computing AIC and BIC within a PPMC framework, $n$ in Equation 9 and $p$ in Equations 9 and 10 are constants across replications for any given model, so AIC and BIC do not provide additional utility above and beyond the deviance term. For this reason, deviance was used in the present study

while AIC and BIC were not. GDDM (Levy & Svetina, 2011) evaluates whether the dimensionality for a given set of items is adequately represented by the specified model:

$$GDDM = \frac{\sum_{j>j'} \left| \frac{\sum_{i=1}^{N} \left( X_{ij} - E\left(\mathrm{X}_{ij} \middle| \boldsymbol{\theta}_i, \boldsymbol{\omega}_j \right) \right) \left( X_{ij'} - E\left(\mathrm{X}_{ij'} \middle| \boldsymbol{\theta}_i, \boldsymbol{\omega}_{j'} \right) \right)}{N} \right|}{J(J-1)} \quad (11),$$

where $X_{ij}$ is the scored value (1 or 0) from examinee $i$ on observable $j$, $\boldsymbol{\theta}_i$ are the student model variables for examinee $i$, $\boldsymbol{\omega}_j$ are the conditional probabilities that govern the distribution of observable $j$, $N$ and $J$ are the number of examinees and observables, respectively, and $E(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j)$ is the model-implied expected value from examinee $i$ on observable $j$, which in the context of dichotomous observables is the model-implied probability that the examinee correctly completes the aspect of the task captured by that observable. Note that a set of observables comparable by the GDDM can consist of the full set of observables, in which case it functions as an assessment of global fit, or a subset of observables (as few as two), in which case it functions as a local fit tool. A standardized version of the GDDM (SGDDM; Levy, Xu, Yel, & Svetina, 2012) has been developed to overcome limitations associated with properties of the covariance metric:

$$SGDDM = \frac{\sum_{j>j'} \left| \frac{\sum_{i=1}^{N} (X_{ij} - E(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j))(X_{ij'} - E(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j))}{N} \right|}{\sqrt{\frac{\sum_{i=1}^{N} (X_{ij} - E(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j))^2}{N}} \sqrt{\frac{\sum_{i=1}^{N} (X_{ij'} - E(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j))^2}{N}}} \Bigg| \Bigg| \quad (12).$$

***Local fit.*** Limited information fit statistics, including univariate and bivariate statistics, have been used in BNs (and other CDMs) to help investigate local dependence

(Levy et al., 2009; Levy, 2011), item fit, and via summation, to help address the need for indices of global fit (Rupp, Templin, & Henson, 2010). A drawback of these statistics within an HT framework is that their reference distributions remain unknown. These statistics retain some heuristic utility, even when the reference distributions are only approximations, but the *p*-values cannot be taken at face value. Future research is needed to clarify the advantages and disadvantages of using statistics heuristically versus committing to a framework which estimates the reference distributions empirically. The framework emphasized in the present study was PPMC, but conceptually related techniques within a frequentist framework (i.e. PB) might yield similar findings.

Chen and Thissen (1997) used a simulation study to compare the effectiveness of four statistics (Yen's $Q_3$, Pearson's $\chi^2$, The Likelihood Ratio $G^2$, and The Standardized φ Coefficient Difference) for detecting local dependence among item pairs in IRT models. $Q_3$, $\chi^2$, and $G^2$ were each found to be preferable to the other three indices under some conditions.

Sinharay and Almond (2007) used a $\chi^2$-type item-fit statistic to help detect misfitting items in a BN with two latent classes fit to Tatsuoka's (1984) mixed-number subtraction data:

$$\chi_j^2 = \sum_k \frac{N_k(O_{kj} - E_{kj})^2}{E_{kj}(N_k - E_{kj})} \quad (13),$$

where $N_k$ is the number of examinees with skill pattern *k*, $O_{kj}$ is the number of examinees with skill pattern *k* that responded correctly to item *j*, and $E_{kj}$ is the product of the expected proportion of correct responses for pattern *k* multiplied by $N_k$ . Note that because equivalence-class membership is not actually observed, $O_{kj}$ is substituted by

$\hat{p}_{kj}N_k$, where $\hat{p}_{kj}$is the median proportion of class membership from the posterior distribution.

The Item Consistency Index (ICI; Lai, Gierl, & Cui, 2012) is an item-fit index for use in CDMs. It was developed from the person-fit analogue Hierarchy Consistency Index (HCI; Cui & Leighton, 2009). The ICI is given by

$$ICI_j = 1 - \frac{2\sum_i \left[\sum_{g\epsilon S_j} X_{i_j}(1 - X_{i_g}) + \sum_{h\epsilon S_j^*} X_{i_h}(1 - X_{i_j})\right]}{N_{c_j}} \quad (14),$$

where $X_{i_j}$ is student *i*'s score for item *j*, $S_j$ is an index set that includes items requiring the subset of attributes measured by item *j*, $X_{i_g}$ is student *i*'s score to item $g$ where item $g$ belongs to $S_j$ , $S_j^*$is an index set that includes items requiring all, but not limited to, the attributes measured by item *j*, $X_{i_h}$ is student *i*'s score to item *h* where item *h* belongs to $S_j^*$, and $N_{c_j}$ is the total number of comparisons for item *j* across all students. The kernel of the ICI counts the number of mismatches between the observed and expected responses to items as dictated by the hypothesized model. This count is then divided by the number of possible comparisons being made, yielding a proportion of mismatched comparisons. The numeric constant "2" in the numerator serves to change the index from a proportion metric ranging from 0 to 1, to a metric ranging from -1 to 1. The resulting quantity is then subtracted from 1 to translate the index into matched comparisons, as opposed to mismatched comparisons.

The next three indices (WPI, RPS, and GLS) belong to a large family of statistical functions known traditionally as scoring rules. These scoring rules were developed historically outside of psychometrics, and have not appeared much in the psychometric

literature. However, given their general structure and broad applicability in other statistical applications, they were considered herein as potentially valuable fit functions. Weaver's Surprise Index (WSI; Weaver, 1948) makes a distinction between "rare" events and "surprising" events, the latter being distinct from the former by virtue of being unusual in relation to alternative outcomes, as opposed to simply being unusual in an absolute sense. Weaver reminds us that in a scenario where all possible outcomes are equally rare, a rare outcome would be inevitable and should therefore be construed as unsurprising. Researchers are cautioned against mistaking rare events for surprising events. The WSI provides a formal computation of surprise, thereby relieving the researcher of embarrassing emotional attributions to rare events. The WSI ranges from unity to infinity, with values indicating surprise as they grow increasingly large. In addition to showing how to compute his surprise index, Weaver also demonstrated how he interpreted its outputted values: "A Surprise Index of 3 or 5 is surely not large; one of 10 begins to be surprising; one of 1,000 is definitely surprising; one of 1,000,000 or larger is very surprising indeed; one of 1,000,000,000,000 would presumably qualify as a miracle" (Weaver, 1948, p. 392).

$$WSI_i = \frac{E(p)}{p_i} = \frac{{p_1}^2 + {p_2}^2 + p)\, {p_n}^2}{p_i} \tag{15}$$

The Ranked Probability Score (RPS; Epstein, 1969) was developed by Edward Epstein in the context of weather forecasting, where categories of potential temperatures were assigned probabilities and forecasts were assessed based on whether observed temperatures fell within specified temperature ranges (categories). Epstein noted that pre-existing indices did not take into account how much "distance" existed between the

observed category and the predicted category, a feature which he incorporated into the RPS. RPS scores range from zero to one, with a score of zero indicating the worst possible prediction (i.e. an outcome which is the polar opposite of the expectation), and a score of one indicating perfect prediction.

$$RPS_j = \frac{3}{2} - \frac{1}{2(K-1)} \sum_{i=1}^{K-1} \left[ \left( \sum_{n=1}^{i} p_n \right)^2 + \left( \sum_{n=i+1}^{K} p_n \right)^2 \right] - \frac{1}{K-1} \quad (16)$$

Williamson et al. (2000) found Good's Logarithmic Score (GLS; Good, 1952) to be unique in its ability to detect errors of node state inclusion or exclusion, although it did not perform as well as the WSI or RPS in terms of detecting other types of errors, namely node inclusion or exclusion, edge inclusion or exclusion, and prior probabilities. The GLS was developed as a tool for quantifying the merit of probabilistic judgments by experts. As a side note, Good also provided a formula by which experts' payments would vary as a function of accuracy. Good described it as "a method of introducing piece-work into the Meteorological Office. The weather forecaster would lose money whenever he made an incorrect forecast." (Good, 1952, p. 112). The GLS is given by

$$GLS = \log (bp_i) \quad (17)$$

when the predicted event occurs, and

$$GLS = \log\ b(1 - p_i) \quad (18)$$

when the predicted event does not occur. The prior probability of event *i* is $p_i$, and *b* is a penalty term which was designed to keep the expert from guessing the marginal expectation instead of considering as much evidence as possible in a particular case. The penalty term is given by

$$b = -\sum_{j=1}^{r} x_j log x_j \quad (19),$$

where $r$ is the number of possible outcomes and $x_j$ is the marginal probability associated with category $j$. The GLS ranges from zero to infinity, with values of zero representing perfect prediction and increasingly large values representing inaccuracy.

Note that while RPS, WSI, and GLS can be implemented as item-fit functions, they exemplify the principle that fit functions can often be used in a variety of ways. Williamson et al (2000) describe how these three functions can be aggregated to assess item fit, person fit, or global fit, depending on the needs of the researcher. This principle of variable use applies to many of the fit functions used in model criticism research.

***Person fit.*** Person fit is a way to evaluate whether a particular model applies adequately to certain response patterns, and by extension, to the respondents represented by those response patterns. Person-fit statistics do not directly test the cause of an aberrant response pattern. Causal interpretations must be investigated and validated separately. In many person-fit applications, it is the misfitting individuals who are of interest. For example, these fit statistics have been used to identify cheating, test anxiety, faking (of personality or clinical diagnoses), or lack of motivation (Karabatsos, 2003; Meijer and Sijtsma, 2001). In other applications, it is the improvement of data/model fit that is of interest. Misfitting people degrade the quality of the estimated model parameters. Removing misfitting people from the sample effectively redefines the population to which the remaining sample will generalize. The loss of generalizability in this process is compensated by improved prediction or understanding of the remaining persons who do adequately fit the employed model.

Person-fit statistics measure the extent to which observed response patterns are deviant from typical response patterns that are expected under the utilized model. The statistics are often model-specific; in such cases, a researcher must painstakingly choose from the large family of person-fit statistics that has been developed. Armstrong and Shi (2009) introduced a model-free approach to person-fit for linear tests, based on likelihood ratios. Emons, Sijtsma, and Meijer (2005) proposed a three-step methodology to initiate an investigation of person fit. The first step is to use a global person-fit statistic to identify questionable response patterns. The second step involves graphical construction of a person response function (PRF), which required nonparametric kernel smoothing of the observed response pattern. The third step involves using a local person-fit statistic to test specific items which appeared to cause irregularities in the PRF. Glas and Meijer (2003) used PPMC in a simulation study to compare the detection rates and false alarm rates of 8 person-fit indices to detect aberrant response patterns in a 3-parameter normal ogive (3PNO) model. "Bayesian *p*-values" were reported as an outcome.

Meijer and Sijtsma (2001) reviewed 40 person-fit indices. The authors grouped the indices into two main categories: group-dependent (a.k.a. non-parametric, model-free) and IRT-based (a.k.a. model-dependent). Karabatsos (2003) compared 36 person-fit indices using simulation study implementing a Rasch model. He characterized $H_i^T$, which is a correlation between one observed response pattern and the remaining observed response patterns, as the top-performing index, although there were a few others that performed nearly as well.

According to Cui and Leighton (2009), the group-dependent category of person-fit statistics assumes unidimensionality. Due to the multi-dimensional nature of CDMs,

existing person-fit statistics are not appropriate for CDMs. The authors consequently introduced the Hierarchy Consistency Index (HCI) for evaluating person-fit in cognitive diagnostic models. HCI values range from -1 to 1, with lower scores indicating response patterns that are expected with lower frequency given the cognitive model. The HCI for student *i* is given by

$$HCI_i = 1 - \frac{2\sum_{j \in S_{correct_i}} \sum_{g \in S_j} X_{i_j}\left(1 - X_{i_g}\right)}{N_{c_i}} \quad (20),$$

where $S_{correct_i}$ is an index set that includes items correctly answered by student *i*, $X_{i_j}$ is student *i*'s score for item *j*, where item *j* belongs to $S_{correct_i}$ , $S_j$ is and index set that includes items requiring the subset of attributes measured by item *j*, $X_{i_g}$ is student *i*'s score to item *g* where item *g* belongs to $S_j$ , and $N_{c_i}$ is the total number of comparisons for all the items that are correctly answered by student *i*. The kernel of the HCI counts the number of mismatches between the observed response vector and the expected response vector as dictated by the Q matrix. This count is then divided by the number of possible comparisons being made, yielding a proportion of mismatched comparisons. The numeric constant "2" in the numerator serves to change the index from a proportion metric ranging from 0 to 1, to a metric ranging from -1 to 1, which its creators preferred on the basis of interpretability. The resulting quantity is then subtracted from 1 to transfer the focus of the index from mismatches to matches.

***Research on discrepancy measures.*** Many different discrepancy measures have been employed in the literature using applied data, but relatively few simulation studies exist where discrepancy measures have been systematically compared and evaluated. Among studies of the latter variety, attention has been divided across different modeling

paradigms. Levy, Mislevy, and Sinharay (2009) and a closely related follow-up study (Levy, 2011) operated within an IRT/PPMC framework. Williamson (2000) and Williamson, Mislevy, & Almond (2000) used a BN/PB framework. Levy (2006) contained two simulation studies within a PPMC framework, one using IRT models and the other using BNs. Both types of sources above—simulation studies using PPMC with alternative psychometric models, and applications of PPMC using BNs—were used to inform the choices of discrepancy measures for the present study.

In a simulation study, Levy (2006) compared the performance of eight discrepancy measures for criticizing the fit (bivariate associations) in BN models which ignored inhibitory relationships in the generated data. Model-based covariance (MBC; Reckase, 1997) and $Q_3$ (Yen, 1993) were found to perform the best. Four discrepancy measures that performed similarly to each other were categorized together as the next best: covariance, residual item covariance (McDonald & Mok, 1995), log odds ratio (Agresti, 2002), and standardized log odds ratio residual (Chen & Thissen, 1997). Finally, $X^2$ and $G^2$ (Chen & Thissen, 1997) were found to be less useful than the other discrepancy measures because they did not indicate the directionality of detected misfit.

In closely related work, Levy, Mislevy, & Sinharay (2009) investigated the utility of several different discrepancy measures to check for multidimensionality when data were generated to have various forms of multidimensionality but were estimated with a (2PL) unidimensional IRT model. They found a Mantel-Haenszel statistic (MH; Agresti, 2002), model-based covariance (MBC; Reckase, 1997), and Yen's $Q_3$ (Yen, 1993) to be most effective at detecting multidimensionality in their conditions. Less effective bivariate measures included the covariance, residual item covariance (Fu et al., 2005),

natural log of the odds-ratio (Agresti, 2002), a standardized odds-ratio residual (Chen & Thissen, 1997), and the item-pair versions of $X^2$ and $G^2$ (e.g. Chen & Thissen, 1997). The latter two are nondirectional measures of association, which may partly explain their poor performance relative to the other bivariate measures. The univariate measures $X^2$, $G^2$, and proportion correct were found to be useless as detectors of multidimensionality. In an extension of the previous studies, Levy (2011) found that the same pattern of results was generally supported for models with conjunctive multidimensionality.

In studies where many hours of computing time are required, choosing among similarly performing discrepancy measures can be influenced by practical considerations such as how much processing time is required. For example, Levy (2006) reported that model-based covariance (MBC; Reckase, 1997) and $Q_3$ (Yen, 1993) performed similarly to each other as discrepancy measures of bivariate association (in MIRT and BN models), while generally performing better than the other investigated discrepancy measures. $Q_3$ possessed the additional benefit of having simpler computational requirements (i.e. less computing time), and was therefore the preferred discrepancy measure of bivariate association in subsequent studies in the same line of research (e.g. Levy, Crawford, Fay, & Poole, 2011).

Li, Cohen, Kim, and Cho (2009) compared five indices of model selection for mixture IRT models. The competing models of interest were non-nested, and therefore a likelihood ratio (LR) test could not be employed for model selection. The authors used a simulation study to investigate the most effective method for selecting the best-fitting model from among a group of candidate models. The competing methods of model selection included:  PPMC using a single discrepancy measure (OR), AIC, BIC, DIC, and

Bayes factor (PsBF). The authors generally recommended BIC as the preferred index, but results were complex. Perhaps their results would have been different if a different discrepancy measure had been used instead of OR. Studies cited in the previous paragraphs found OR to be inferior to other discrepancy measures for detecting multidimensionality in IRT models, but it remains an open question whether the preferred discrepancy measures in those studies would have performed better than OR in IRT mixture models in the Li et al. (2009) study. Although the focus of the present study was not model selection, the results of the present study could potentially help to inform researchers about which discrepancy measures to select when conducting studies of model selection. The choice of discrepancy measures is a crucial decision when implementing a PPMC framework, yet the number of studies devoted to recommending different discrepancy measures for different modeling purposes is underdeveloped.

In a simulation study that used a parametric bootstrapping framework to investigate the utility of various indices for detecting model misspecifications in BNs, Williamson, Almond, and Mislevy (2000) found Weaver's Surprise Index (WSI; Weaver, 1948), Ranked Probability Score (RPS; Epstein, 1969), and Good's Logarithmic Score (GLS; Good, 1952) to be the most effective fit functions. Overall, RPS was judged to be the most effective index, and was recommended for detecting the following model misspecifications (see Figure 2 for illustrations of applicable misspecifications): node inclusion (adding a variable that should not be in the model), node exclusion (omitting a variable that should be in the model, strong edge inclusion (including a strong dependency in the model between two variables that are not strongly associated in the data), strong edge exclusion (omitting a dependency from the model between two

variables that are strongly associated in the data, node state exclusion (omitting from the model a level of the variable that exists in the data), and prior probability errors (specifying prior probabilities in the model which are do not accurately represent the true population probabilities).

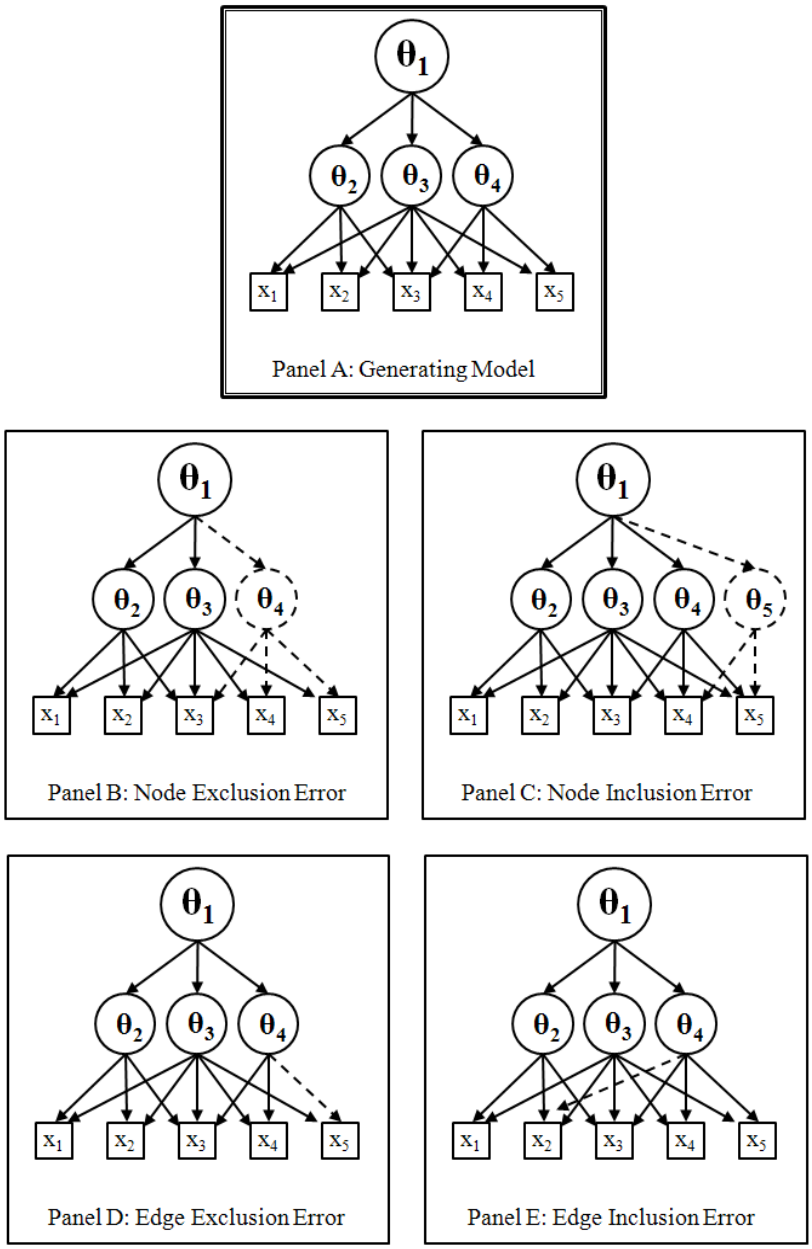



*Figure 2.* BN misspecifications. Panel A shows the generating model. Subsequent panels include dashed lines to illustrate how misspecified models differ with respect to the generating model. Adapted from Williamson, Almond, and Mislevy (2000).

Judged to be next most effective was WSI, which was recommended for detecting a similar slew of model misspecifications: node inclusion and exclusion, strong edge inclusion and exclusion, and weak edge exclusion. GLS was deemed third most useful despite being relatively less effective overall, because it detected types of misspecifications that were not detected by the other indices, namely node *state* inclusion and exclusion. The effectiveness of all indices was improved to some degree as sample size increased from 100 to 1000. Williamson and his coauthors called for future research on the generalizability of their findings to different BN structures---which is a contribution of the current study.

**Label Switching**

The ordering, naming, or numbering of categories of respondents within discrete models such as BNs or LCAs is arbitrary and unimportant within a given context, but it must remain consistent lest complications arise. Label switching refers to the problematic situation where alternative forms of the otherwise nominal assignment process are mixed together within the same analytical context. Label switching in BNs can obscure the underlying story that is told by parameter estimates, fit functions, and graphs (an example is provided in the results section). Results from any method that aggregates information across alternatively labeled solutions can be impacted. Previous research devoted to label switching (e.g. Chung, Loken, & Schafer, 2004; Stephens, 2000) has discussed a number of alternative procedures for fixing (avoiding) the problem, the most common of which are identifiability constraints and relabeling algorithms. This is an active area of research with much yet to be learned about the tradeoffs associated with various approaches.

## Summary

The current study aimed to help meet the growing demand for psychometric model checking tools for BNs by exploring the utility of several different types of fit functions for critiquing the fit of complex multidimensional BNs. This study differed from previous studies in important ways. First, the generating and scoring models in the present study were multidimensional, so it was unknown whether fit functions that had successfully detected evidence of multidimensionality in unidimensional models would be successful in this new role. Second, the BNs in this study were more complex than BNs reported in previous PPMC research. This complexity was not included simply to extend previous research, but was based on existing models being used in an innovative operational performance assessment (Rupp et al., 2012).

The potential toolkit for PPMC users is limitless due to the flexibility of PPMC to incorporate any fit function that may be of theoretical use, but the current toolkit for BN users is limited by the sparsity of examples in the literature for models like the ones included in this study. Simulation studies are particularly useful for investigating methodological tools because they allow researchers to know (and control) the properties of the data. The current study represented an exploratory step into a vast methodological space. Many of the design features could have been implemented in so many different ways and still have forged new ground. This study would have looked much different if it had been designed only from a methodological perspective. However, the study was motivated within the context of specific modeling experiences, based on repeated efforts to critique related BNs with a limited number of tools and wanting to know if additional tools could improve our ability to critique those models.

# Method

## Simulation Study

A simulation study was conducted to investigate the utility of conducting PPMC with a variety of fit functions to detect different types of misfit in complex BN models. The following subsections describe the features of the models used to simulate and fit the data. Later sections describe the fit functions and outcome variables.

**Manipulated Factors.** The various models described here are variations on a common theme, motivated by an existing complex performance assessment (Rupp et al., 2012), and briefly described here. The general latent structure consisted of three discretized latent variables, each measured by a subset of 33 dichotomously scored observed variables. The three latent variables represented sequentially-offered educational content, with mastery of subsequent content somewhat dependent upon mastery of previous content. The first latent variable ($\theta_1$) was the foundational latent construct. It was relatively easy for students to master, but was important for mastery of subsequent constructs ($\theta_2$ and $\theta_3$). The dependence among the latent variables will be discussed in further detail later, after other features of the models have been presented. Across all models, the theoretical importance of $\theta_1$ was evidenced by the relatively large number of observed variables devoted to its measurement compared to the numbers of observed variables measuring $\theta_2$ and $\theta_3$. Each observed variable represented specific aspects of a broad series of behaviors on an open-ended performance assessment.

BN models were manipulated along two factors: latent variable dependency structure and number of latent classes. The “latent variable dependency structure” factor had 3 levels (“simple”, “contextual”, “complex”,) and the “number of latent classes”

factor had 2 levels (“2 latent classes”, “3 latent classes”), resulting in a total of 6 different BNs. Figures 3-8 show diagrams of the six models, with accompanying conditional probability table (CPT) templates, which express the modeled dependencies among variables. The CPT templates are also presented independently in Tables 2-5, 7-9, 11-16, and 19-21.

***Simple structure: Models 1 and 2.*** As can be seen in Figures 3 and 4 respectively, Models 1 and 2 exhibited simple structure, which means that each observed variable measured only a single latent variable. The three latent variables ($\theta_1$, $\theta_2$, $\theta_3$) were measured respectively by fifteen, twelve, and six observables. The decision to have different numbers of measured variables per latent variable reflected a desire to retain fidelity to real-world models that motivated this study. It would be unlikely for task designers in this applied setting to restrict themselves to a uniform number of observed variables per latent construct. It was therefore of interest to investigate how discrepancies in the number of observed variables per latent variable might impact model criticism tools. Models 1 and 2 differed from each other along the second manipulated factor, with Model 1 having two classes per latent variable, and Model 2 having three classes per latent variable.

The *Q*-matrix for Models 1 and 2 is provided in Table 1. A *Q*-matrix features the complete list of observed variables as rows and the complete list of latent variables as columns.

Table 1

*Q-matrix for Models 1 and 2*

| Observable ($x_j$) | Number of Parents | $\theta_1$ | $\theta_2$ | $\theta_3$ |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 3 | 1 | 1 | 0 | 0 |
| 4 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 |
| 6 | 1 | 1 | 0 | 0 |
| 7 | 1 | 1 | 0 | 0 |
| 8 | 1 | 1 | 0 | 0 |
| 9 | 1 | 1 | 0 | 0 |
| 10 | 1 | 1 | 0 | 0 |
| 11 | 1 | 1 | 0 | 0 |
| 12 | 1 | 1 | 0 | 0 |
| 13 | 1 | 1 | 0 | 0 |
| 14 | 1 | 1 | 0 | 0 |
| 15 | 1 | 1 | 0 | 0 |
| 16 | 1 | 0 | 1 | 0 |
| 17 | 1 | 0 | 1 | 0 |
| 18 | 1 | 0 | 1 | 0 |
| 19 | 1 | 0 | 1 | 0 |
| 20 | 1 | 0 | 1 | 0 |
| 21 | 1 | 0 | 1 | 0 |
| 22 | 1 | 0 | 1 | 0 |
| 23 | 1 | 0 | 1 | 0 |
| 24 | 1 | 0 | 1 | 0 |
| 25 | 1 | 0 | 1 | 0 |
| 26 | 1 | 0 | 1 | 0 |
| 27 | 1 | 0 | 1 | 0 |
| 28 | 1 | 0 | 0 | 1 |
| 29 | 1 | 0 | 0 | 1 |
| 30 | 1 | 0 | 0 | 1 |
| 31 | 1 | 0 | 0 | 1 |
| 32 | 1 | 0 | 0 | 1 |
| 33 | 1 | 0 | 0 | 1 |

*Note.* A value of 0 indicates that the latent skill is not required to correctly complete the observed task, while a value of 1 indicates that it is.

The entries in the *Q*-matrix specify whether a latent skill is required by each observable variable. A value of "0" in the *Q*-matrix indicates that the latent skill is not required for successful completion of the observed variable, while a value of "1" indicates that it is required. Rows in the *Q*-matrix thus summarize the patterns of latent skills required for each item, and columns summarize the groups of observables requiring each latent skill.

All observed variables in Model 1 followed the same dependency structure with respect to their latent parent (see Figure 3).

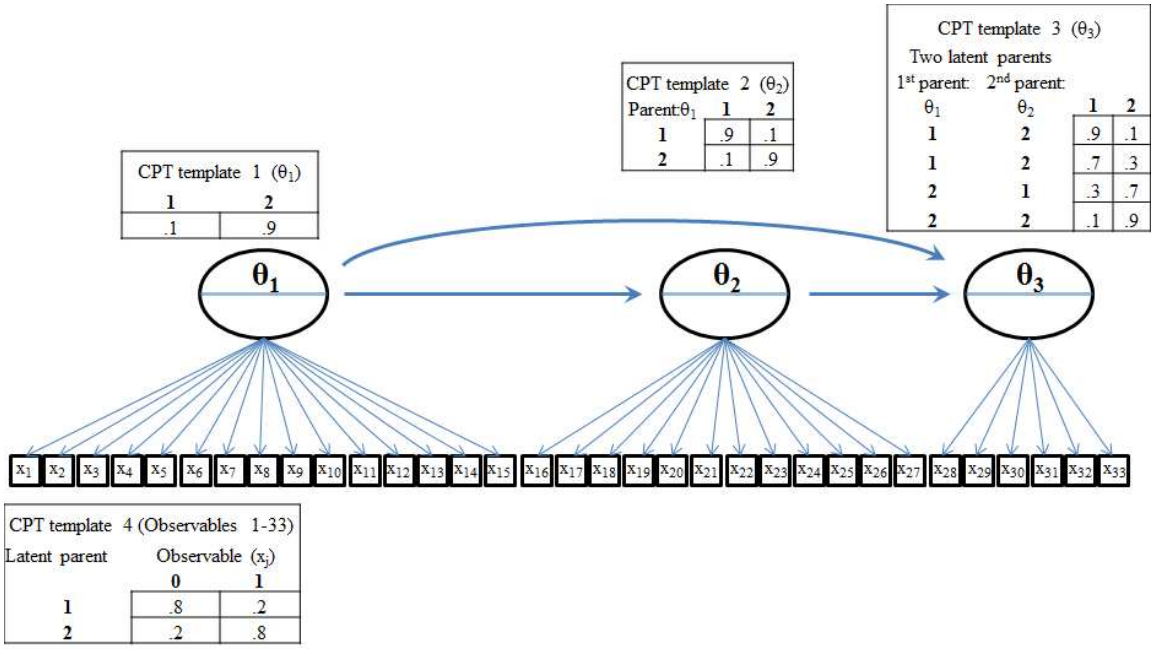



*Figure 3*. BN Generating Model 1: Simple structure, 3 latent variables, 2 latent classes.

Specifically, examinees with a value of 1 on the latent parent had a 20% probability of correctly completing the observable and an 80% probability of not completing the observable correctly, while examinees with a value of 2 on the latent parent had an 80% probability of correctly completing the observable and a 20% probability of not completing the observable correctly (see Table 2).

Table 2

*CPT Template 4*

| | Child value | |
|---|---|---|
| Parent value | 0 | 1 |
| 1 | .8 | .2 |
| 2 | .2 | .8 |

*Note*. This template applies to Observables 1-33 in Model 1 (see Figure 3), Observables 4-5, 9-10, 14-15, 19-21, and 25-27 in Model 3 (see Figure 5), and Observables 4-15, 19-27, and 31-33 in Model 5 (see Figure 7).

This represented a two-class solution in which the class with greater probability of success was conceived as relative masters of the construct, while the class with lower probability of success was conceived as relative non-masters. These observables discriminated strongly between the two classes of examinees, because there was a large difference between the conditional probabilities of a correct (or incorrect) response for the classes (0.8 - 0.2 = .6). The decision to hold constant the "quality" of the observables represented a choice of convenience. Observables in practice would be expected to vary with respect to this property. However, task designers always strive to create observables of high quality (discriminating power), so it was reasonable to investigate the properties of an assessment that held this desirable, albeit ambitious, property. In the context of this simulation study, varying the discrimination between observables would have created undesirable noise that could have obscured effects of greater interest. It was therefore believed that sacrificing this type of fidelity was worth the increased clarity with respect to prioritized purposes.

As can be seen from Figure 4, the observed variables in Model 2 followed one of two dependency structures in relation to their associated latent variables.

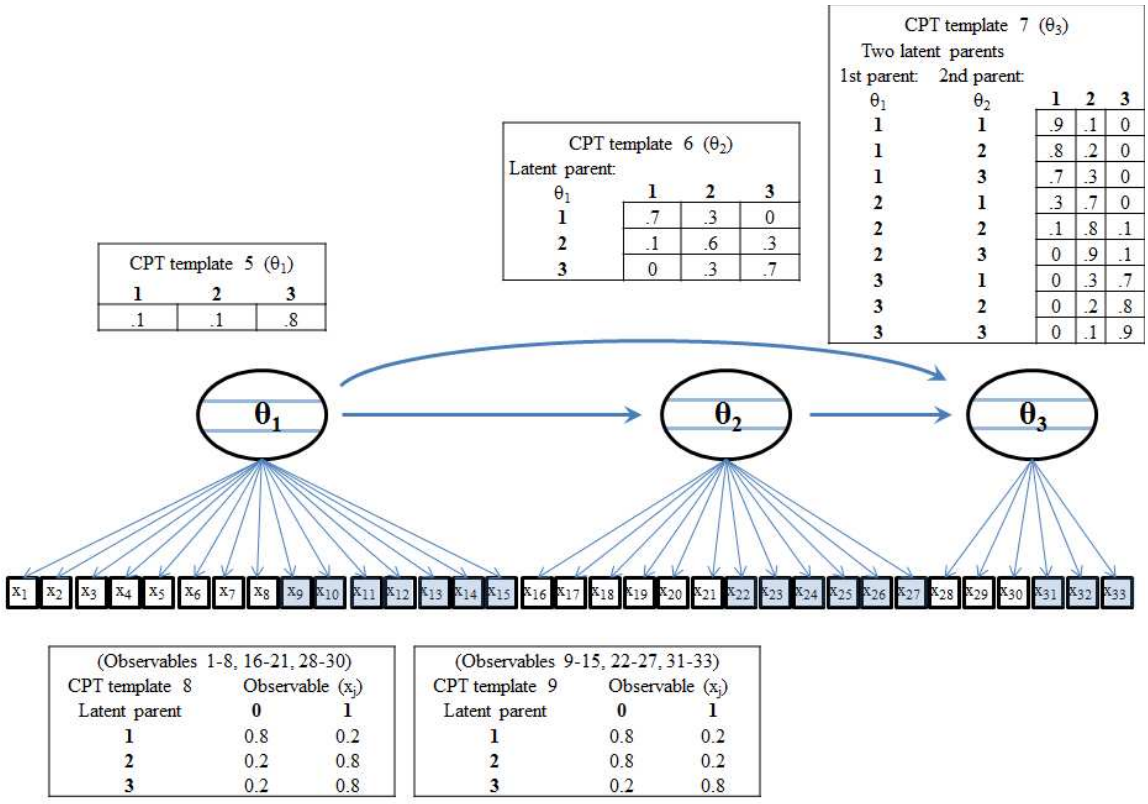



*Figure 4*. BN Generating Model 2: Simple structure, 3 latent variables, 3 latent classes.

The increased complexity compared to Model 1 was due to the addition of a third class of examinees per latent variable ($\theta_1$, $\theta_2$, $\theta_3$). The number of latent classes represented a convenient and theoretically relevant way to alter model complexity. A model with fewer classes is more parsimonious and more restrictive because it classifies students into a smaller number of distinct categories even though their patterns of responses retain the same variability as in the comparison model. For example, a model with two classes (e.g., content master and non-master) posits that students can only be categorized into these two groups, according to their propensity to correctly complete the observed tasks. Additional classes allow for greater flexibility regarding the classification of response patterns (e.g., Mastery, Partial-mastery, and Non-mastery levels). Holding

constant the other factor (dependency structure), models with three latent classes per latent variable were expected to fit better than their more restrictive, two-class counterparts. The number of latent classes was convenient to manipulate in the sense that it did not require modifications to the DAG. Speaking generally, additional classes cause the number of estimated conditional probabilities to increase precipitously, which can impede or even prevent estimation.

Note in Tables 3 and 4 that Class 1 had the same 20% probability of success across all observables as was the case in Model 1, and Class 1 retained its interpretability as a low-performing or non-mastery class. Similarly, Class 3 represented the high-performing or mastery class having 80% probability of successfully completing each observable task (as did Class 2 in Model 1).

Table 3

*CPT Template 8*

| | Child value | |
|---|---|---|
| Parent value | 0 | 1 |
| 1 | .8 | .2 |
| 2 | .2 | .8 |
| 3 | .2 | .8 |

*Note*. This template applies to Observables 1-8, 16-21, and 28-30 in Model 2 (see Figure 4), Observables 4, 5 and 19-21 in Model 4 (see Figure 6), and Observables 4-8 and 19-21 in Model 6 (see Figure 8).

Table 4

*CPT Template 9*

| | Child value | |
|---|---|---|
| Parent value | 0 | 1 |
| 1 | .8 | .2 |
| 2 | .8 | .2 |
| 3 | .2 | .8 |

*Note*. This template applies to Observables 9-15, 22-27, and 31-33 in Model 2 (see Figure 4), Observables 9, 10, 14-15, and 25-27 in Model 4 (see Figure 6), and Observables 9-15, 22-27, and 31-33 in Model 6 (see Figure 8).

The additional class was the middle-performing or partial mastery class. This middling class performed as the mastery class on some observables but performed as the non-mastery class on the remaining observables. Specifically, examinees within Class 2 (middle class) in Model 2 had an 80% probability of correctly completing Observables 1-8, 16-21, and 28-30 (represented in Figure 4 by white squares), and a 20% probability of correctly responding on Observables 9-15, 22-27, and 31-33 (represented in Figure 4 by shaded squares). Any single observable discriminated strongly between two classes of examinees but was unable to distinguish the third class. It was the performance *across* observables that distinguished the additional class in Model 2 (see also Models 4 and 6), not relative performance on any single observable. This pattern of performance across observables represented a particular hypothesis of partial mastery, namely that partial mastery consisted of the ability to do well on some constituent tasks but not others. By contrast, an alternative conception of partial mastery (not represented in the present study) might consist of in-between probabilities of success across all (or some) constituent observables. For example, on a given observable the mastery class might have

an 80% probability of success, the non-mastery class a 20% probability of success, and the partial-mastery class a 50% probability of success.

***Context effects: Models 3 and 4.*** As can be seen in Figures 5 and 6 respectively, Models 3 and 4 had seven additional latent variables compared to Models 1 and 2. The additional latent variables ($\theta_4 - \theta_{10}$) in Models 3 and 4 were measured by non-overlapping subsets of the same observed variables that measured the three latent variables common to all the models ($\theta_1$, $\theta_2$, $\theta_3$).

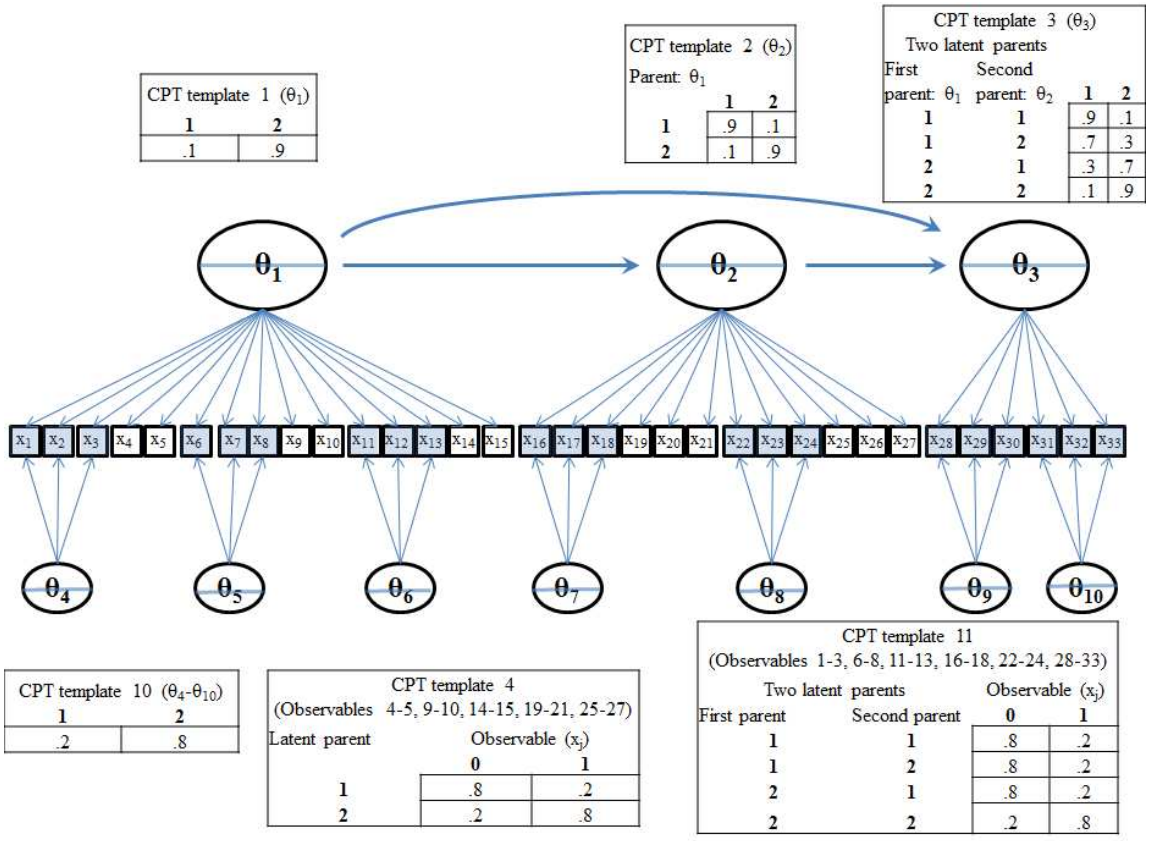



*Figure 5*. BN Generating Model 3: Context effects, 10 latent variables, 2 latent classes.

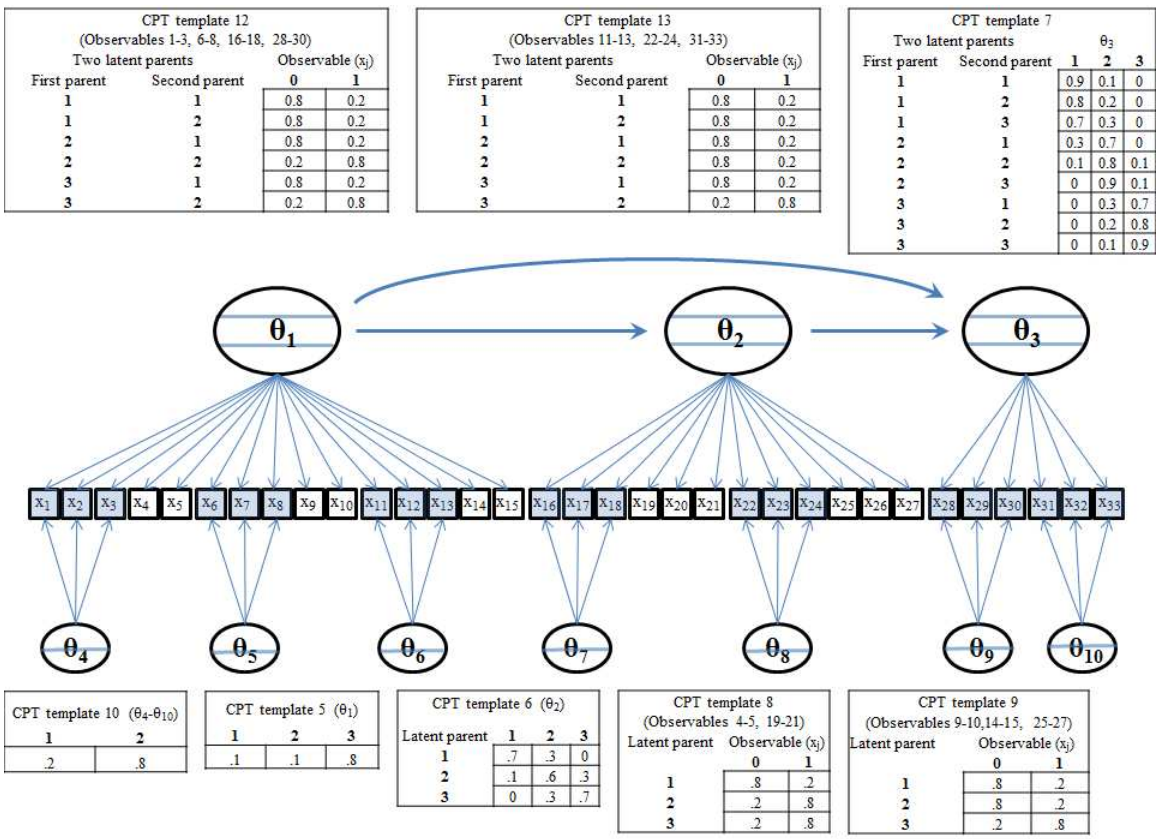



*Figure 6.* BN Generating Model 4: Context effects, 10 latent variables, 3 latent classes.

These additional latent variables were conceived as contextual variables that shared some residual dependence not measured by the "primary" latent constructs. For example, in the domain of computer networking that motivated this study, the observed variables were open-ended tasks that simulated real-world situations that computer networking technicians are faced with when configuring computing devices as part of a network. Clusters of observed variables might be associated by device (e.g. router, switch, personal computer, printer, server, etc.) or by instructional prompts that are a function of a specific testing environment. Rimjen (2010) showed that bi-factor models can be constrained into testlet models and second-order models, which are shown to be formally equivalent. The bi-factor model was implemented as part of the present study

because it represented the most general (flexible) of these variants. Having a high value on the contextual variables might be thought of as the answers to questions like: "Did the examinee understand how to apply their knowledge and skills in the context of this specific device typology?" or "Did the examinee understand how the assessment instructions applied to this cluster of tasks?". It can be seen in Figures 5 and 6 that the contextual latent variables $\theta_4$ - $\theta_{10}$ had no parents, so the proportions provided in CPT template 10 (see Table 5) were the marginal class memberships for these variables.

Table 5

*CPT Template 10*

| | Latent value | |
|---|---|---|
| | 1 | 2 |
| Probability | .2 | .8 |

*Note*. This template applies to $\theta_4$- $\theta_{10}$ in Model 3 (see Figure 5) and Model 4 (see Figure 6).

These proportions indicate that 80% of the examinees (in the population) possess the knowledge and skills implied by a contextual latent variable, and that 20% of the examinees do not. The choice to have a relatively large proportion of students possess each context variable reflects the theoretical position that context variables in practice are not usually designed to impede students. Contextual variables are conceived as representing challenges to some students, but generally aligning with proficiency on the primary latent variable. The choice to hold this proportion constant across all contextual variables reflected a desire to simplify this component of the design, as opposed to the more realistic option of letting contextual effects vary across latent variables. Future research could explore alternatives of these decisions. For a more detailed account of some different types of contextual effects that have been modeled in CDMs, including

inhibitory effects like those modeled in the present study, see Almond, Mulder, Hemat, and Yan (2009). The *Q*-matrix for Models 3 and 4 is provided in Table 6.

Table 6

*Q-matrix for Models 3 and 4*

| Observable ($x_j$) | Number of Parents | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ | $\theta_9$ | $\theta_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 7 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 8 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 12 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 13 | 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 14 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 17 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 18 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 19 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 22 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 23 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 24 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 25 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 26 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 27 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 29 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 30 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 31 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 32 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 33 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

*Note*. A value of 0 indicates that the latent skill is not required to correctly complete the observed task, while a value of 1 indicates that it is.

Models 3 and 4 differed from each other along the second manipulated factor, with Model 3 having two classes per latent variable and Model 4 having three classes per latent variable. Note that two classes were always estimated for each contextual latent variable ($\theta_4$-$\theta_{10}$) regardless of whether there were two or three classes per $\theta_1$-$\theta_3$. The decision to hold constant the number of classes per contextual variable was due to a theoretical conception which viewed them as being present or absent, but not varying categorically within each context. By contrast, varying the strength of the context effects *across* $\theta_4$-$\theta_{10}$ could be sensible theoretically, but it was not manipulated in the present study. Future research could explore this issue.

As can be seen from Table 7, all observed variables with two latent parents in Model 3 (Observables 1-3, 6-8, 11-13, 16-18, 22-24, and 28-33, represented in Figure 5 by shaded squares) followed a more complex dependency structure in relation to their associated latent variables than did the observables with one latent parent (Observables 4-5, 9-10, 14-15, 19-21, and 25-27, represented in Figure 5 by white squares).

Table 7

*CPT Template 11*

| | | Child value | |
|---|---|---|---|
| Parent 1 | Parent 2 | 0 | 1 |
| 1 | 1 | .8 | .2 |
| 1 | 2 | .8 | .2 |
| 2 | 1 | .8 | .2 |
| 2 | 2 | .2 | .8 |

*Note*. This template applies to Observables 1-3, 6-8, 11-13, 16-18, 22-24, and 28-33 in Model 3 (see Figure 5), and Observables 1, 3, 16, 18, 28, and 30 in Model 5 (see Figure 7).

The modeled relationships here were conjunctive, meaning that both latent constructs were required for having a strong (80%) probability of correctly completing the observed

task. Examinees with only one of the requisite abilities, or neither, had a lower (20%) probability of correctly completing the task.

As was the case in Model 2, observables with one latent parent in Model 4 (see Figure 6) followed one of two dependency patterns (specified by Tables 3 and 4) depending on whether the partial mastery class responded as the mastery or non-mastery class on a particular observable. Similarly, observables with two latent parents in Model 4 followed one of two dependency structures according to the differential behavior of the partial mastery class, but with the necessary level of added complexity due to the role of the additional latent variables (specified by Tables 8 and 9).

Table 8

*CPT Template 12*

| | | Child value | |
|---|---|---|---|
| Parent 1 | Parent 2 | 0 | 1 |
| 1 | 1 | .8 | .2 |
| 1 | 2 | .8 | .2 |
| 2 | 1 | .8 | .2 |
| 2 | 2 | .2 | .8 |
| 3 | 1 | .8 | .2 |
| 3 | 2 | .2 | .8 |

*Note*. This template applies to Observables 1-3, 6-8, 16-18, and 28-30 in Model 4 (see Figure 6).

Table 9

*CPT Template 13*

| | | Child value | |
|---|---|---|---|
| Parent 1 | Parent 2 | 0 | 1 |
| 1 | 1 | .8 | .2 |
| 1 | 2 | .8 | .2 |
| 2 | 1 | .8 | .2 |
| 2 | 2 | .8 | .2 |
| 3 | 1 | .8 | .2 |
| 3 | 2 | .2 | .8 |

*Note*. This template applies to Observables 11-13, 22-24, and 31-33 in Model 4 (see Figure 6).

Note that because there was no partial mastery class with respect to the contextual variables (those variables had two classes across all models), examinees lacking the contextual skill always performed as the non-mastery class regardless of their value for the primary latent variable. In other words, examinees with skill profiles [1,1], [2,1], and [3,1] each had the same 20% probability of success across all observables with two latent parents. By contrast, examinees who did possess the contextual skill differed in their probabilities of success according to their skill level on the primary latent variable, such that the middle class performed as the mastery class on Observables 1-3, 6-8, 16-18, and 28-30 and as the non-mastery class on Observables 11-13, 22-24, and 31-33.

***Complex Structure: Models 5 and 6.*** As can be seen in Figures 7 and 8 respectively, Models 5 and 6 exhibited complex structure, meaning some observed variables measured more than one primary latent variable. Observables 1, 3, 16, 18, 28, and 30 had two latent parents (represented in Figures 7 and 8 by lighter shading), while Observables 2, 17, and 29 had three latent parents (represented in Figures 7 and 8 by darker shading). The three latent variables ($\theta_1$, $\theta_2$, $\theta_3$) were measured respectively by

nineteen, sixteen, and ten observables; the increased number of measured variables per latent variable relative to Models 1 and 2 was due to added cross-loadings between the aforementioned observables and their latent parents. Models 5 and 6 differed from each other along the second manipulated factor, with Model 5 having two classes per latent variable, and Model 6 having three classes per latent variable. The *Q*-matrix for Models 5 and 6 is provided in Table 10.

Table 10

*Q-matrix for Models 5 and 6*

| Observable ($x_j$) | Number of Parents | $\theta_1$ | $\theta_2$ | $\theta_3$ |
|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 0 |
| 2 | 3 | 1 | 1 | 1 |
| 3 | 2 | 1 | 0 | 1 |
| 4 | 1 | 1 | 0 | 0 |
| 5 | 1 | 1 | 0 | 0 |
| 6 | 1 | 1 | 0 | 0 |
| 7 | 1 | 1 | 0 | 0 |
| 8 | 1 | 1 | 0 | 0 |
| 9 | 1 | 1 | 0 | 0 |
| 10 | 1 | 1 | 0 | 0 |
| 11 | 1 | 1 | 0 | 0 |
| 12 | 1 | 1 | 0 | 0 |
| 13 | 1 | 1 | 0 | 0 |
| 14 | 1 | 1 | 0 | 0 |
| 15 | 1 | 1 | 0 | 0 |
| 16 | 2 | 1 | 1 | 0 |
| 17 | 3 | 1 | 1 | 1 |
| 18 | 2 | 0 | 1 | 1 |
| 19 | 1 | 0 | 1 | 0 |
| 20 | 1 | 0 | 1 | 0 |
| 21 | 1 | 0 | 1 | 0 |
| 22 | 1 | 0 | 1 | 0 |
| 23 | 1 | 0 | 1 | 0 |
| 24 | 1 | 0 | 1 | 0 |
| 25 | 1 | 0 | 1 | 0 |
| 26 | 1 | 0 | 1 | 0 |
| 27 | 1 | 0 | 1 | 0 |
| 28 | 2 | 1 | 0 | 1 |

| | | | | |
|---|---|---|---|---|
| 29 | 3 | 1 | 1 | 1 |
| 30 | 2 | 0 | 1 | 1 |
| 31 | 1 | 0 | 0 | 1 |
| 32 | 1 | 0 | 0 | 1 |
| 33 | 1 | 0 | 0 | 1 |

*Note*. A value of 0 indicates that the latent skill is not required to correctly complete the observed task, while a value of 1 indicates that it is.

As can be seen from Figure 7, all observed variables with one latent parent in Model 5 followed the same dependency structure in relation to their associated latent variables as did the observables with one latent parent in Models 1 and 3 (see Table 2).

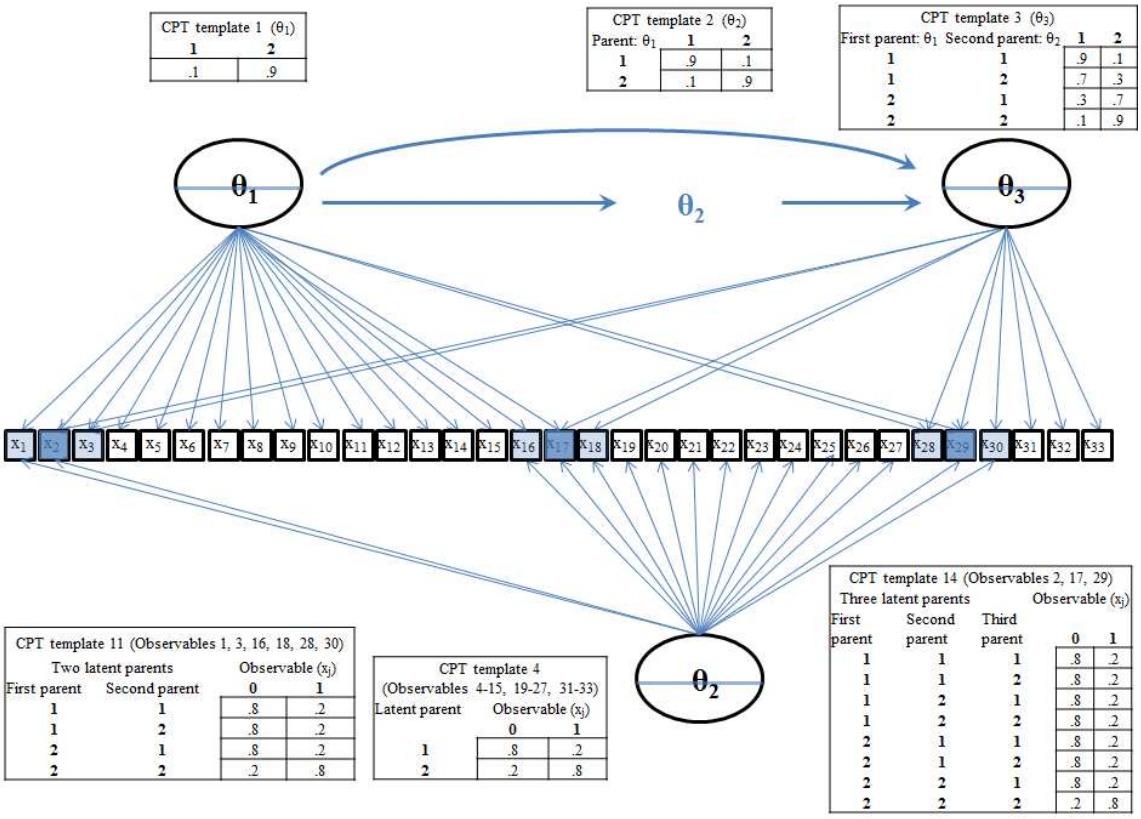



*Figure 7*. BN Generating Model 5: Complex structure, 3 latent variables, 2 latent classes.

Similarly, all observed variables with two latent parents in Model 5 followed Table 7 as did the observables with two latent parents in Model 3 (Model 1 did not have any observables with multiple parents). The observables with three latent parents followed the

specifications shown in Table 11, where it can be seen that these observables discriminated strongly between the examinees that did or did not possess all three latent parent variables.

Table 11

*CPT Template 14*

| | | | Child value | |
|---|---|---|---|---|
| Parent 1 | Parent 2 | Parent 3 | 0 | 1 |
| 1 | 1 | 1 | .8 | .2 |
| 1 | 1 | 2 | .8 | .2 |
| 1 | 2 | 1 | .8 | .2 |
| 1 | 2 | 2 | .8 | .2 |
| 2 | 1 | 1 | .8 | .2 |
| 2 | 1 | 2 | .8 | .2 |
| 2 | 2 | 1 | .8 | .2 |
| 2 | 2 | 2 | .2 | .8 |

*Note*. This template applies to Observables 2, 17, and 29 in Model 5 (see Figure 7).

Examinees possessing all three latent skills had a strong (80%) probability of completing the observables correctly, while examinees with two, one, or none of the requisite skills had a low (20%) probability of success.

For Model 6 (see Figure 8) observables with one latent parent in followed one of two dependency patterns (specified by Tables 3 and 4) depending on whether the partial mastery class responded as the mastery or non-mastery class on a particular observable(as was the case in Models 2 and 4). Examinees were required to have at least partial mastery (a value of 2 or 3) on all requisite skills in order to have an 80% probability of correctly completing an observable with multiple parents.

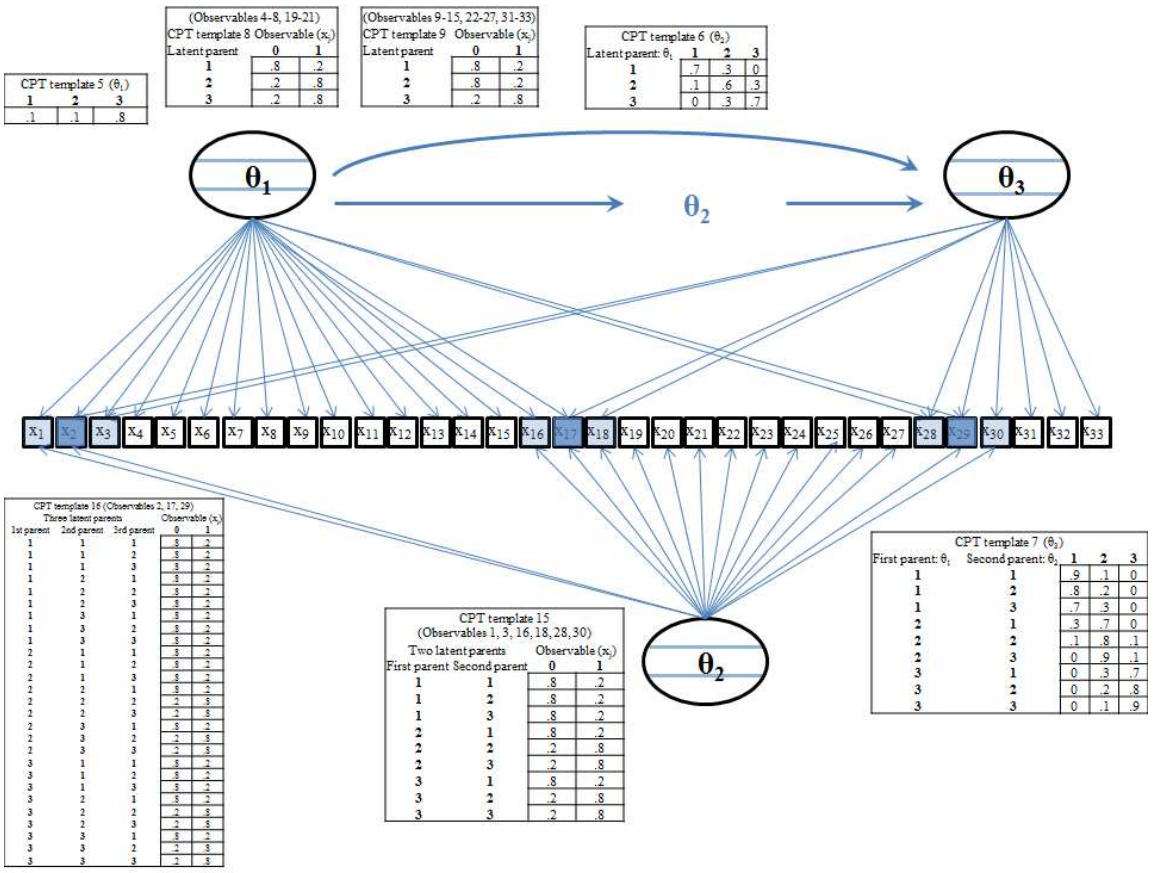



*Figure 8*. BN Generating Model 6: Complex structure, 3 latent variables, 3 latent classes.

Note in Model 6 that the cross-loadings were associated with observables where the partial mastery class responded as the mastery class. Consequently, observables with two latent parents in Model 6 followed a single dependency structure specified by Table 12, while observables with three latent parents followed a single dependency structure specified by Table 13.

Table 12

*CPT Template 15*

| | | Child value | |
|---|---|---|---|
| Parent 1 | Parent 2 | 0 | 1 |
| 1 | 1 | .8 | .2 |
| 1 | 2 | .8 | .2 |
| 1 | 3 | .8 | .2 |
| 2 | 1 | .8 | .2 |
| 2 | 2 | .2 | .8 |
| 2 | 3 | .2 | .8 |
| 3 | 1 | .8 | .2 |
| 3 | 2 | .2 | .8 |
| 3 | 3 | .2 | .8 |

*Note*. This template applies to Observables 1, 3, 16, 18, 28, and 30 in Model 6 (see Figure 8).

Table 13

*CPT Template 16*

| | | | Child value | |
|---|---|---|---|---|
| Parent 1 | Parent 2 | Parent 3 | 0 | 1 |
| 1 | 1 | 1 | .8 | .2 |
| 1 | 1 | 2 | .8 | .2 |
| 1 | 1 | 3 | .8 | .2 |
| 1 | 2 | 1 | .8 | .2 |
| 1 | 2 | 2 | .8 | .2 |
| 1 | 2 | 3 | .8 | .2 |
| 1 | 3 | 1 | .8 | .2 |
| 1 | 3 | 2 | .8 | .2 |
| 1 | 3 | 3 | .8 | .2 |
| 2 | 1 | 1 | .8 | .2 |
| 2 | 1 | 2 | .8 | .2 |
| 2 | 1 | 3 | .8 | .2 |
| 2 | 2 | 1 | .8 | .2 |
| 2 | 2 | 2 | .2 | .8 |
| 2 | 2 | 3 | .2 | .8 |
| 2 | 3 | 1 | .8 | .2 |
| 2 | 3 | 2 | .2 | .8 |
| 2 | 3 | 3 | .2 | .8 |
| 3 | 1 | 1 | .8 | .2 |
| 3 | 1 | 2 | .8 | .2 |
| 3 | 1 | 3 | .8 | .2 |
| 3 | 2 | 1 | .8 | .2 |
| 3 | 2 | 2 | .2 | .8 |

| | | | | |
|---|---|---|---|---|
| 3 | 2 | 3 | .2 | .8 |
| 3 | 3 | 1 | .8 | .2 |
| 3 | 3 | 2 | .2 | .8 |
| 3 | 3 | 3 | .2 | .8 |

*Note*. This template applies to Observables 2, 17, and 29 in Model 6 (see Figure 8).

***Latent dependency structures: Models 1, 3, and 5.*** Looking across Models 1, 3, and 5, note that the same latent dependency structure was maintained among $\theta_1$, $\theta_2$, and $\theta_3$. Generally speaking, the choices and specifications detailed hereafter regarding the latent dependency structures were motivated by previous findings within a research program at the Cisco Networking Academy. A hierarchy was implied by this structure, with $\theta_1$ functioning as a parent of $\theta_2$ and $\theta_3$, and $\theta_2$ as a parent of $\theta_3$. It can be seen that $\theta_1$ had no parents, so the proportions provided in Table 14 were the marginal class memberships for this variable.

Table 14

*CPT Template 1*

| | Latent value | |
|---|---|---|
| | 1 | 2 |
| Probability | .1 | .9 |

*Note*. This template applies to $\theta_1$ in Model 1 (see Figure 3), Model 3 (see Figure 5), and Model 5 (see Figure 7).

These proportions indicate that 90% of the examinees possessed the knowledge and skills implied by this latent variable, and that 10% of the examinees did not. Shifting attention to Table 15, it can be seen that the knowledge and skills represented by $\theta_1$ were important for acquiring the knowledge and skills represented by $\theta_2$:

Table 15

*CPT Template 2*

| | Child value | |
|---|---|---|
| Parent value | 1 | 2 |
| 1 | .9 | .1 |
| 2 | .1 | .9 |

*Note*. This template applies to $\theta_2$ in Model 1 (see Figure 3), Model 3 (see Figure 5), and Model 5 (see Figure 7).

Among students who possessed $\theta_1$, 90% also possessed $\theta_2$, while 10% lacked $\theta_2$. Similarly, of the students who lacked $\theta_1$, 90% also lacked $\theta_2$, while 10% possessed $\theta_2$. As can be seen in Table 16, the knowledge and skills represented by $\theta_1$ were also important for acquiring the knowledge and skills represented by $\theta_3$, but the relationship was complicated by the influence of $\theta_2$, which was also useful for acquiring $\theta_3$, but not as strongly as $\theta_1$.

Table 16

*CPT Template 3*

| | | Child value | |
|---|---|---|---|
| Parent 1 | Parent 2 | 1 | 2 |
| 1 | 1 | .9 | .1 |
| 1 | 2 | .7 | .3 |
| 2 | 1 | .3 | .7 |
| 2 | 2 | .1 | .9 |

*Note*. This template applies to $\theta_3$ in Model 1 (see Figure 3), Model 3 (see Figure 5), and Model 5 (see Figure 7).

One consequence of retaining the same latent dependencies across these models was that the marginal model-implied latent class memberships remained constant as well (see Table 17), with the caveat that for Model 3 the addition of the contextual latent

variables created many additional subclasses (Table 18 may be useful for this conception).

Table 17

*Marginal latent variable proficiencies for Generating Models 1, 3, and 5*

| Latent Profile | $\theta_1$ | $\theta_2$ | $\theta_3$ | marginal membership |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | .081 |
| 2 | 1 | 1 | 2 | .009 |
| 3 | 1 | 2 | 1 | .007 |
| 4 | 1 | 2 | 2 | .003 |
| 5 | 2 | 1 | 1 | .027 |
| 6 | 2 | 1 | 2 | .063 |
| 7 | 2 | 2 | 1 | .081 |
| 8 | 2 | 2 | 2 | .729 |

*Note*. For Model 3 the 8 latent profiles shown here represent aggregations across the contextual latent variables (see Table 18).

Table 18

*Marginal latent variable proficiencies for Generating Model 3*

| Latent Profile | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ | $\theta_9$ | $\theta_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 129 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 257 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 385 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 513 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 641 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 769 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 897 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| … | … | … | … | … | … | … | … | … | … | … |
| 1,024 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

*Note*. Rows have been collapsed (…) due to space considerations.

These model-implied memberships were calculated by multiplying the three conditional probability values that were relevant to each latent profile. For example, the first row in Table 17 corresponds to Latent Profile 1, where students lacked each of the latent variables $\theta_1$, $\theta_2$, and $\theta_3$. The probability associated with Level 1 in Table 14 (0.1) was multiplied by the analogous probabilities from Tables 15 and 16 (0.9 and 0.9 respectively), yielding 0.1 * 0.9 * 0.9 = .081, or 8.1%. These same model-implied memberships can be imposed upon Model 3 if one thinks of the various configurations of contextual latent proficiencies as subsets within the eight latent profiles characterized by proficiency patterns on the three primary latent variables. Table 18 illustrates that for each primary proficiency profile there were 128 contextual proficiency profiles, resulting in a total of 1,024 proficiency profiles for Model 3. Within each primary profile, the memberships were uniformly distributed due to the fact that each contextual latent variable was exogenous and was governed by CPT Template 10 (see Table 5).

***Latent dependency structures: Models 2, 4, and 6.*** Looking across Models 2, 4, and 6, note that the same latent dependency structures were maintained among $\theta_1$, $\theta_2$, and $\theta_3$. The patterns described in the previous section for Models 1, 3, and 5 generally hold for these models as well, with the added complexity of a third latent class (see Tables 19-21).

Table 19

*CPT Template 5*

| | Latent value | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Probability | .1 | .1 | .8 |

*Note*. This template applies to $\theta_1$ in Model 2 (see Figure 4), Model 4 (see Figure 6), and Model 6 (see Figure 8).

Table 20

*CPT Template 6*

| | Child value | | |
|---|---|---|---|
| Parent value | 1 | 2 | 3 |
| 1 | .7 | .3 | 0 |
| 2 | .1 | .6 | .3 |
| 3 | 0 | .3 | .7 |

*Note*. This template applies to $\theta_2$ in Model 2 (see Figure 4), Model 4 (see Figure 6), and Model 6 (see Figure 8).

Table 21

*CPT Template 7*

| | | Child value | | |
|---|---|---|---|---|
| Parent 1 | Parent 2 | 1 | 2 | 3 |
| 1 | 1 | .9 | .1 | 0 |
| 1 | 2 | .8 | .2 | 0 |
| 1 | 3 | .7 | .3 | 0 |
| 2 | 1 | .3 | .7 | 0 |
| 2 | 2 | .1 | .8 | .1 |
| 2 | 3 | 0 | .9 | .1 |
| 3 | 1 | 0 | .3 | .7 |
| 3 | 2 | 0 | .2 | .8 |
| 3 | 3 | 0 | .1 | .9 |

*Note*. This template applies to $\theta_3$ in Model 2 (see Figure 4), Model 4 (see Figure 6), and Model 6 (see Figure 8).

The marginal model-implied latent class memberships for these three models are provided in Table 22.

Table 22

*Marginal latent variable proficiencies for Generating Models 2, 4, and 6*

| Latent Profile | $\theta_1$ | $\theta_2$ | $\theta_3$ | marginal membership |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | .063 |
| 2 | 1 | 1 | 2 | .007 |
| 3 | 1 | 1 | 3 | 0 |

| | | | | |
|---|---|---|---|---|
| 4 | 1 | 2 | 1 | .024 |
| 5 | 1 | 2 | 2 | .006 |
| 6 | 1 | 2 | 3 | 0 |
| 7 | 1 | 3 | 1 | 0 |
| 8 | 1 | 3 | 2 | 0 |
| 9 | 1 | 3 | 3 | 0 |
| 10 | 2 | 1 | 1 | .003 |
| 11 | 1 | 1 | 2 | .007 |
| 12 | 2 | 1 | 3 | 0 |
| 13 | 2 | 2 | 1 | .006 |
| 14 | 2 | 2 | 2 | .048 |
| 15 | 2 | 2 | 3 | .006 |
| 16 | 2 | 3 | 1 | 0 |
| 17 | 2 | 3 | 2 | .027 |
| 18 | 2 | 3 | 3 | .003 |
| 19 | 3 | 1 | 1 | 0 |
| 20 | 3 | 1 | 2 | 0 |
| 21 | 3 | 1 | 3 | 0 |
| 22 | 3 | 2 | 1 | 0 |
| 23 | 3 | 2 | 2 | .048 |
| 24 | 3 | 2 | 3 | .192 |
| 25 | 3 | 3 | 1 | 0 |
| 26 | 3 | 3 | 2 | .056 |
| 27 | 3 | 3 | 3 | .504 |

*Note.* For Model 4 the 27 latent profiles shown here represent aggregations across the contextual latent variables (see Table 23).

Finally, Table 23 illustrates that for each of the 27 primary proficiency profiles for Model 4 there were 128 contextual proficiency profiles, resulting in a total of 3,456 proficiency profiles. The model-implied memberships shown in Table 22 represent the collective memberships of the 128 contextual profiles within each primary profile. Within each primary profile, the memberships were uniformly distributed due to the fact that each contextual latent variable was exogenous and was governed by CPT Template 10 (see Table 5).

Table 23

*Marginal latent variable proficiencies for Generating Model 4*

| Latent Profile | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ | $\theta_9$ | $\theta_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 129 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 257 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 385 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 513 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 641 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 769 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 897 | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1025 | 1 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1153 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1281 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1409 | 2 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1537 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1665 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1793 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1921 | 2 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2049 | 2 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2177 | 2 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2305 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2433 | 3 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2561 | 3 | 1 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2689 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2817 | 3 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2945 | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3073 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3201 | 3 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3329 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3456 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

*Note*. Most rows have been omitted due to space considerations.

In summary, all investigated models shared the same number of observed variables (33), but varied with respect to the structure of latent variables. Models with simple or complex structure had three latent variables, while context-effect models had ten latent variables. Latent structure was manipulated because it represented an important

type of model modification that researchers employ in practice. Due in part to computational demands, it was beyond the scope of the current study to manipulate latent structure more extensively. Ideally it would be beneficial to vary the latent structures to a greater degree, and to investigate the implications of various model modification strategies, but those developments were left to future research. In the present study, the choices were intended to represent the most common and important strategies that would have relevance to the applied researchers who motivated the study.

**Conditions.** Each of the six BN models was used as a generating model, and for each generating model a subset of the same BN models was used as scoring models, resulting in a total of 11 conditions (see Table 24).

Table 24

*Table of Conditions*

| Model Description | Generating Model | Scoring Model 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Simple structure 2 latent classes | 1 | | ■ | ■ | ■ | ■ | ■ |
| Simple structure 3 latent classes | 2 | | | ■ | ■ | ■ | ■ |
| Context effects 2 latent classes | 3 | | ■ | | ■ | ■ | ■ |
| Context effects 3 latent classes | 4 | | ■ | ■ | | ■ | ■ |
| Complex structure 2 latent classes | 5 | | ■ | ■ | ■ | | ■ |
| Complex structure 3 latent classes | 6 | | ■ | ■ | ■ | ■ | |

*Note*. White square indicates condition is included in the study, shaded square indicates condition is not included in the study.

Conditions were denoted by abbreviations for the generating and scoring models separated by a period. For example, “1.1” indicates that Model 1 was used as the

generating model and the scoring model, while “5.1” indicates that Model 5 was used as the generating model and Model 1 was used as the scoring model.

Fully crossing the manipulated factors was not deemed necessary because some conditions resulting from a fully crossed design offered mostly redundant information. For example, consider the first row of Table 24, where data were generated using the most restrictive model (Model 1). Using the same model (Model 1) as the scoring model was a necessary step because this condition served as a control group. However, using Model 2 as a scoring model for data generated from Model 1would not provide valuable fit information. The more restricted version (Model 1) was a special case which can be obtained from Model 2 by restricting membership in the third latent class to zero. There was perhaps something to be learned in such conditions about the efficiency of estimation routines, the impact of maintaining a constant sample size when estimating increasing numbers of parameters, etc., but there would have been diminishing returns with respect to the performance of data-model fit techniques. Given the relatively steep cost in computing time per condition in this study, using a scoring model that was known to be a more general case of a generating model did not represent an efficient use of resources. The discussion section provides approximations of the computing time required to complete the simulation component of this study.

**Replications.** Each condition was replicated 100 times. Replications within the same condition differed from each other due only to sampling variability, which refers to the effects of using random processes to obtain a sample from a population of potential values. The purpose of replication in this context was to mitigate the effects of sampling variability by obtaining a larger sample of exchangeable studies drawn randomly from

the population of studies to which they belong. The choice of how many replications to conduct was a tradeoff between resources and generalizability. Each replication was somewhat costly in terms of computing time, yet it was important to have enough replications to ameliorate the influence of sampling variability on interpretations made from the study. It was believed that 100 replications struck an appropriate balance between these opposing considerations, influenced by previous research and available computing resources.

**Sample size.** Sample size in the context of the present study refers to the number of simulees used to estimate the parameters of the models, which was 1000. Varying the sample size would have increased the study's ability to generalize its findings to studies using other (most likely smaller) sample sizes. However, adding even one additional sample size would have greatly increased the total computational time required for this study, so it did not seem justified relative to the inclusion of other elements (e.g. more model variants or fit functions) that were more central to the purposes of the study. Previous studies have well established the finding that model criticism tools perform better as sample size increases (e.g. Williamson et al., 2000), so it was believed that computational resources were better utilized for other design considerations. A sample size of 1000 might be considered large in the context of some research settings, but was relatively small in the context of the Cisco networking academy, and therefore represented a very realistic baseline from which to assess these model-checking procedures.

**Estimation.** WinBUGS version 1.4.3 (Spiegelhalter, Thomas, Best, & Lunn, 2007) was used to conduct the MCMC estimation via the R2Winbugs package in R

version 3.0.0 (R Core Development Team, 2013). Three independent chains were used, with start values drawn randomly from probability distributions spanning the range of potential parameter values (when possible; see label switching subsections of method and results sections for more details). Convergence was assessed using a criterion of approximately 1.0 on the Brooks-Gelman-Rubin diagnostic (BGR; Brooks and Gelman, 1998) in conjunction with visual inspection of trace plots from pilot replications. Autocorrelations from pilot replications were inspected to determine the necessity of thinning.

**Label Switching.** Label switching was handled using a strategy of assigning the most unambiguous response patterns from each data set to theoretically appropriate latent classes, as opposed to estimating the latent class memberships for those simulees. For example, a response pattern of all 1's (i.e. a perfect score on the exam) was assigned to the mastery class for each primary latent variable. Due to the fact that samples (of N=1000) did not always contain enough perfect scores to anchor each latent class in the "correct" labeling orientation, a variety of the most unambiguous response patterns were included. Across all replications and latent variables, the average number of memberships assigned in this way per latent variable was approximately 78 (out of 1000).

For Models 3 and 4, which included seven contextual latent variables in addition to the three primary latent variables, a more complicated strategy was necessary. Assigning values on the contextual latent variables based on response patterns alone was not sufficient because there were only three observables per contextual latent variable. In these models, constraints were imposed upon the conditional probabilities of observables with a contextual latent variable parent, such that the estimated probability of

successfully completing an observable for the non-mastery class could not exceed that of a partial-mastery class, and the estimated probability of success for a partial-mastery class could not exceed that of a mastery class. These constraints, in conjunction with the strategy of assignments on the primary latent variables for unambiguous response patterns, were sufficient to prevent label switching in most replications. In the results section, a modification is described that eliminated the observed label switching in all subsequent replications.

**Fit Functions**

A total of thirteen fit functions were included in this study (SGDDM was used in five ways). Functions were selected to address different levels of misfit. Table 25 lists the fit functions and their levels of analysis within the PPMC framework implemented in the present study.

Table 25

*Fit functions and their levels of analysis*

| Fit function | Level of analysis |
|---|---|
| Deviance | global |
| Proportion Correct | observable |
| $Q_3$ | pairs of observables |
| SGDDM | global; subscales; pairs of observables |
| $\chi^2$-type index | observable |
| Ranked Probability Score (RPS) | observable |
| Good's Logarithmic Scale (GLS) | observable |
| Hierarchical Consistency Index (HCI) | person |
| Item Consistency Index (ICI) | observable |

**Deviance.** Deviance is a global measure of data-model fit. Evaluating deviance within an NRD framework is possible for model comparison purposes, but not for evaluating the fit of a single model in isolation. The deviance scale does not lend itself to absolute interpretations because the properties of the likelihood function vary with respect to model features. Within an HT framework, deviance is computed as a single number per replication and evaluated as a chi-square test with approximate degrees of freedom n-(p+1), where n is the number of independent observations and p is the number of estimated parameters. Within PPMC, there is a distribution of realized deviance values, based on the observed data and the posterior distribution of the parameters, and there is a distribution of posterior predicted values, based on the posterior predictive data and the posterior distribution of parameters. Both distributions (realized and posterior predictive) are represented by the same sample of posterior draws.

**Proportion correct.** Proportion correct is computed at the observable level directly from data, as opposed to requiring model parameters. The inclusion of this fit function was primarily for verifying that PPMC programming code was functioning properly. Speaking generally, proportion correct is a feature of model fit that is easy to reproduce, even for models that fit poorly according to other DMs.

$\boldsymbol{Q_3}$. $Q_3$ was the only fit function in this study that could have been evaluated within all of the model-checking frameworks. There was some known redundancy with SGDDM in the sense that both indices evaluate associations between observables as a test of local dependence. $Q_3$ has been a popular choice in past PPMC research, so comparing the performance of $Q_3$to SGDDM within this study helped to establish the utility of SGDDM and helped to expand the generalizability of previous $Q_3$findings.

**Standardized generalized dimensionality discrepancy measure (SGDDM).** SGDDM was applied at three different levels along the global-local spectrum. At the most local level, it was applied to each pair of observables while aggregating across examines. Secondly, it was applied to the sets of observables associated with each of three primary latent variables. This second level of aggregation was akin to a subscale level. Thirdly, SGDDM was aggregated at the global level, meaning that the full set of 33 observables was included.

**$\chi^2$-type item fit index.** This fit function was included primarily because it was one of the few fit functions that had been demonstrated in the BN literature. Additionally, $\chi^2$ tests have been used commonly in IRT for item fit, so the performance of this function may be of interest to a broader audience.

**Ranked probability score (RPS).** RPS was an appealing fit function because it performed well in a previous BN simulation study (Williamson, Almond, and Mislevy, 2000), and because it can be aggregated across observables or examinees. In the present study, RPS was aggregated at the observable level. Evaluating RPS within an NRD framework is only possible when adopting a model-comparison approach due to the fact that cutoff values have not been established in relevant modeling contexts. Furthermore, an analytical reference distribution has not been proposed for RPS, so evaluation within an HT framework is not yet possible.

**Good's logarithmic scale (GLS).** In previous BN research (Williamson, Almond, and Mislevy, 2000) GLS was more successful at detecting state misspecification errors than RPS, despite better overall performance by RPS. Given that one of the manipulated factors in this study was the number of states per latent variable (latent classes), it was of

interest to assess whether GLS would be effective in the current study. Evaluating GLS within an NRD framework is only possible when adopting a model-comparison approach due to the fact that cutoff values have not been established in relevant modeling contexts. Furthermore, an analytical reference distribution has not been proposed for GLS, so evaluation within an HT framework is not yet possible.

**Hierarchy consistency index (HCI).** HCI and ICI (below) are analogues of each other, aggregated across different units. HCI assesses person fit by aggregating across observables, while the ICI assesses observable fit by aggregating across examinees. HCI and ICI were included in this study because they were recently developed for use in CDMs. Their utility for BNs has not yet been established, but conceptually they seemed well-suited for the present application. These indices were designed for use in conjunctive models only. The dependency relationships in the present study are not strictly conjunctive, but they can be viewed as approximately conjunctive.

**Item consistency index (ICI).** ICI assesses fit at the level of observables. The creators of ICI proposed a criterion of .5 (Lai, Gierl, and Cui, 2012) for evaluating whether an observable fits, with values above .5 (i.e., from .5 to 1) indicating adequate fit and values below .5 (i.e., -1 to .5) indicating misfit. This criterion corresponds to an observable with at least 75% of its observed responses matching the responses expected by *Q*-matrix specifications.

## Outcome Variables

The fit functions in this study were conceptualized along two dimensions: effectiveness and efficiency. Effectiveness was defined as the propensity to correctly identify data-model misfit, while efficiency referred to the amount of computing time

required. If two fit functions took the same amount of time to compute, then the function with increased effectiveness was preferred. Similarly, if two fit functions were equivalent in terms of their effectiveness, then the function requiring less time would be preferred. If two fit functions differed with respect to their effectiveness and efficiency, then deciding between them became more situation-specific and user-dependent. The following outcomes, with the exception of computing time, were intended to help researchers evaluate the effectiveness of the fit functions with regard to detecting data-model misfit. The inclusion of computing time was intended to help researchers evaluate efficiency, and therefore to inform a researcher about the tradeoffs of using various fit functions.

**PPP-values.** The primary outcome measure of this study was the distribution of PPP-values. In addition to graphical presentations, these distributions were summarized using median values across replications, and proportions of replications in which the PPP-value was "extreme". From a Bayesian perspective, PPP-values should not be interpreted with respect to a cutoff value. However, to facilitate comparisons to other frameworks, extreme PPP-values were defined as $< .025$ or $> .975$, or in other words the 5% most extreme PPP-values (akin to $\alpha = .05$). Note that in null conditions (i.e. when the scoring model was the same as the generating model) this outcome measure represented an empirical Type-I error rate, and in misspecified conditions it reflected observed power. For localized fit functions, heat maps were used to summarize findings across observables or observable pairs. Squares in the heat maps were shaded to represent categorical ranges of values.

**Effect Size.** An effect size measure was created to help summarize information not revealed by the PPP-values, namely the magnitude of the differences between

realized and posterior predicted values. The mean difference between realized and posterior predicted values was divided by the standard deviation of those same differences:

$$ES = \frac{\sum_{n=1}^{N} D_n(\boldsymbol{y}, \boldsymbol{\Theta}) - D_n(\boldsymbol{y^{rep}}, \boldsymbol{\Theta})}{N} \Big/ \sqrt{\frac{\sum_{n=1}^{N} (D_n(\boldsymbol{y}, \boldsymbol{\Theta}) - D_n(\boldsymbol{y^{rep}}, \boldsymbol{\Theta}))^2}{N}} \qquad (21),$$

where *n* was one of *N* draws from the posterior distribution, $D(\boldsymbol{y}, \boldsymbol{\Theta})$ were values of the discrepancy measure using the observed data and $D(\boldsymbol{y^{rep}}, \boldsymbol{\Theta})$ were values of the discrepancy measure using replicated data. The metric was therefore standard deviation units of the differences, which varied across fit functions. Conceptually, the PPP-value is a measure of how often posterior predicted values exceed realized values, with no distinction made for the degree of excess. The effect size is meant to quantify the magnitude of the differences between realized and posterior predicted values on a scale that is standardized with respect to the variability of those differences.

Larger effect sizes are driven either by larger numerator terms (holding constant the denominator), or by smaller denominator terms (holding constant the numerator), or by both factors in conjunction. The main reason for a large numerator is systematically large differences between realized and posterior predicted values. The main reason for a small denominator is small variability in the differences between realized and posterior predicted values, irrespective of the size of those differences.

**Computing time.** Computing time was evaluated descriptively, with representative examples drawn for illustrative purposes. Including computing time as a formal factor would have created many logistical problems, including standardization of computing resources across conditions. The inclusion of computing time as an outcome

was meant to help characterize the results with respect to the investment of practical resources. General statements including approximate computing time were included, but precise computational comparisons were not attempted.

## Results

### MCMC

A burn-in of 100 iterations was used for most conditions; the exceptions were Conditions 3.3 and 4.4 which had a burn-in of 700, and Condition 6.6 which had a burn-in of 3000. A thinning factor of 10 was adequate to minimize autocorrelations for most conditions, while a factor of 20 was used for Conditions 3.3, 4.4, and 6.6. In all conditions, sufficient iterations were run to yield 100 draws from each of three chains. A total of 300 draws was used to represent the posterior distribution in the PPMC analysis conducted in R.

### Label Switching

As described in the method section, the practice of assigning top-performing and bottom-performing response patterns (simulees) to mastery and non-mastery latent classes respectively was theoretically sufficient to prevent label switching (e.g. Chung, Loken, & Schafer, 2004), but label switching nevertheless occurred intermittently in a minority of replications (the number of affected replications ranged from 0 to 63 across conditions with a mean of 29). The problem first presented itself as two distinct clusters of points in deviance PPP-scatterplots where a single cluster was expected (examples for comparison are shown in Figures 9 and 10), though other indications were subsequently discovered elsewhere.

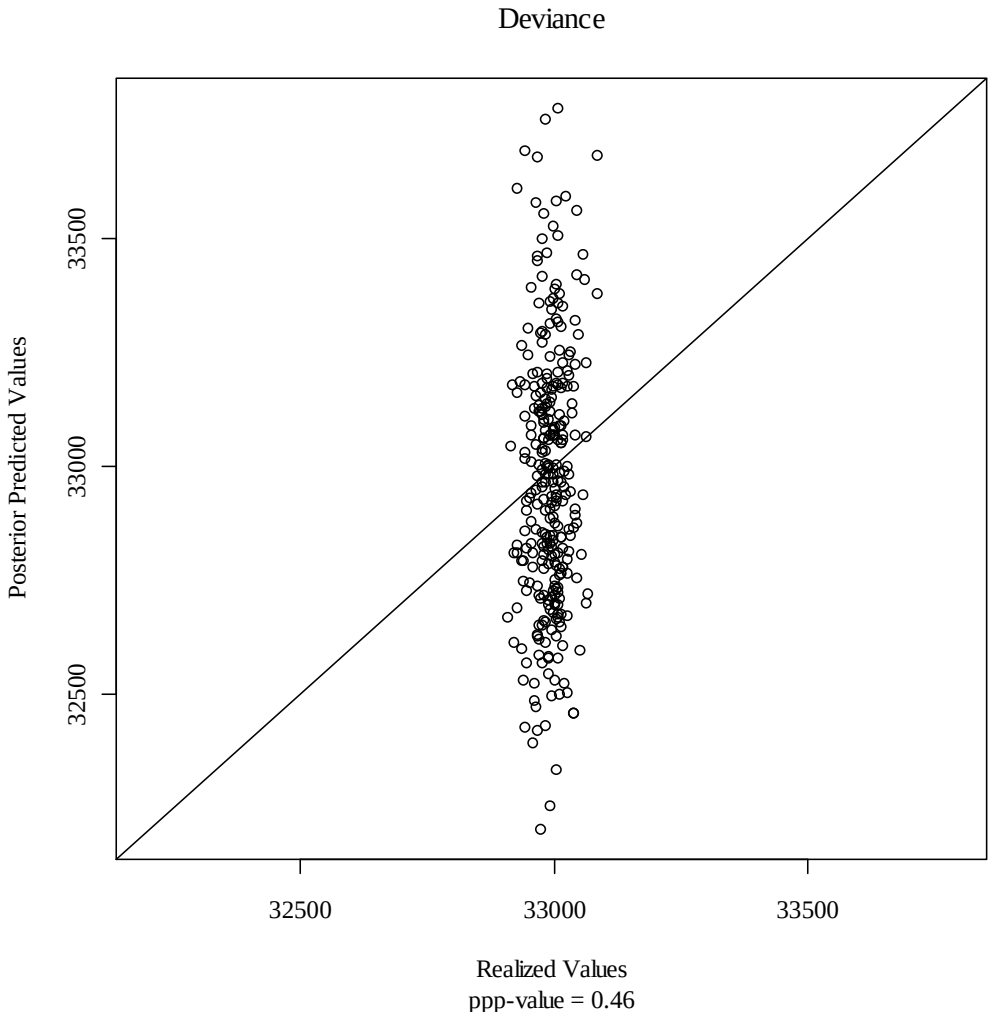



*Figure 9.* Scatterplot of deviance values from a typical replication of Condition 1.1.

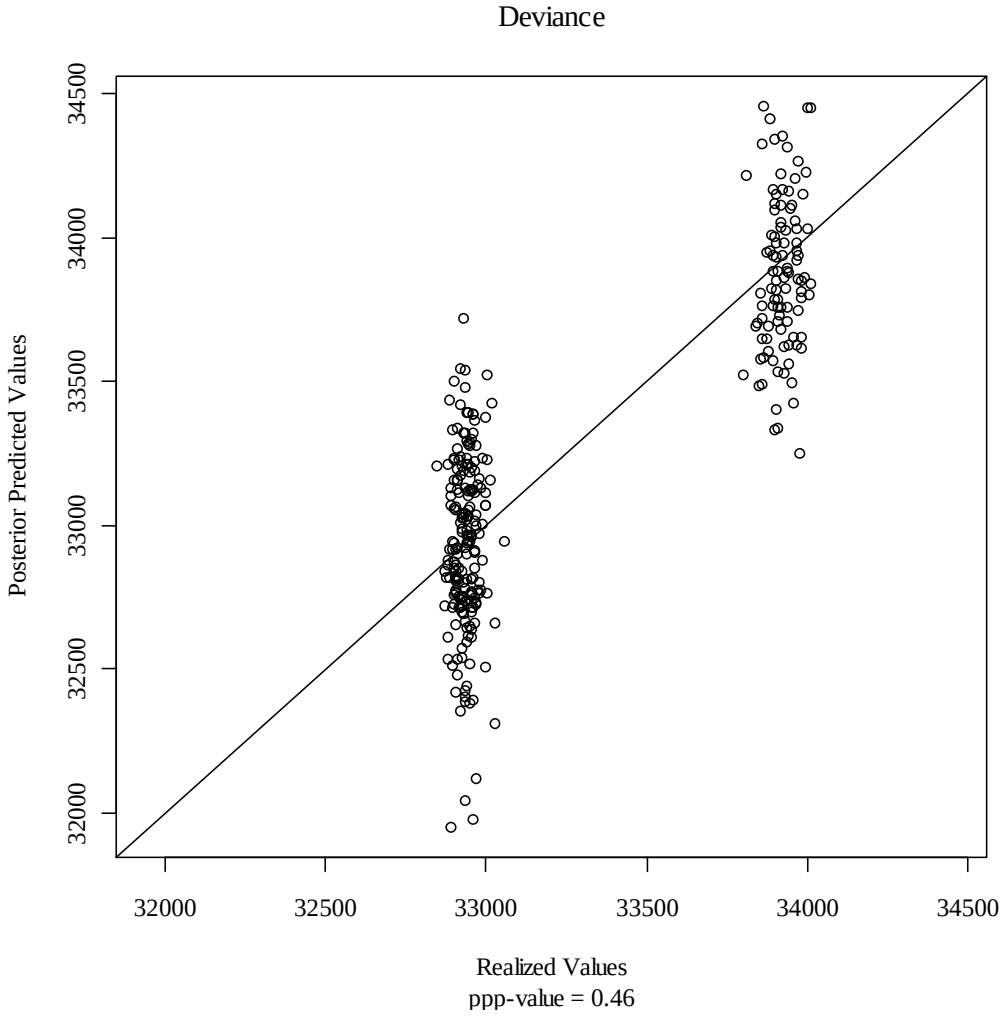



*Figure 10*. Scatterplot of deviance values from a replication of Condition 1.1 with "partial label switching".

Upon further investigation it was determined that label switching was affecting one or more (but not all) of the latent variables. The affected variable(s) was not consistent. We began referring to this phenomenon as "partial label switching" to differentiate it from the type of label switching we had warded against in previous research where all latent variables were affected. Theoretically, all of the MCMC chains in this study would have moved to and from these alternatively labeled solutions if allowed an infinite number of

iterations. A sudden switch from one labeling system to another would be caused by an extreme draw of candidate values. However, in practice no such within-chain switching was observed. Any given chain remained internally consistent with regard to its labeling in the finite number of observed draws, but across chains it was evident that alternative labeling existed. In affected replications, typically two chains stabilized on the "correctly" labeled solution and the third chain stabilized on a partially label-switched solution, although rarely it was observed that two chains exhibited partial label switching and the third was "correctly" labeled.

Alternative methodologies were explored for eliminating the occurrence of partially mislabeled solutions (e.g. stronger a priori class assignments, different MCMC updater methods, restricting latent variable CPT parameter values, restricting observable CPT parameter values, and post-hoc relabeling). It was beyond the scope of this study to investigate label switching methodologies systematically, but the issues mentioned briefly here could be investigated in future research. Suffice it to say that the approach adopted here was to place restrictions on the start values for some parameters (in addition to retaining the initial methodology of assigning extremely unambiguous response patterns to specific classes). This approach compared favorably to other attempted methods in terms of its effectiveness and timeliness, and seemed to come at a reasonable price in terms of assumptions.

In the initial methodology, all start values had been drawn from uniform distributions that spanned the entire range of possible parameter values. In hindsight this choice was less desirable because it permitted the label-switched solutions in some replications, whereas a less conservative approach would have avoided them altogether.

However, even after imposing range restrictions on some start values, start values were still more widely dispersed than in previous research with similar models (e.g. Levy et. al, 2011), so it could be argued that even the modified methodology was relatively conservative. The parameters were still allowed to vary over a comparatively wide range of the possible values; they were just restricted relative to the initial settings. Note that restricting the start values in this way did not (further) restrict the parameter values; it simply restricted the locations where the searches for the posterior distribution were allowed to begin. The replications that exhibited partial label switching were stored for reference purposes and for potential future research, and additional replications were run using the restricted start value methodology.

**Distributions of PPP-values**

Figure 11 shows smoothed density plots of the distributions of PPP-values for each of the 13 fit functions, pooled into two groups, defined as the six null conditions and the five misspecified conditions.

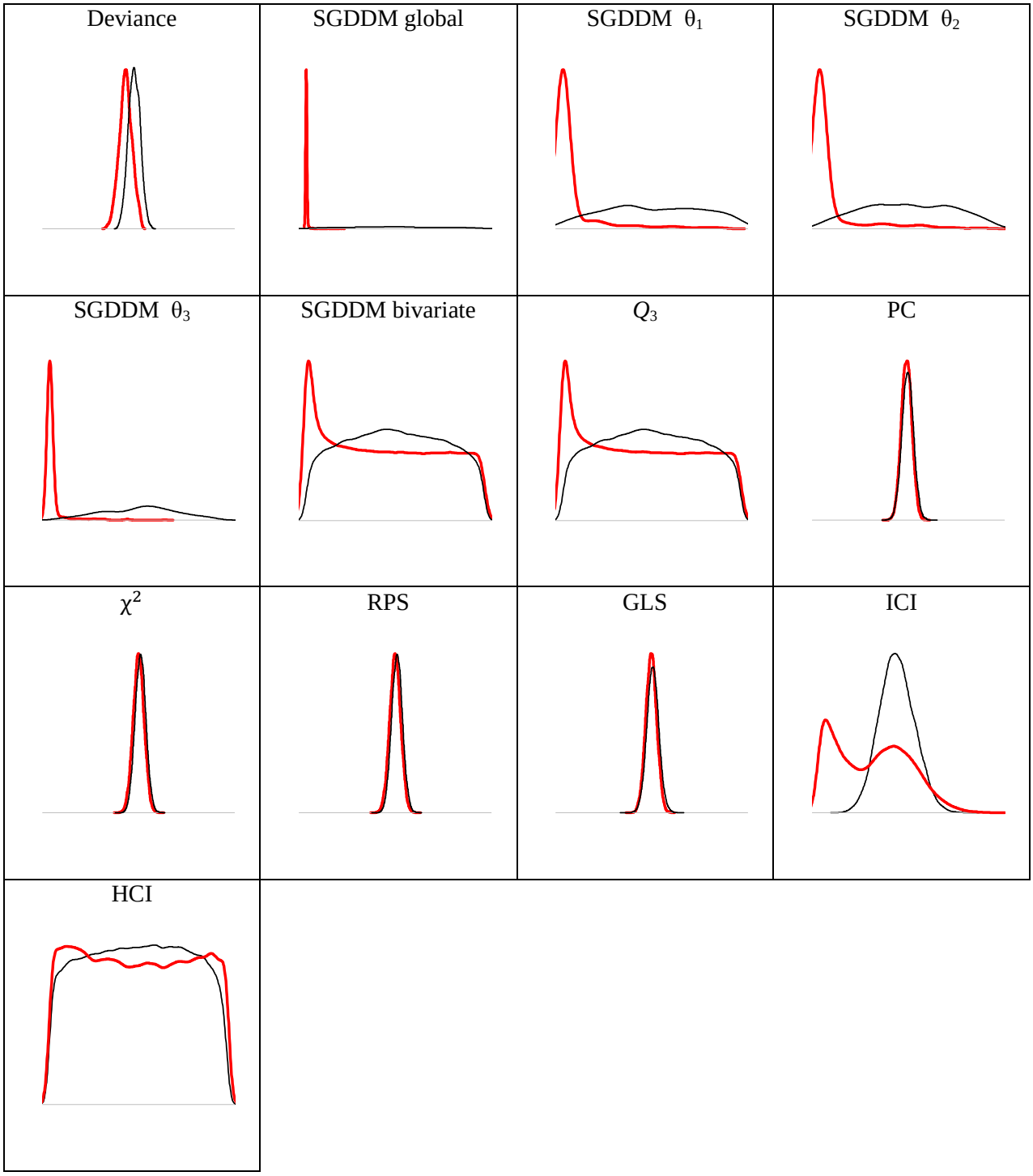



*Figure 11*. Distributions of PPP-values pooled across conditions. Misspecified conditions are represented by thicker lines and null conditions are represented by thinner lines. The x-axis of each panel spans the full range of possible PPP- values (0 to 1). The y-axis of each panel is proportional to frequency.

Within each panel, the x-axis represents the full range of possible PPP-values from 0 to 1 and the y-axis is proportional to frequency. For the first five panels representing deviance and the global and subscale aggregations of SGDDM respectively, each condition is represented by 100 PPP-values (one per replication), for a total of 600 values when pooled across the six null conditions, or 500 values when pooled across the five misspecified conditions. Additional pooling exists for the other fit functions, which consist of finer grain sizes and have multiple PPP-values per replication. The observable-level fit functions (PC, $\chi^2$, RPS, GLS, and ICI) contribute one PPP-value per observable in each replication, or 3300 values per condition. The densities pooled across null and misspecified conditions represent 19,800 and 16,500 PPP-values respectively. By comparison, the bivariate fit functions SGDDM and $Q_3$ contribute one PPP-value for each of 528 unique pairings of observables per replication, or 52,800 values per condition. The densities pooled across null and misspecified conditions represent 316,800 and 264,000 PPP-values respectively. Lastly, the person-level fit function HCI contributes one PPP-value for each of 1000 simulees per replication, or 100,000 values per condition. The densities pooled across null and misspecified conditions represent 600,000 and 500,000 PPP-values respectively. This level of aggregation is not ideal for most of the fit functions, but it is useful for highlighting the low relative utility of some of the fit functions before moving on to more appropriate views for the more promising functions.

In a hypothesis-testing framework, uniformity in the null distribution offers a number of attractive features, including producing Type-I error rates at the nominal level. The benefits of uniformity have also been advocated from a Bayesian, non-hypothesis-

testing viewpoint (e.g. Berkhof, van Mechelen, & Gelman, 2004). Therefore one criterion of good performance for each of the null distributions in Figure 11 is to be uniform throughout the range of possible values (0 to 1). However, the most important feature when comparing the two densities within a panel of Figure 11 is the extent to which they can be distinguished from one another, because even if the shapes are far from ideal, separation between the two indicates the potential for cutoff values to be developed, albeit perhaps heuristically. In practice, researchers obtain a single PPP-value that summarizes the relationship between realized DM values and posterior predicted DM values, but the observed PPP-value is itself a member of a different (meta) sampling distribution that can only be viewed in a simulation study where it is known how the realized data were generated. Hypothetically speaking, if a given pair of PPP-value sampling distributions were completely separate, then any observed PPP-value would with certainty indicate whether or not a model misspecification existed, irrespective of the degree of fit indicated by the PPP-value itself. Conversely, if a pair of PPP-value sampling distributions overlapped completely, then any observed PPP-value would be useless as an indicator of model misspecification because sampling variability alone would be equally likely to have produced the observed PPP-value (i.e. there exists no difference to detect between the sampling distributions of PPP-values).

For example, consider the densities of PPP-values for deviance in the first panel. It can be seen that all observed PPP-values were located near the center of the possible range, but that misspecified conditions tended to produce smaller PPP-values than null conditions.

Using conventional hypothesis-testing criteria in practice would result in no power to detect the misfit that existed in misspecified conditions (see first column of Table 26).

Table 26

*Proportion of PPP-values flagged as extreme across replications by condition and fit function*

| | Fit Function | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | Deviance | SGDDM global | SGDDM $\theta_1$ | SGDDM $\theta_2$ | SGDDM $\theta_3$ | SGDDM bivariate | $Q_3$ | PC | $\chi^2$ | RPS | GLS | ICI | HCI |
| 1.1 | 0 | .02 | .10 | .05 | .01 | .04 | .04 | 0 | 0 | 0 | 0 | 0 | .05 |
| 2.1 | 0 | 1 | 1 | 1 | 1 | .09 | .09 | 0 | 0 | 0 | 0 | .09 | .06 |
| 2.2 | 0 | .10 | .04 | 0 | 0 | .03 | .03 | 0 | 0 | 0 | 0 | 0 | .03 |
| 3.1 | 0 | 1 | 1 | 1 | 1 | .06 | .06 | 0 | 0 | 0 | 0 | .03 | .07 |
| 3.3 | 0 | 0 | 0 | .01 | 0 | .02 | .02 | 0 | 0 | 0 | 0 | 0 | .04 |
| 4.1 | 0 | 1 | 1 | 1 | 1 | .08 | .08 | 0 | 0 | 0 | 0 | .28 | .07 |
| 4.4 | 0 | 0 | .02 | 0 | 0 | .01 | .01 | 0 | 0 | 0 | 0 | 0 | .03 |
| 5.1 | 0 | .93 | .18 | .19 | .58 | .05 | .05 | 0 | 0 | 0 | 0 | 0 | .05 |
| 5.5 | 0 | .07 | .12 | .07 | .02 | .03 | .04 | 0 | 0 | 0 | 0 | 0 | .05 |
| 6.1 | 0 | 1 | 1 | 1 | 1 | .10 | .10 | 0 | 0 | 0 | 0 | .06 | .07 |
| 6.6 | 0 | .02 | .04 | .01 | 0 | .03 | .03 | 0 | 0 | 0 | 0 | 0 | .03 |

By contrast, alternative criteria could theoretically be constructed by considering the location where the null and misspecified densities cross (see Hjort, Dahl, & Steinbakk, 2006). Observed PPP-values below that threshold would suggest that the source of the realized data was a misspecified model because sampling variability alone was less likely to produce PPP-values that low.

The distributions for the poorest-performing group of fit functions (PC, $\chi^2$, RPS, and GLS) were far from uniform across all null conditions. The PPP-values for these functions were centered properly near .5 but barely dispersed, and there was virtually no separation between null and alternative distributions. Consequently, these functions were excessively conservative in null conditions and powerless in misspecified conditions. Though not apparent from the viewpoint offered by Figure 11, the distributions of PPP-values *per observable* (a more meaningful aggregation for observable-level functions) were all similarly shaped. These functions did not show differential performance across observables. Further presentation of the results for these fit functions was therefore omitted.

The distributions of deviance PPP-values were shaped similarly to the previous group of poor-performing functions, with the important distinction that there was some separation between null and alternative distributions in terms of location. In other words, despite a dramatic departure from the ideal of uniformity in the null case, the separation between distributions would make it possible to specify a cutoff value for use in practice. It was beyond the goals of this study to investigate recommended cutoff values for fit functions, but these results suggest that it would be possible to do so if deviance was needed as a global fit function for some theoretical reason. However, across all misspecified conditions in this study, the deviance PPP-value in every replication was less extreme (closer to .5) than the SGDDM global PPP-value. This indicates that for the conditions studied here, there were no situations in which deviance was sensitive to misfit but SGDDM was not. Given the superior performance of the global SGDDM fit function, there seems little reason for including deviance as an assessor of global fit for the types of

violations simulated here when SGDDM is available. For this reason, further details of the deviance results have been omitted. ICI was the only investigated observable-level fit function to display any power for detecting the types of misfit modeled in this study. Figure 11 was suggestive of ICI's utility, but aggregation across observables obscures the underlying results. When viewed at the observable level (see Figure 12), the performance of ICI can be understood more clearly.

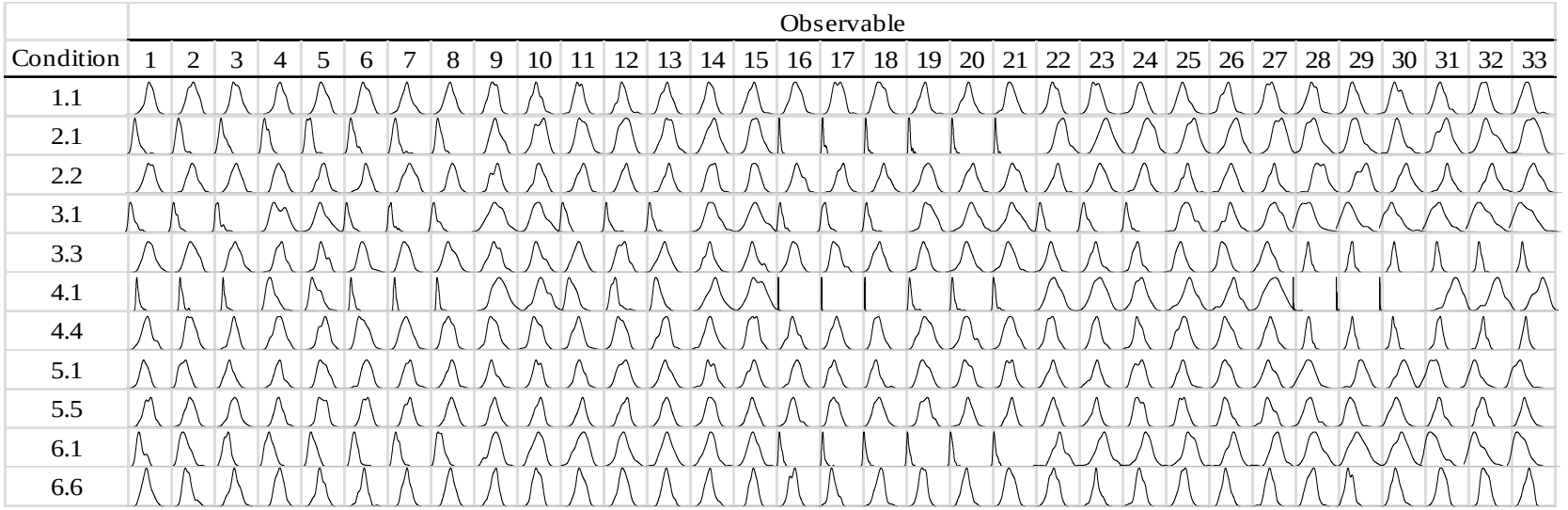



*Figure 12*. PPP-value distributions for the ICI fit function by condition and observable. Each density in the matrix represents 100 PPP-values (1 per replication).

While it is true that all the ICI PPP-value distributions in null conditions were far from uniform, there was sufficient power in some misspecified conditions to indicate that ICI could be useful as part of a PPMC toolkit. Specifically, ICI had its greatest

power in conditions where the model misspecifications included an additional latent class. However, even within these conditions with a misspecified number of latent classes (2.1, 4.1, and 6.1), the observed power of ICI was dramatically greater for some observables than others, even holding constant the CPT template. For example, the second row of Figure 12 shows the distributions of ICI PPP-values for each observable within Condition 2.1. In this row, the most extreme distributions correspond to Observables 16-21, which are the observables parented by $\theta_2$ and governed by CPT Template 8 where the partial mastery class behaved as the mastery class. This finding is complicated by the fact that the remainder of the observables governed by the same CPT template and parented by a different latent variable exhibited minimal power. The location of the PPP-value distributions for Observables 1-8 (governed by the same CPT template but parented by $\theta_1$) and Observables 28-30 (governed by the same CPT template but parented by $\theta_3$) were in the same direction as those from Observables 1-8 but were less extreme. These results are examined in greater detail later.

**Global SGDDM**

As depicted in the second panel of Figure 11, the distributions of global SGDDM PPP-values were dramatically different for null and misspecified conditions. The distributions from null conditions approached uniformity, while the distributions from misspecified conditions were located almost exclusively within the extreme lower tail. Condition 5.1 was the only misspecified condition with any non-zero PPP-values.

Three different ways to summarize the distributions of PPP-values were implemented in this study. For the global SGDDM PPP-values, each summary told much the same story. Looking down the second column of Table 26, it can be seen that there

was a large disparity between the proportions of extreme PPP-values from null and misspecified conditions. The proportions of PPP-values flagged as extreme in the null conditions (1.1, 2.2, 3.3, 4.4, 5.5, and 6.6) ranged from .00 to .10, and were all much smaller than the proportions from the misspecified conditions (2.1, 3.1, 4.1, 5.1, and 6.1), which ranged from .93 to 1.00. The median PPP-values displayed a similar pattern of disparity (see the second column of Table 27): medians in the null conditions ranged from .41 to .50, while the medians from the misspecified conditions were all zero.

Table 27

*Median PPP-value across replications by condition and fit function*

| | Fit Function | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | Deviance | SGDDM global | SGDDM $\theta_1$ | SGDDM $\theta_2$ | SGDDM $\theta_3$ | SGDDM bivariate | $Q_3$ | PC | $\chi^2$ | RPS | GLS | ICI | HCI |
| 1.1 | .48 | .41 | .52 | .47 | .53 | .50 | .50 | .49 | .51 | .51 | .50 | .45 | .49 |
| 2.1 | .43 | 0 | 0 | 0 | 0 | .47 | .47 | .49 | .50 | .50 | .49 | .27 | .49 |
| 2.2 | .47 | .54 | .52 | .56 | .52 | .50 | .50 | .49 | .51 | .51 | .50 | .46 | .50 |
| 3.1 | .42 | 0 | 0 | 0 | 0 | .50 | .50 | .49 | .50 | .50 | .50 | .20 | .48 |
| 3.3 | .50 | .49 | .47 | .52 | .53 | .50 | .50 | .51 | .51 | .51 | .52 | .41 | .50 |
| 4.1 | .39 | 0 | 0 | 0 | 0 | .47 | .47 | .49 | .49 | .49 | .49 | .12 | .49 |
| 4.4 | .48 | .48 | .44 | .48 | .52 | .50 | .50 | .50 | .51 | .51 | .51 | .42 | .50 |
| 5.1 | .46 | 0 | .19 | .22 | .02 | .49 | .49 | .49 | .50 | .50 | .50 | .42 | .50 |
| 5.5 | .47 | .54 | .60 | .49 | .55 | .50 | .50 | .50 | .51 | .51 | .51 | .42 | .50 |
| 6.1 | .43 | 0 | 0 | 0 | 0 | .49 | .49 | .49 | .50 | .50 | .49 | .27 | .50 |
| 6.6 | .47 | .52 | .49 | .47 | .50 | .50 | .50 | .49 | .51 | .51 | .50 | .44 | .50 |

This pattern continued for the median effect size outcome (see the second column of Table 28). The median effect sizes in the null conditions ranged from -.11 to .22, and were dramatically smaller than the median effect sizes from the misspecified conditions, which ranged from 3.60 to 15.06.

Table 28

*Median effect size across replications by condition and fit function*

| | Fit Function | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | Deviance | SGDDM global | SGDDM $\theta_1$ | SGDDM $\theta_2$ | SGDDM $\theta_3$ | SGDDM bivariate | $Q_3$ | PC | $\chi^2$ | RPS | GLS | ICI | HCI |
| 1.1 | .06 | .22 | -.05 | .05 | -.08 | 0 | 0 | 0 | .01 | -.01 | -.03 | .01 | .02 |
| 2.1 | .17 | 13.02 | 9.20 | 9.18 | 5.32 | .06 | .06 | 0 | .03 | .02 | 0 | .03 | .05 |
| 2.2 | .09 | -.10 | -.07 | -.16 | -.09 | .01 | .01 | 0 | 0 | -.01 | -.01 | .01 | .05 |
| 3.1 | .19 | 9.41 | 9.97 | 7.35 | 13.02 | -.01 | -.01 | 0 | .03 | .02 | -.02 | .06 | .11 |
| 3.3 | .03 | .02 | .07 | -.08 | -.07 | .01 | .01 | 0 | -.01 | -.02 | -.06 | .01 | .06 |
| 4.1 | .27 | 15.06 | 14.02 | 12.52 | 8.49 | .07 | .07 | 0 | .05 | .04 | .01 | .09 | .17 |
| 4.4 | .04 | .06 | .18 | .09 | -.06 | .01 | .01 | 0 | -.01 | -.02 | -.02 | .01 | .10 |
| 5.1 | .11 | 3.60 | .89 | .78 | 2.21 | .01 | .01 | 0 | .01 | 0 | -.02 | .01 | .05 |
| 5.5 | .08 | -.11 | -.28 | .01 | -.16 | -.01 | -.01 | 0 | 0 | -.01 | -.04 | .01 | .04 |
| 6.1 | .19 | 13.20 | 8.27 | 8.64 | 6.09 | .02 | .02 | 0 | .03 | .02 | .01 | .04 | .07 |
| 6.6 | .09 | -.04 | -.01 | .05 | .01 | 0 | 0 | 0 | 0 | -.02 | -.01 | .01 | .06 |

Taken collectively, these three columns of results suggest that the global SGDDM fit function performed well in terms of distinguishing between null and misspecified conditions. One notable difference among the three outcomes is that the proportion-

flagged and median-PPP-values outcomes displayed a kind of ceiling effect. All misspecified conditions had a median PPP-value of zero, so comparative judgments of misfit across conditions were not possible. Similarly, the proportion flagged was 1 for all misspecified conditions except Condition 5.1 (proportion = .93), indicating that the degree of misfit in Condition 5.1 was less than the other four conditions, but no further distinctions were possible. By comparison, a useful feature of the effect size outcome was that it did not have a ceiling. The scale of the effect size outcome permitted distinctions among conditions in terms of overall degree of misfit that were not apparent using the proportion flagged and median PPP-value outcomes. Specifically, the degree of misfit across replications as characterized by largest median effect size to smallest median effect size was Condition 4.1 (ES = 15.06), Conditions 6.1 (ES = 13.20), Condition 2.1 (ES = 13.02), Condition 3.1 (ES = 9.41), and Condition 5.1 (ES = 3.60). This pattern is telling because the three conditions with the greatest misfit all had the partial mastery class misspecification. Additionally, the contextual variables misspecification produced greater misfit than the cross-loadings misspecification, as evidenced by the worse fit for Condition 3.1 relative to Condition 5.1 as well as Condition 4.1 relative to Condition 6.1.

Figures 13-14 depict scatterplots of realized and posterior predicted SGDDM values across all 100 replications of each condition. The figures are paneled by condition, with null conditions and misspecified conditions grouped together to facilitate comparisons of the manipulated factors across conditions.

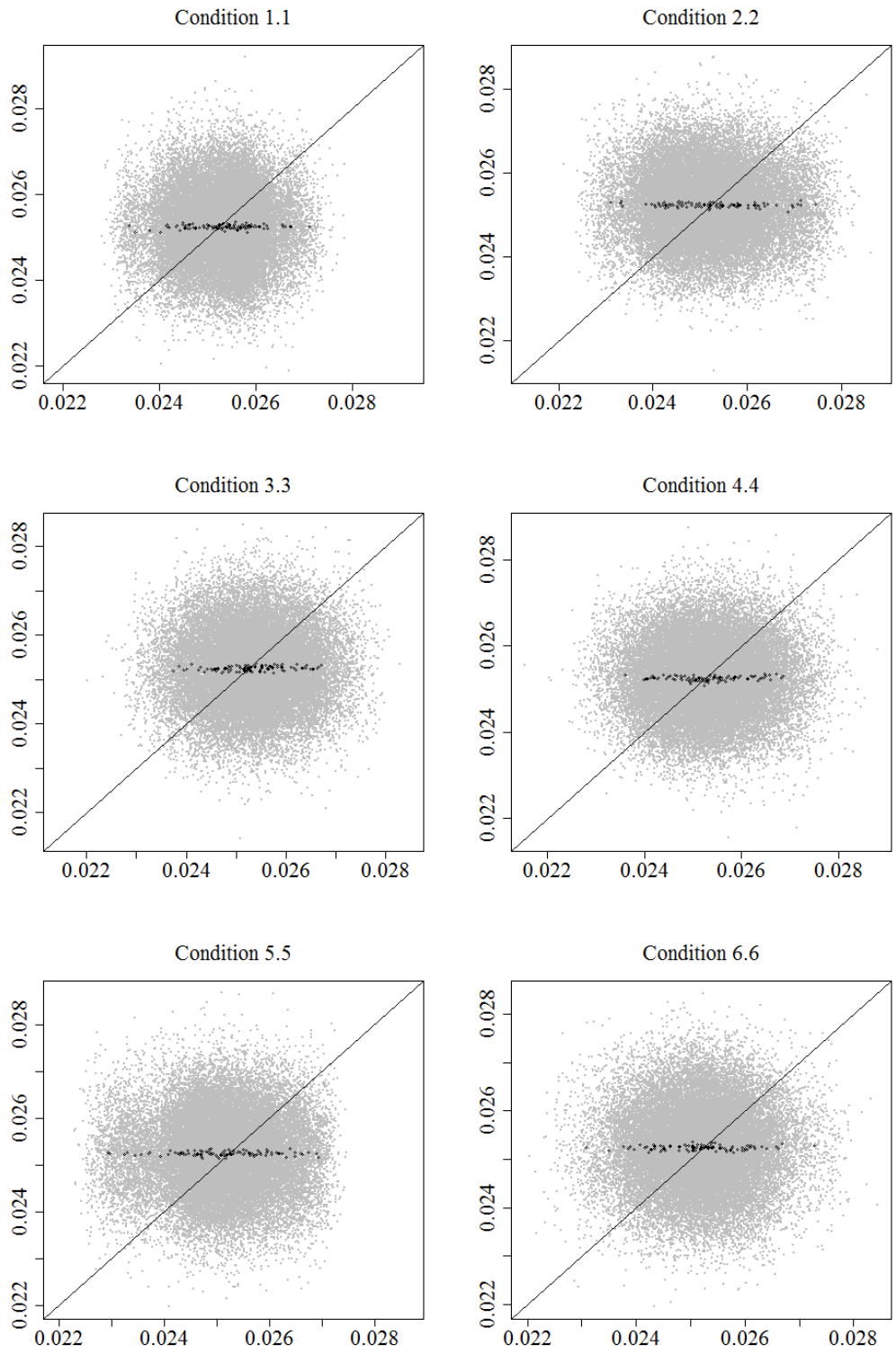



*Figure 13*. Scatterplots of SGDDM global values in null conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

The x-axis of each scatterplot contains the realized values, while posterior predicted values are plotted along the y-axis. Each gray dot in the scatterplots represents one draw from a posterior distribution (300 draws were used from each of 100 replications for a total of 30,000 draws per condition). The open black circles represent the centroids (i.e. the mean realized and posterior predicted values across the 300 draws within a given replication). Each centroid can be thought of as a summary of the 300 draws from the replication it represents. The centroids are included for graphical purposes only, to help the reader perceive patterns when viewing the graphs. PPMC analyses do not make use of the centroids. The centroids are simply auxiliary information to facilitate digestion of the results in the current context of a simulation study. Figure 13 suggests that global SGDDM performed similarly across the six null conditions, although some slight differences are perceptible. The shape of the scatter in each panel is best described as roughly spherical. Generally speaking, the SGDDM values ranged from .022 to .028 for realized and posterior predicted data (Condition 1.1 can be seen to have a slightly narrower range of values). By comparison, the misspecified conditions featured in Figure 14 displayed patterns that were different from their null counterparts and from each other.

In each misspecified condition, posterior predicted values had less variability than did realized values, and were generally smaller in magnitude. These tendencies were less severe for Condition 5.1 than for the other misspecified conditions. Generally speaking, these patterns held for the subscale aggregations of SGDDM as well (see Figures 15-20), though contrasts between conditions (e.g. between Condition 5.1 and the other misspecified conditions) were more striking for some of the subscales than for others.

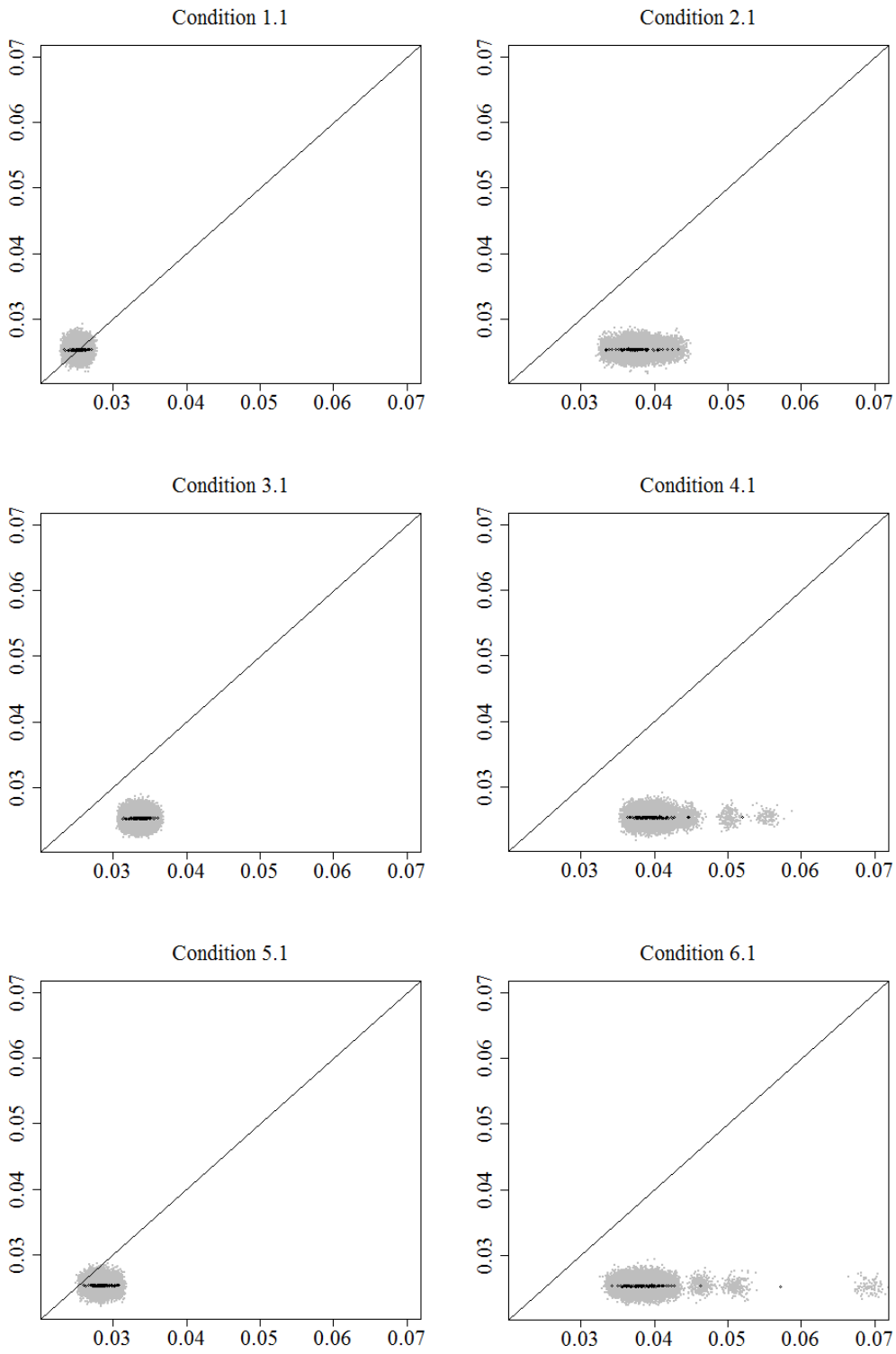



*Figure 14.* Scatterplots of SGDDM global values in misspecified conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

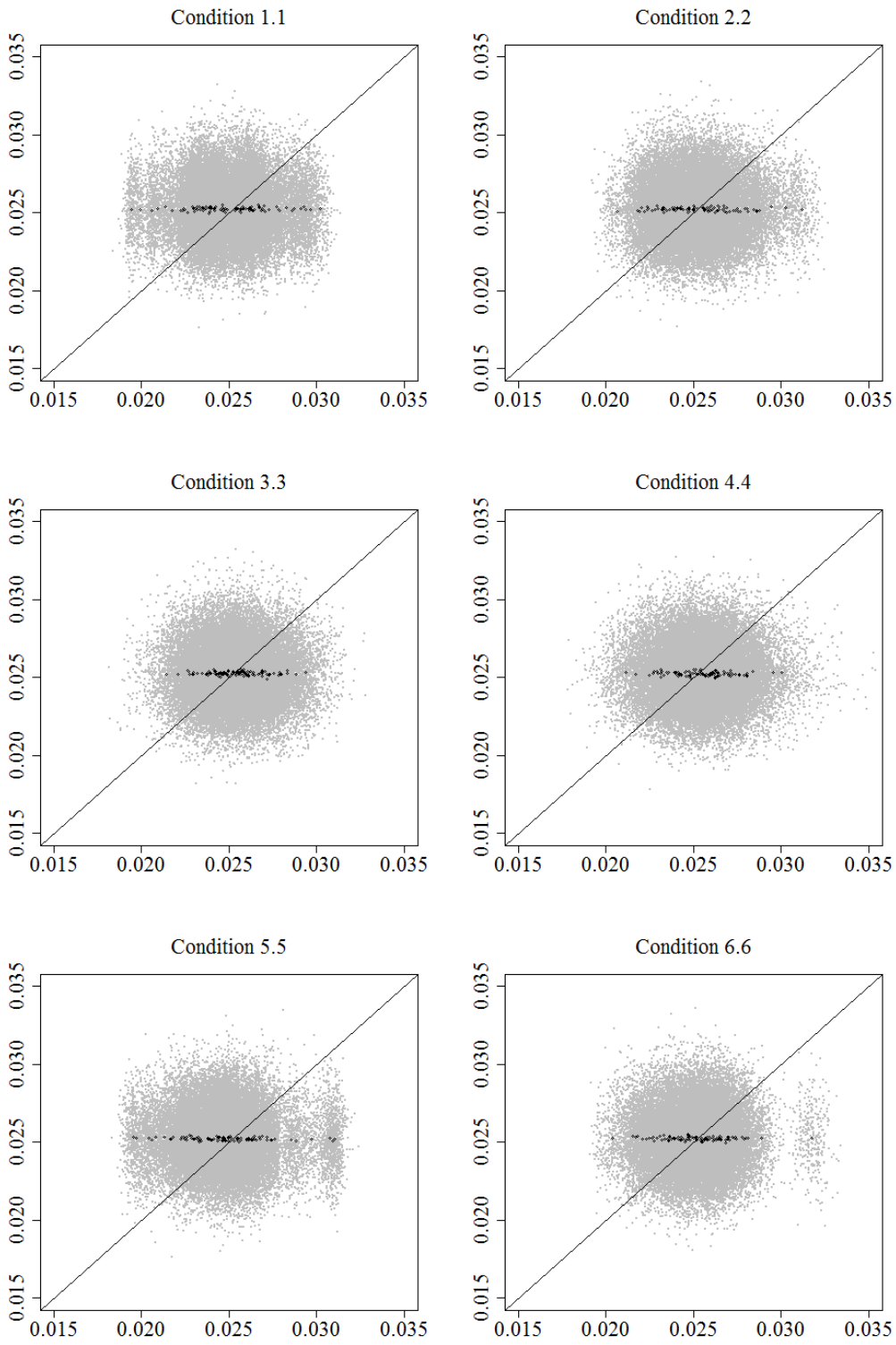



*Figure 15*. Scatterplots of SGDDM subscale $\theta_1$ values in null conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

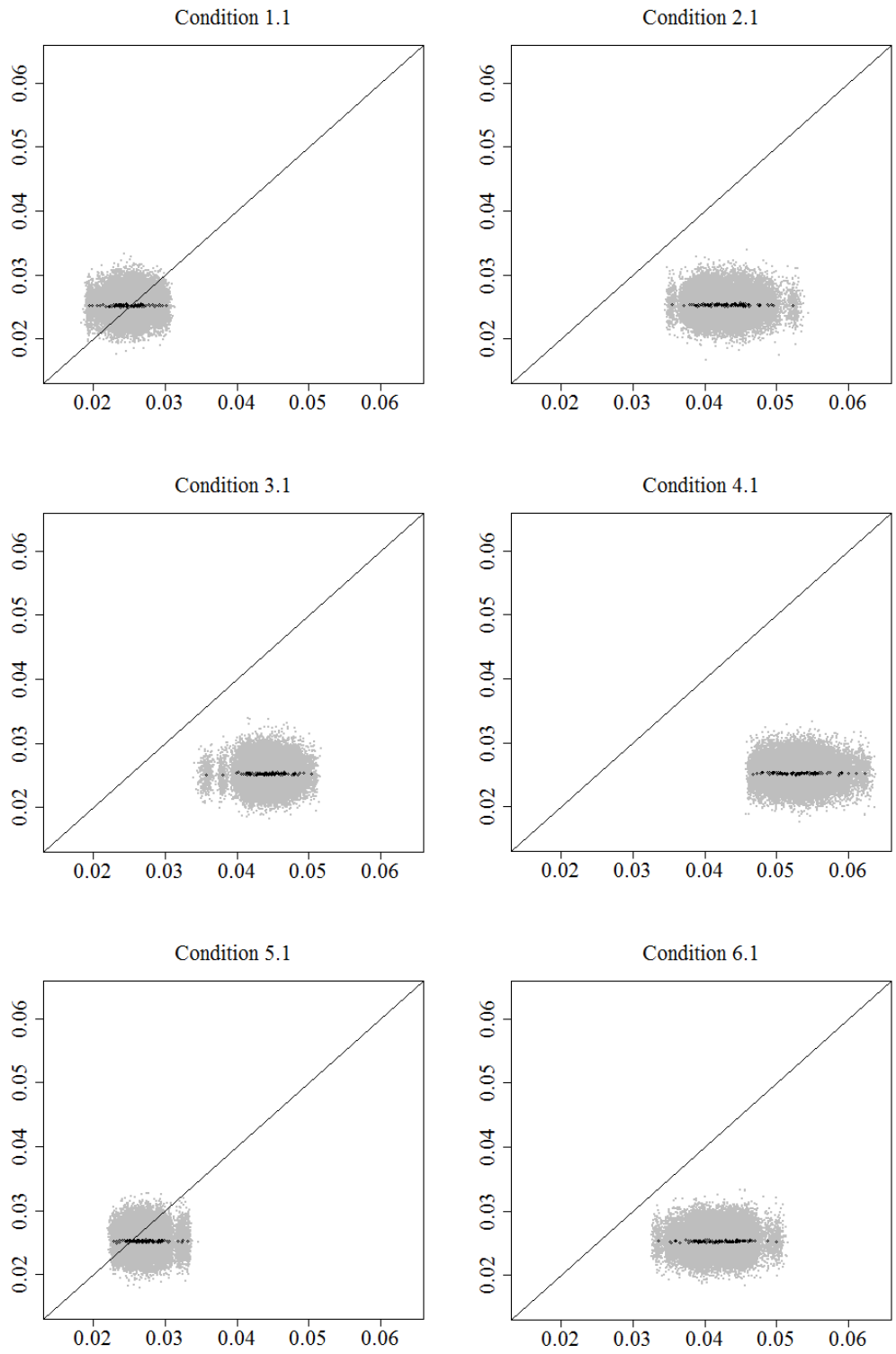



*Figure 16.* Scatterplots of SGDDM subscale $\theta_1$ values in misspecified conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

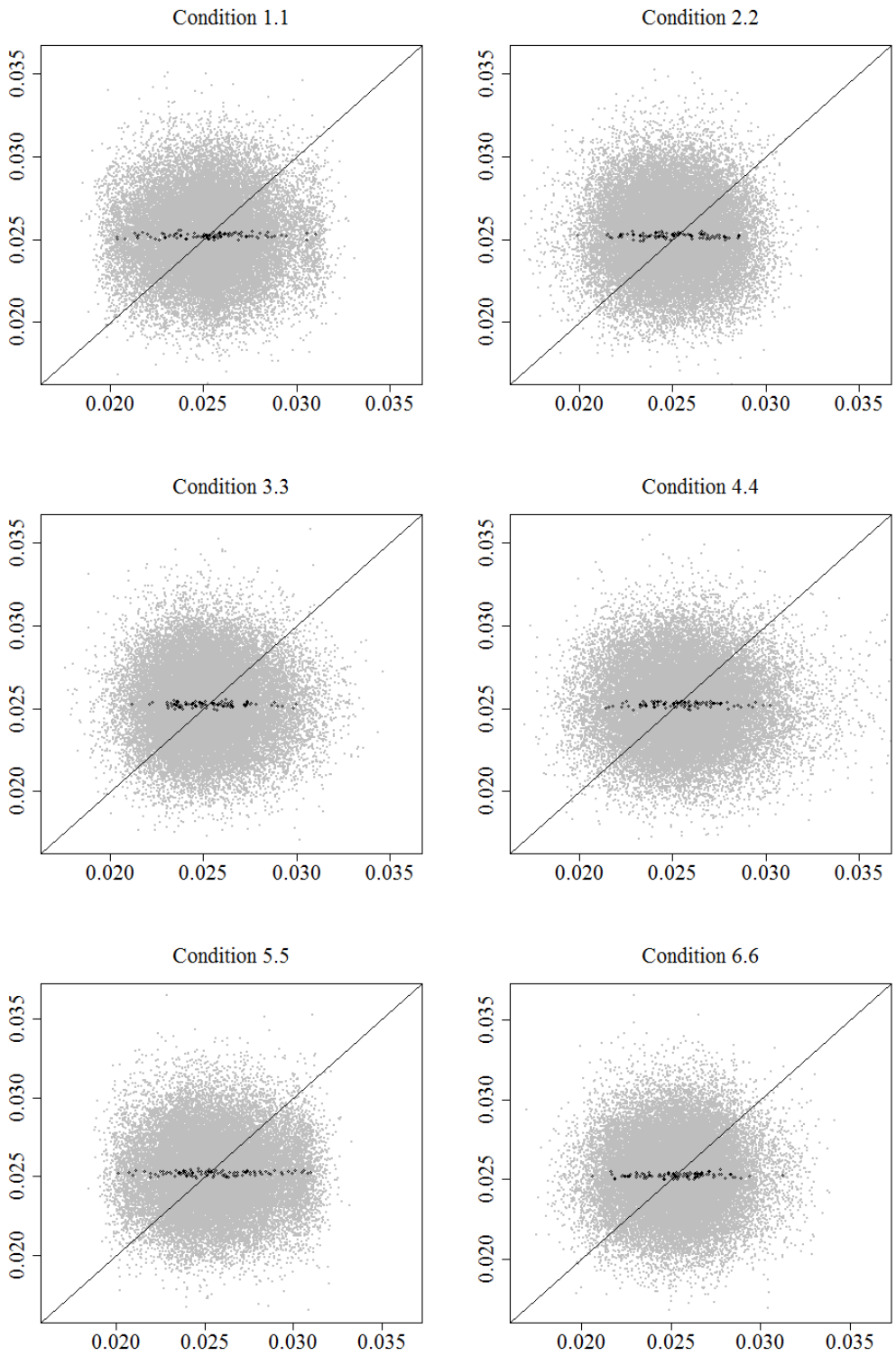



*Figure 17.* Scatterplots of SGDDM subscale $\theta_2$ values in null conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

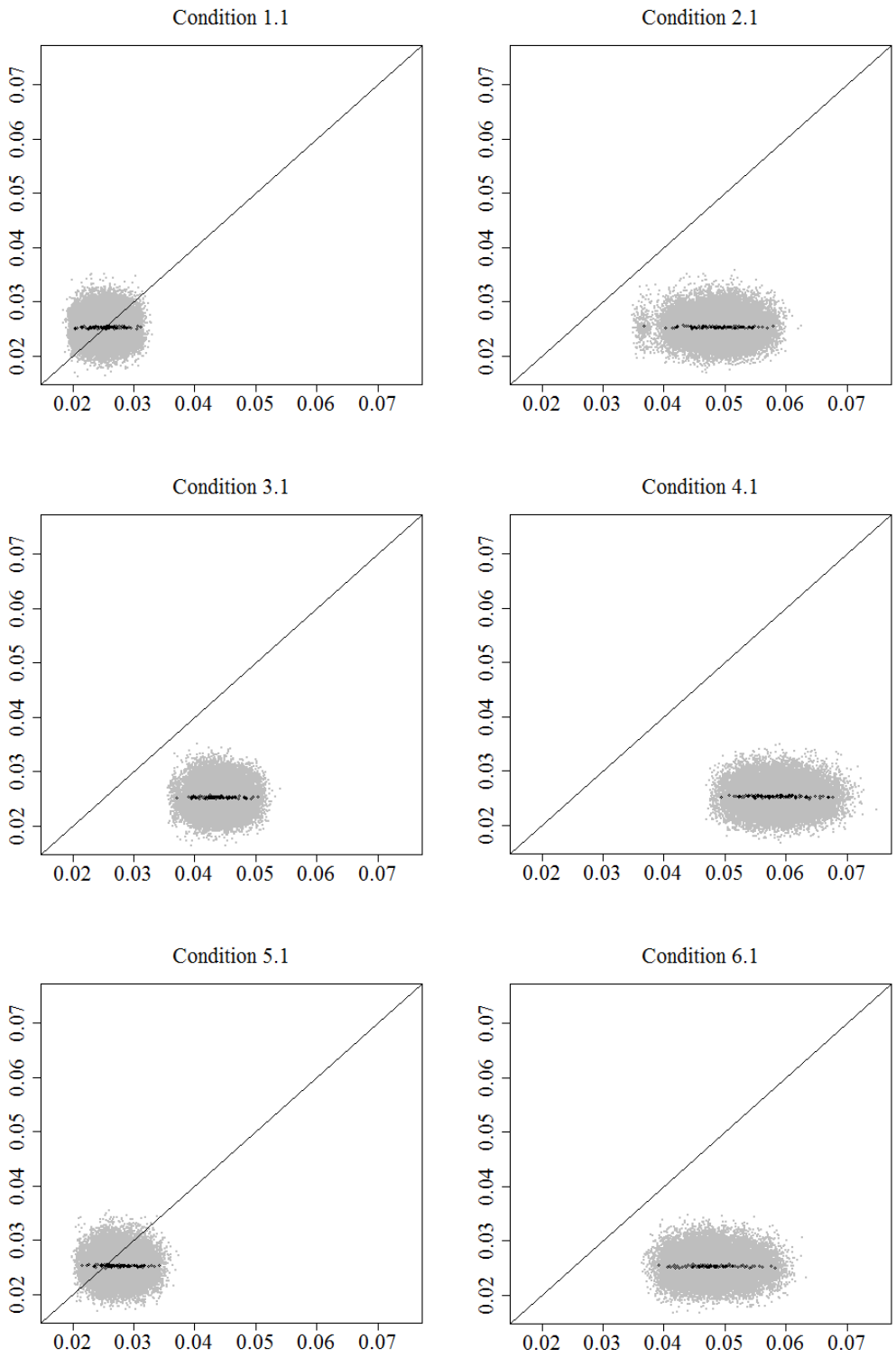



*Figure 18*. Scatterplots of SGDDM subscale $\theta_2$ values in misspecified conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

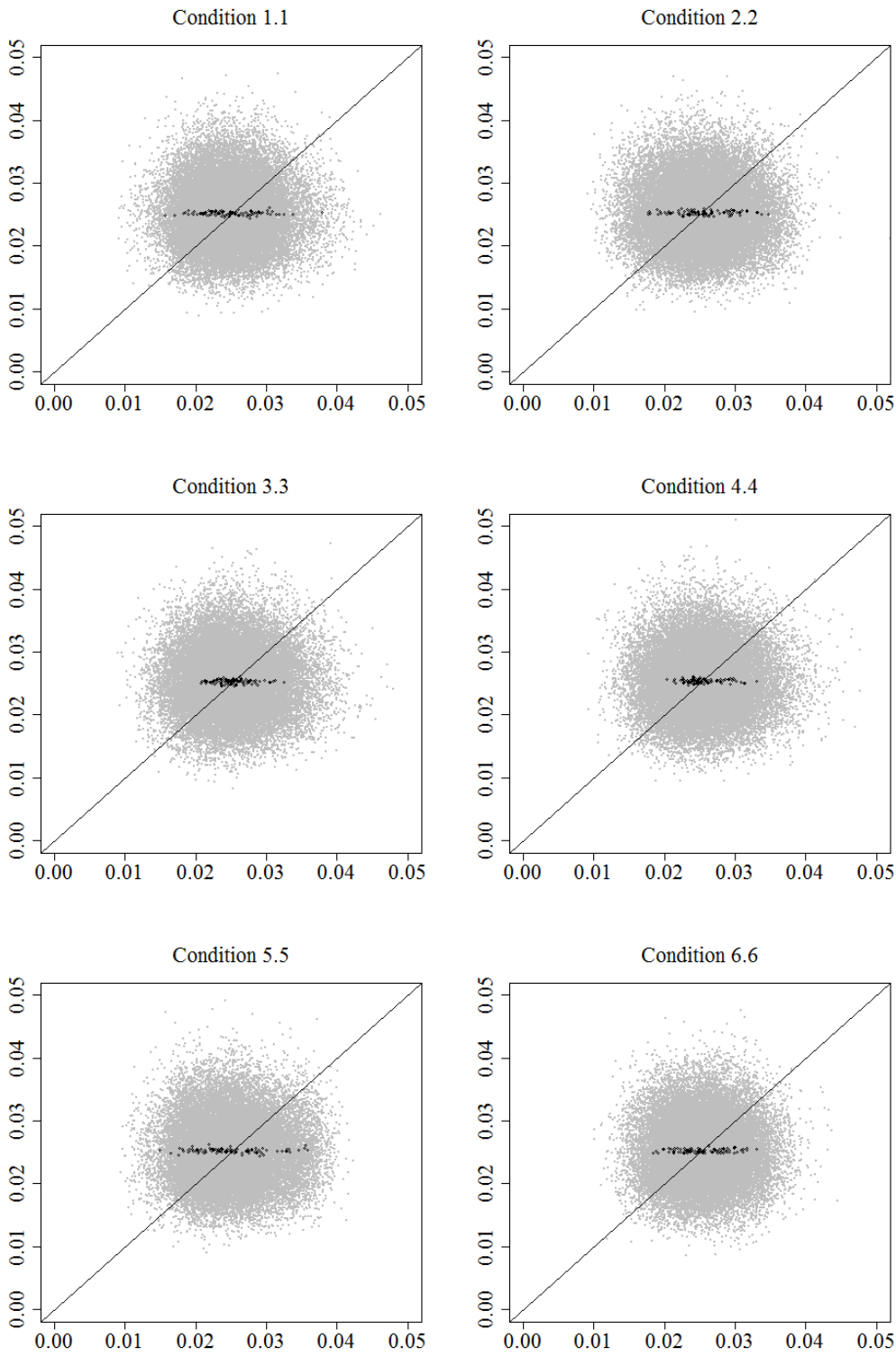



*Figure 19.* Scatterplots of SGDDM subscale $\theta_3$ values in null conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

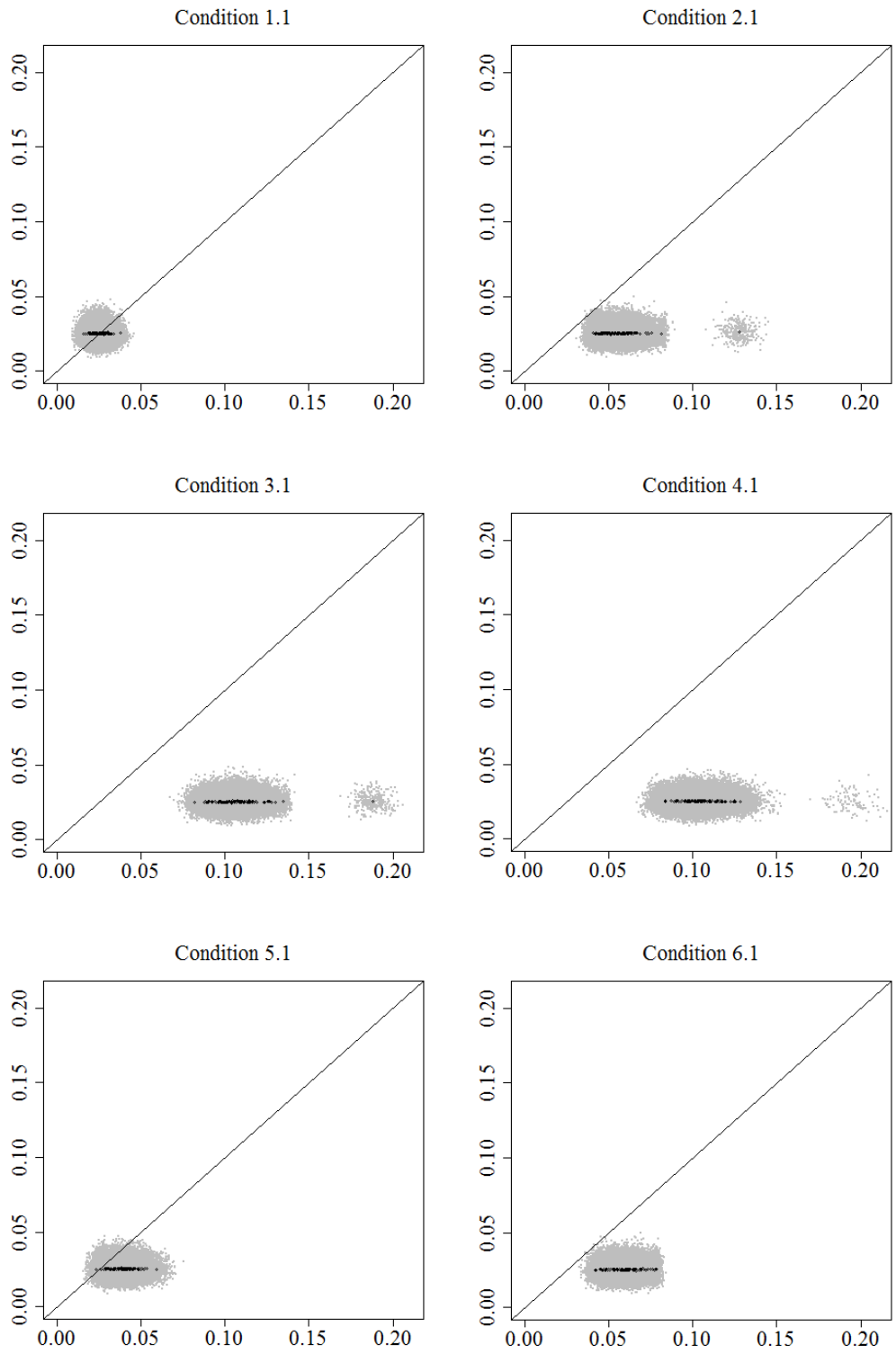



*Figure 20.* Scatterplots of SGDDM subscale $\theta_3$ values in misspecified conditions. Posterior predicted values are on the y-axis and realized values are on the x-axis.

**SGDDM Subscale $\theta_1$**

As depicted in the third panel of Figure 11, the distributions of PPP-values for SGDDM subscale $\theta_1$ tended to be different for null and misspecified conditions. The distributions from null conditions approached uniformity, while the distributions from misspecified conditions were located almost exclusively within the extreme lower tail except for Condition 5.1. Condition 5.1 was the only misspecified condition with any non-zero PPP-values. Figure 16 makes it easy to see that the misfit in Condition 5.1 was to a smaller degree than the other misspecified conditions, although it tended to be in the same direction in the aggregate. Condition 5.1 was the only misspecified condition in which any of the sampled MCMC iterations had posterior predicted values exceeding realized values (i.e. grey dots above the identity line); indeed Condition 5.1 even had some replications with PPP-values above .5, indicating that the majority of iterations within those replications exhibited over-prediction.

The proportions of PPP-values flagged as extreme (third column of Table 26) in the null conditions ranged from .02 to .12. The proportions from the misspecified conditions were 1.00 except for Condition 5.1, which was .18. The median PPP-values (third column of Table 27) displayed a similar pattern: medians in the null conditions ranged from .44 to .60, while the medians from the misspecified conditions were .00 except for Condition 5.1 which was .19. This pattern continued for the median effect size outcome (see the third column of Table 28). The median effect sizes in the null conditions ranged from -.28 to .18, and were smaller than the median effect sizes from the misspecified conditions, which ranged from .89 to 14.02.

**SGDDM Subscale $\theta_2$**

As depicted in the fourth panel of Figure 11, the distributions of PPP-values for SGDDM subscale $\theta_2$ were quite different for null and misspecified conditions. The distributions from null conditions approached uniformity, while the distributions from misspecified conditions tended toward the extreme lower tail. Condition 5.1 was the only misspecified condition with any non-zero PPP-values. Figure 18 shows that the misfit in Condition 5.1 was not as severe. Condition 5.1 was the only misspecified condition in which any of the sampled MCMC iterations had posterior predicted values exceeding realized values (i.e. grey dots above the identity line). Condition 5.1 had some replications with PPP-values above .5, indicating that the majority of iterations within those replications exhibited this characteristic.

The proportions of PPP-values flagged as extreme (fourth column of Table 26) in the null conditions ranged from .00 to .07, while proportions from the misspecified conditions were 1.00 except for Condition 5.1 which was .19. The median PPP-values (fourth column of Table 27) displayed a similar pattern: medians in the null conditions ranged from .47 to .56, while the medians from the misspecified conditions were .00 except for Condition 5.1 which was .22. This pattern of results continued for the median effect size outcome (fourth column of Table 28). The median effect sizes in the null conditions ranged from -.16 to .09, and were smaller than the median effect sizes from the misspecified conditions, which ranged from .78 to 12.52.

**SGDDM Subscale $\theta_3$**

As depicted in the fifth panel of Figure 11, the distributions of PPP-values for SGDDM subscale $\theta_3$ were different for null and misspecified conditions. The

distributions from null conditions approached uniformity, while the distributions from misspecified conditions were located almost exclusively within the extreme lower tail. Condition 5.1 was the only misspecified condition with any non-zero PPP-values. Figure 20 shows that Condition 5.1 was the only misspecified condition in which any of the sampled MCMC iterations had posterior predicted values exceeding realized values (i.e. grey dots above the identity line).

The proportions of PPP-values flagged as extreme (see Table 26) in the null conditions ranged from .00 to .02, while the proportions from misspecified conditions were 1.00 except for Condition 5.1, which was .58. The median PPP-values (see the fifth column of Table 27) displayed a similar pattern: medians in the null conditions ranged from .50 to .55, while the medians from the misspecified conditions were .00 except for Condition 5.1 which was .02. The pattern of results continued for the median effect size outcome (see the fifth column of Table 28). The median effect sizes in the null conditions ranged from -.16 to .01, and were smaller than the median effect sizes from the misspecified conditions, which ranged from 2.21 to 13.02.

To better understand why subscale SGDDM detected misfit more often for $\theta_3$ than for $\theta_2$ or $\theta_1$, consider Figure 21, which shows the conditional probability of a correct response by observable for each latent proficiency profile. The upper panel refers to the realized data and the lower panel refers to the posterior predicted data. The middle panel shows Model 1 generating parameters for reference purposes, i.e. to help illustrate which conditional probabilities were affected by the model misspecification in Condition 5.1.

Conditional probability of a correct response (Model 5 generating parameters)

| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .08 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 |
| [1,2,1] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .80 | .80 | .80 | .80 |
| [2,1,1] | .03 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .06 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .80 | .80 | .80 |
| [2,2,1] | .08 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |

Conditional probability of a correct response Model 1 generating parameters)

| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .08 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [1,2,1] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,1,1] | .03 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .06 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,2,1] | .08 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |

Conditional probability of a correct response (Condition 5.1 estimated parameters)

| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .09 | .20 | .19 | .21 | .22 | .22 | .21 | .21 | .21 | .21 | .21 | .22 | .21 | .22 | .21 | .22 | .20 | .20 | .20 | .22 | .21 | .22 | .21 | .21 | .21 | .21 | .21 | .21 | .26 | .19 | .19 | .29 | .30 | .30 |
| [1,1,2] | .00 | .20 | .19 | .21 | .22 | .22 | .21 | .21 | .21 | .21 | .21 | .22 | .21 | .22 | .21 | .22 | .20 | .20 | .20 | .22 | .21 | .22 | .21 | .21 | .21 | .21 | .21 | .21 | .80 | .78 | .78 | .80 | .80 | .80 |
| [1,2,1] | .01 | .20 | .19 | .21 | .22 | .22 | .21 | .21 | .21 | .21 | .21 | .22 | .21 | .22 | .21 | .22 | .80 | .74 | .74 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .26 | .19 | .19 | .29 | .30 | .30 |
| [1,2,2] | .00 | .20 | .19 | .21 | .22 | .22 | .21 | .21 | .21 | .21 | .21 | .22 | .21 | .22 | .21 | .22 | .80 | .74 | .74 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .78 | .78 | .80 | .80 | .80 |
| [2,1,1] | .06 | .74 | .69 | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .22 | .21 | .22 | .21 | .21 | .21 | .21 | .21 | .21 | .26 | .19 | .19 | .29 | .30 | .30 |
| [2,1,2] | .03 | .74 | .69 | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .22 | .21 | .22 | .21 | .21 | .21 | .21 | .21 | .21 | .80 | .78 | .78 | .80 | .80 | .80 |
| [2,2,1] | .08 | .74 | .69 | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .74 | .74 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .26 | .19 | .19 | .29 | .30 | .30 |
| [2,2,2] | .72 | .74 | .69 | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .74 | .74 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .78 | .78 | .80 | .80 | .80 |

*Figure 21*. Conditional probability of a correct response by latent proficiency. Upper panel shows values as generated from Model 5. Middle panel shows values as generated from Model 1. Lower panel shows the mean of 30,000 estimated values (300 posterior draws from each of 100 replications) from Condition 5.1.

The key point in the upper panel is that the three latent variables parented a different number of observables, yet each latent variable was impacted by the same number of crossloadings in the misspecification. These crossloadings translated to seven conditional probabilities being affected within each latent variable. For $\theta_3$, the seven impacted conditional probabilities represent a larger proportion of the corresponding response patterns, meaning that the correlations (SGDDM values) among the relevant response patterns would be affected more by the misspecification for $\theta_3$ than for $\theta_2$ or $\theta_1$. This interpretation is further evidenced in the lower panel by noting the consequences on the estimated conditional probabilities. The discrepancies between the upper and lower panel are proportionally more prevalent in the third block of observables relative to the first two blocks. In other words, one can see from the patterns in the figure why the correlations within the third block of variables changed more in the posterior predicted data relative to the realized data.

Global and subscale SGDDM discrepancy measures were less suited to detect the crossloadings misspecification in Condition 5.1 relative to the other misspecified conditions because this misspecification produced more localized effects. The global and subscale aggregations associated with this misspecification included larger portions of the data that were unaffected by the misspecification than did the other misspecifications. In other words, the summary statistics were diluted to a greater extent by well-fitting data due to the aggregation process. Looking again at Figures 3 and 7, only nine observables had conditional probability tables that differed between the generating and scoring models in Condition 5.1, meaning that for 24 of the 33 observables sampling variability was the only factor responsible for differences between observed and model-implied

responses. For the nine observables that did have different CPTs between Model 1 and Model 5, only the latent profiles with mastery on the first latent parent and non-mastery on the second latent parent were impacted by the differences in these conditional probabilities (see Figure 21), leaving a large proportion of simulees with expected probabilities of success that were equal across all observables in both models. Compared to the other misspecifications, Condition 5.1 appeared to be the most localized model misspecification in the sense that the matrix of expected response probabilities across all simulees and observables appeared least disturbed relative to the same matrix from the generating model (a comparison of Figure 21 to Figures 22-25 is suggestive of this principle). This interpretation is further supported by bivariate evidence presented in the next section. At the bivariate level greater detail was afforded by the aggregation across 528 pairs of observables as opposed to a single model, three subscales, or 33 observables, which better isolated areas of fit and misfit.

| Conditional probability of a correct response (Model 2 generating parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .06 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [1,1,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [1,2,1] | .02 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [1,2,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [1,3,1] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,3,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 |
| [1,3,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,1,1] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .01 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [2,1,3] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,2,1] | .01 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .05 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [2,2,3] | .01 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,3,1] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,3,2] | .03 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 |
| [2,3,3] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |
| [3,1,1] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,1,2] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [3,1,3] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [3,2,1] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,2,2] | .05 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [3,2,3] | .19 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [3,3,1] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,3,2] | .06 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 |
| [3,3,3] | .51 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |

| Conditional probability of a correct response (Condition 2.1 estimated parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .15 | .45 | .45 | .44 | .44 | .45 | .44 | .45 | .45 | .19 | .19 | .19 | .19 | .19 | .19 | .19 | .67 | .67 | .67 | .67 | .67 | .67 | .19 | .19 | .19 | .19 | .20 | .20 | .51 | .52 | .51 | .19 | .19 | .18 |
| [1,1,2] | .01 | .45 | .45 | .44 | .44 | .45 | .44 | .45 | .45 | .19 | .19 | .19 | .19 | .19 | .19 | .19 | .67 | .67 | .67 | .67 | .67 | .67 | .19 | .19 | .19 | .19 | .20 | .20 | .81 | .81 | .81 | .76 | .75 | .75 |
| [1,2,1] | .01 | .45 | .45 | .44 | .44 | .45 | .44 | .45 | .45 | .19 | .19 | .19 | .19 | .19 | .19 | .19 | .80 | .80 | .81 | .80 | .81 | .81 | .79 | .79 | .79 | .79 | .79 | .79 | .51 | .52 | .51 | .19 | .19 | .18 |
| [1,2,2] | .00 | .45 | .45 | .44 | .44 | .45 | .44 | .45 | .45 | .19 | .19 | .19 | .19 | .19 | .19 | .19 | .80 | .80 | .81 | .80 | .81 | .81 | .79 | .79 | .79 | .79 | .79 | .79 | .81 | .81 | .81 | .76 | .75 | .75 |
| [2,1,1] | .03 | .80 | .80 | .80 | .80 | .80 | .80 | .81 | .80 | .79 | .79 | .78 | .79 | .79 | .79 | .79 | .67 | .67 | .67 | .67 | .67 | .67 | .19 | .19 | .19 | .19 | .20 | .20 | .51 | .52 | .51 | .19 | .19 | .18 |
| [2,1,2] | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .81 | .80 | .79 | .79 | .78 | .79 | .79 | .79 | .79 | .67 | .67 | .67 | .67 | .67 | .67 | .19 | .19 | .19 | .19 | .20 | .20 | .81 | .81 | .81 | .76 | .75 | .75 |
| [2,2,1] | .04 | .80 | .80 | .80 | .80 | .80 | .80 | .81 | .80 | .79 | .79 | .78 | .79 | .79 | .79 | .79 | .80 | .80 | .81 | .80 | .81 | .81 | .79 | .79 | .79 | .79 | .79 | .79 | .51 | .52 | .51 | .19 | .19 | .18 |
| [2,2,2] | .55 | .80 | .80 | .80 | .80 | .80 | .80 | .81 | .80 | .79 | .79 | .78 | .79 | .79 | .79 | .79 | .80 | .80 | .81 | .80 | .81 | .81 | .79 | .79 | .79 | .79 | .79 | .79 | .81 | .81 | .81 | .76 | .75 | .75 |

*Figure 22*. Conditional probability of a correct response by latent proficiency. Upper panel shows the values as generated from Model 2. Lower panel shows the mean of 30,000 estimated values from Condition 2.1.

| Conditional probability of a correct response (Model 3 generating parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .08 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [1,2,1] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .68 | .68 | .68 |
| [2,1,1] | .03 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .06 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [2,2,1] | .08 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .73 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .68 | .68 | .68 |

| Conditional probability of a correct response (Model 1 generating parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .08 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [1,2,1] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,1,1] | .03 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .06 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,2,1] | .08 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .73 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |

| Conditional probability of a correct response (Condition 3.1 estimated parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .08 | .21 | .20 | .20 | .23 | .22 | .21 | .20 | .21 | .22 | .22 | .21 | .21 | .21 | .22 | .22 | .21 | .20 | .19 | .23 | .23 | .22 | .20 | .20 | .21 | .23 | .22 | .22 | .27 | .26 | .26 | .25 | .25 | .25 |
| [1,1,2] | .01 | .21 | .20 | .20 | .23 | .22 | .21 | .20 | .21 | .22 | .22 | .21 | .21 | .21 | .22 | .22 | .21 | .20 | .19 | .23 | .23 | .22 | .20 | .20 | .21 | .23 | .22 | .22 | .73 | .73 | .73 | .74 | .73 | .73 |
| [1,2,1] | .01 | .21 | .20 | .20 | .23 | .22 | .21 | .20 | .21 | .22 | .22 | .21 | .21 | .21 | .22 | .22 | .69 | .68 | .68 | .80 | .80 | .80 | .69 | .68 | .69 | .80 | .80 | .80 | .27 | .26 | .26 | .25 | .25 | .25 |
| [1,2,2] | .00 | .21 | .20 | .20 | .23 | .22 | .21 | .20 | .21 | .22 | .22 | .21 | .21 | .21 | .22 | .22 | .69 | .68 | .68 | .80 | .80 | .80 | .69 | .68 | .69 | .80 | .80 | .80 | .73 | .73 | .73 | .74 | .73 | .73 |
| [2,1,1] | .04 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .21 | .20 | .19 | .23 | .23 | .22 | .20 | .20 | .21 | .23 | .22 | .22 | .27 | .26 | .26 | .25 | .25 | .25 |
| [2,1,2] | .06 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .21 | .20 | .19 | .23 | .23 | .22 | .20 | .20 | .21 | .23 | .22 | .22 | .73 | .73 | .73 | .74 | .73 | .73 |
| [2,2,1] | .18 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .69 | .68 | .68 | .80 | .80 | .80 | .69 | .68 | .69 | .80 | .80 | .80 | .27 | .26 | .26 | .25 | .25 | .25 |
| [2,2,2] | .61 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .69 | .68 | .68 | .80 | .80 | .80 | .69 | .68 | .69 | .80 | .80 | .80 | .73 | .73 | .73 | .74 | .73 | .73 |

*Figure 23*. Conditional probability of a correct response by latent proficiency. Upper panel shows the values as generated from Model 3 (marginalized over contextual latent proficiencies). Middle panel shows the values as generated from Model 1. Lower panel shows the mean of 30,000 estimated values (300 posterior draws from each of 100 replications) from Condition 3.1.

| Conditional probability of a correct response (Model 4 generating parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .06 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .20 | .20 | .20 |
| [1,1,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [1,2,1] | .02 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .20 | .20 | .20 |
| [1,2,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [1,3,1] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,3,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 |
| [1,3,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .68 | .68 | .68 |
| [2,1,1] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .01 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .20 | .20 | .20 |
| [2,1,3] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [2,2,1] | .01 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .05 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .20 | .20 | .20 |
| [2,2,3] | .01 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [2,3,1] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,3,2] | .03 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 |
| [2,3,3] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .68 | .68 | .68 |
| [3,1,1] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,1,2] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .20 | .20 | .20 |
| [3,1,3] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [3,2,1] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,2,2] | .05 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .20 | .20 | .20 |
| [3,2,3] | .19 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .68 | .68 | .68 | .68 | .68 | .68 |
| [3,3,1] | .00 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,3,2] | .06 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .20 | .20 | .20 |
| [3,3,3] | .51 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .80 | .80 | .80 | .68 | .68 | .68 | .68 | .68 | .68 |

| Conditional probability of a correct response (Condition 4.1 estimated parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .16 | .38 | .39 | .39 | .45 | .45 | .38 | .38 | .38 | .21 | .21 | .19 | .19 | .19 | .21 | .21 | .57 | .57 | .57 | .68 | .68 | .68 | .19 | .19 | .20 | .25 | .24 | .24 | .51 | .51 | .51 | .19 | .19 | .19 |
| [1,1,2] | .01 | .38 | .39 | .39 | .45 | .45 | .38 | .38 | .38 | .21 | .21 | .19 | .19 | .19 | .21 | .21 | .57 | .57 | .57 | .68 | .68 | .68 | .19 | .19 | .20 | .25 | .24 | .24 | .71 | .71 | .71 | .76 | .76 | .76 |
| [1,2,1] | .02 | .38 | .39 | .39 | .45 | .45 | .38 | .38 | .38 | .21 | .21 | .19 | .19 | .19 | .21 | .21 | .69 | .69 | .69 | .81 | .81 | .81 | .70 | .70 | .70 | .79 | .79 | .79 | .51 | .51 | .51 | .19 | .19 | .19 |
| [1,2,2] | .01 | .38 | .39 | .39 | .45 | .45 | .38 | .38 | .38 | .21 | .21 | .19 | .19 | .19 | .21 | .21 | .69 | .69 | .69 | .81 | .81 | .81 | .70 | .70 | .70 | .79 | .79 | .79 | .71 | .71 | .71 | .76 | .76 | .76 |
| [2,1,1] | .08 | .69 | .69 | .68 | .80 | .80 | .69 | .69 | .69 | .79 | .79 | .67 | .67 | .67 | .79 | .79 | .57 | .57 | .57 | .68 | .68 | .68 | .19 | .19 | .20 | .25 | .24 | .24 | .51 | .51 | .51 | .19 | .19 | .19 |
| [2,1,2] | .19 | .69 | .69 | .68 | .80 | .80 | .69 | .69 | .69 | .79 | .79 | .67 | .67 | .67 | .79 | .79 | .57 | .57 | .57 | .68 | .68 | .68 | .19 | .19 | .20 | .25 | .24 | .24 | .71 | .71 | .71 | .76 | .76 | .76 |
| [2,2,1] | .13 | .69 | .69 | .68 | .80 | .80 | .69 | .69 | .69 | .79 | .79 | .67 | .67 | .67 | .79 | .79 | .69 | .69 | .69 | .81 | .81 | .81 | .70 | .70 | .70 | .79 | .79 | .79 | .51 | .51 | .51 | .19 | .19 | .19 |
| [2,2,2] | .41 | .69 | .69 | .68 | .80 | .80 | .69 | .69 | .69 | .79 | .79 | .67 | .67 | .67 | .79 | .79 | .69 | .69 | .69 | .81 | .81 | .81 | .70 | .70 | .70 | .79 | .79 | .79 | .71 | .71 | .71 | .76 | .76 | .76 |

*Figure 24*. Conditional probability of a correct response by latent proficiency. Upper panel shows the values as generated from Model 4 (marginalized over contextual latent proficiencies). Lower panel shows the mean of 30,000 estimated values (300 posterior draws from each of 100 replications) from Condition 4.1.

| Conditional probability of a correct response (Model 6 generating parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .06 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,1,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 |
| [1,2,1] | .02 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,2,2] | .01 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .20 |
| [1,2,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 |
| [1,3,1] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [1,3,2] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .80 | .20 | .20 | .20 |
| [1,3,3] | .00 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .80 | .80 | .80 | .80 |
| [2,1,1] | .00 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,1,2] | .01 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .20 | .20 | .20 |
| [2,1,3] | .00 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .80 | .80 | .80 |
| [2,2,1] | .01 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,2,2] | .05 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [2,2,3] | .01 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [2,3,1] | .00 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [2,3,2] | .03 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 |
| [2,3,3] | .00 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |
| [3,1,1] | .00 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,1,2] | .00 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .20 | .20 | .20 |
| [3,1,3] | .00 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .20 | .20 | .80 | .80 | .80 |
| [3,2,1] | .00 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,2,2] | .05 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .20 | .20 | .20 |
| [3,2,3] | .19 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 |
| [3,3,1] | .00 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 | .20 | .20 | .20 |
| [3,3,2] | .06 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .20 | .20 | .20 |
| [3,3,3] | .51 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 |

| Conditional probability of a correct response (Condition 6.1 estimated parameters) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Proficiencies | Prop. of | Observable ($x_j$) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| $[\theta_1,\theta_2,\theta_3]$ | simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| [1,1,1] | .15 | .41 | .39 | .41 | .44 | .45 | .44 | .44 | .45 | .19 | .19 | .19 | .19 | .19 | .20 | .19 | .62 | .61 | .62 | .67 | .66 | .67 | .20 | .20 | .20 | .20 | .20 | .20 | .44 | .42 | .43 | .18 | .18 | .18 |
| [1,1,2] | .01 | .41 | .39 | .41 | .44 | .45 | .44 | .44 | .45 | .19 | .19 | .19 | .19 | .19 | .20 | .19 | .62 | .61 | .62 | .67 | .66 | .67 | .20 | .20 | .20 | .20 | .20 | .20 | .81 | .81 | .81 | .74 | .74 | .74 |
| [1,2,1] | .01 | .41 | .39 | .41 | .44 | .45 | .44 | .44 | .45 | .19 | .19 | .19 | .19 | .19 | .20 | .19 | .80 | .81 | .80 | .80 | .80 | .80 | .78 | .78 | .77 | .78 | .78 | .78 | .44 | .42 | .43 | .18 | .18 | .18 |
| [1,2,2] | .01 | .41 | .39 | .41 | .44 | .45 | .44 | .44 | .45 | .19 | .19 | .19 | .19 | .19 | .20 | .19 | .80 | .81 | .80 | .80 | .80 | .80 | .78 | .78 | .77 | .78 | .78 | .78 | .81 | .81 | .81 | .74 | .74 | .74 |
| [2,1,1] | .02 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .78 | .78 | .78 | .78 | .78 | .78 | .78 | .62 | .61 | .62 | .67 | .66 | .67 | .20 | .20 | .20 | .20 | .20 | .20 | .44 | .42 | .43 | .18 | .18 | .18 |
| [2,1,2] | .21 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .78 | .78 | .78 | .78 | .78 | .78 | .78 | .62 | .61 | .62 | .67 | .66 | .67 | .20 | .20 | .20 | .20 | .20 | .20 | .81 | .81 | .81 | .74 | .74 | .74 |
| [2,2,1] | .02 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .78 | .78 | .78 | .78 | .78 | .78 | .78 | .80 | .81 | .80 | .80 | .80 | .80 | .78 | .78 | .77 | .78 | .78 | .78 | .44 | .42 | .43 | .18 | .18 | .18 |
| [2,2,2] | .56 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .80 | .78 | .78 | .78 | .78 | .78 | .78 | .78 | .80 | .81 | .80 | .80 | .80 | .80 | .78 | .78 | .77 | .78 | .78 | .78 | .81 | .81 | .81 | .74 | .74 | .74 |

*Figure 25*. Conditional probability of a correct response by latent proficiency. Upper panel shows the values as generated from Model 6. Lower panel shows the mean of 30,000 estimated values from Condition 6.1.

**Bivariate SGDDM**

The detail of the feedback provided by the bivariate fit functions (SGDDM bivariate and $Q_3$) gets obscured when aggregated across variable pairs as was done in Figure 11 and Tables 26-28. Those high-level aggregations were included to facilitate comparisons across fit functions, but to really appreciate the effectiveness of the bivariate fit functions, they must be viewed at the more specific and appropriate grain-size of the variable pair. This is important because not all variable pairs were expected to show poor fit in misspecified conditions.

For all null conditions the results for the bivariate SGDDM fit function were highly similar, so they are represented collectively by Figure 26, which depicts a heat map of median PPP-values from Condition 1.1.

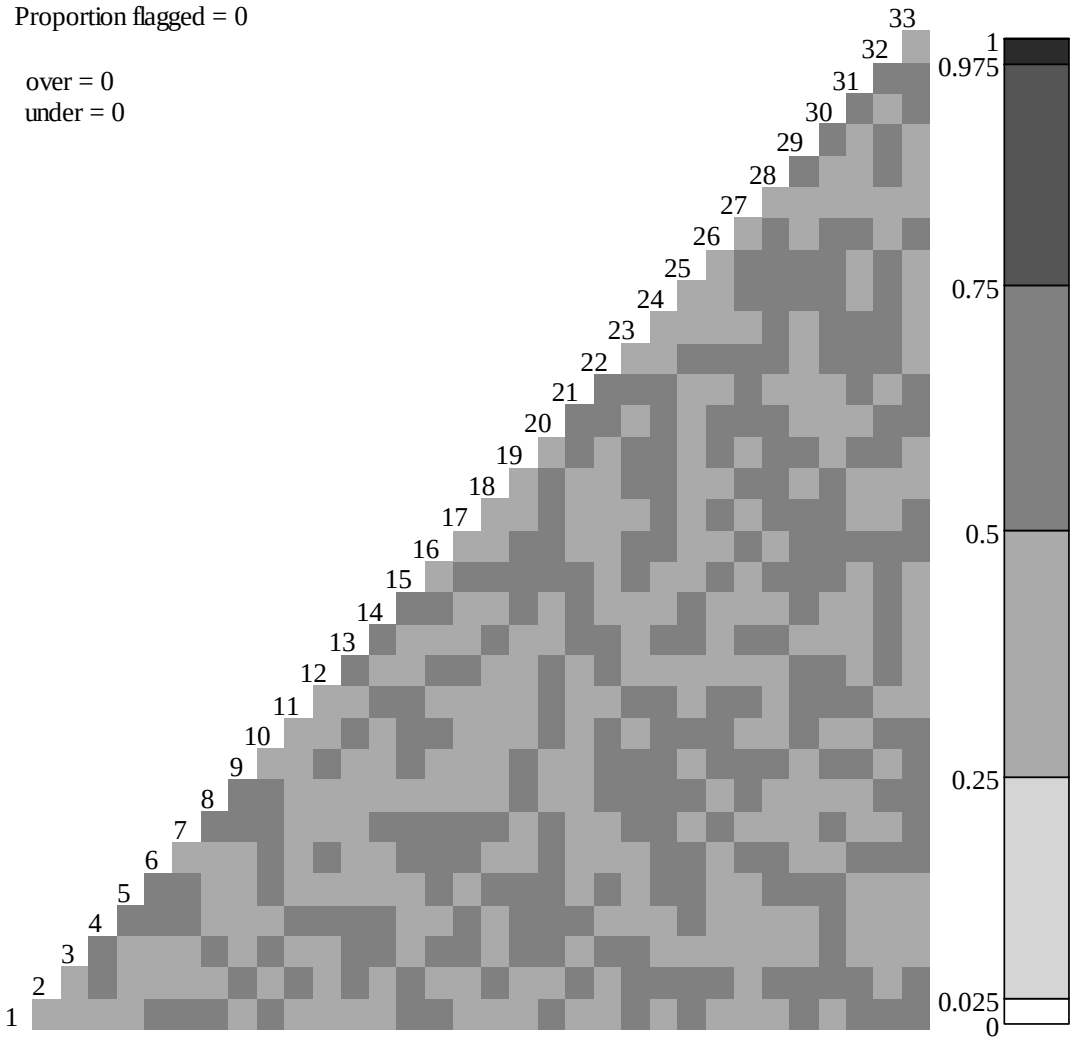



*Figure 26*. Heat map of median PPP-values for bivariate SGDDM for Condition 1.1. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for one pair of observables. White (black) squares indicate that the posterior predicted values were systematically lower (higher) than the realized values. This figure also represents the similar figures produced for bivariate SGDDM and $Q_3$ fit functions for Conditions 1.1, 2.2, 3.3, 4.4, 5.5, and 6.6.

Each of the 528 squares in the figure represents the median of 100 PPP-values for one pair of observables. White squares indicate that the posterior predicted values were systematically lower than the realized values, which means that the association between

the pair of observables was stronger in the observed data than was accounted for by the model (i.e. positive local dependence). Black squares indicate that the posterior predicted values were systematically higher than the realized values, which means that the association between the pair of observables was stronger according to the model than was observed in the realized data (i.e. negative local dependence). Grey squares of varying shades indicate that the median PPP-values were not extreme enough to warrant a flag (i.e. more moderate levels of positive or negative local dependence). As can be seen from this figure, none of the observable pairs had median PPP-values below .025 or above .975 in any of the null conditions. Note that *within* any given replication, it was typical to observe about 21variable pairs with values this extreme, approximately 4% (see Figure 27). However, the identity of those flagged pairs changed across replications, suggesting that the cause was random variation and not systematic misfit.

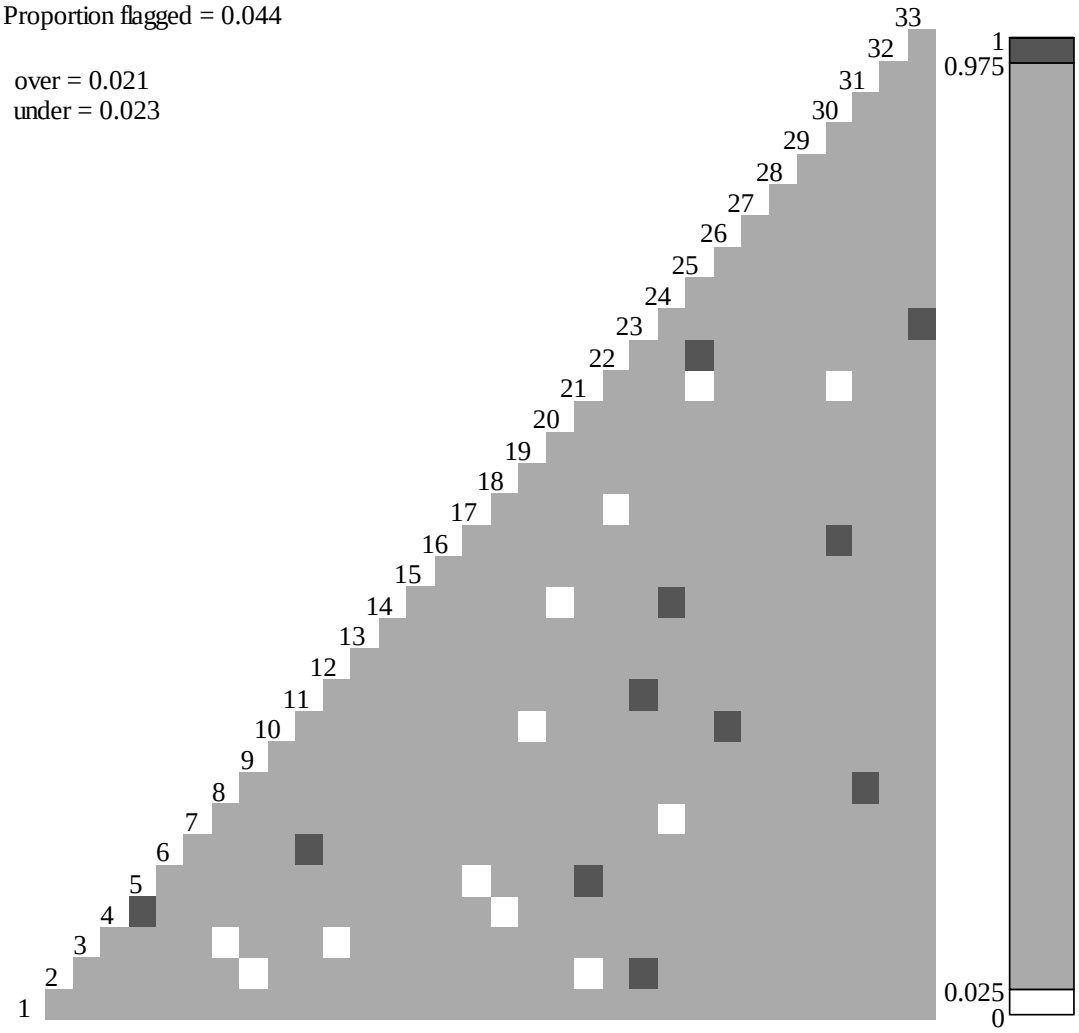



*Figure 27*. Heat map of PPP-values within a single replication. The PPP-values are represented categorically by shaded squares. Each square in this heat map represents the PPP-value from a pair of observables within Replication 1 of Condition 1.1.

From a hypothesis-testing perspective the observed Type-I error rate was around 4%, which is slightly conservative compared to the traditional alpha level of 5%. The pattern of grey squares in this heat map is consistent with a sampling variability explanation, and supports a cautious approach to the interpretation of flagged variable pairs when working with a single observed data set. Clusters of flagged squares, especially among variables

with theoretical connections, are more likely to represent true misfit than isolated flagged squares.

Figure 28 depicts results from Condition 2.1, an example of a heat map with evidence of systematic error (misfit) as opposed to sampling variability.

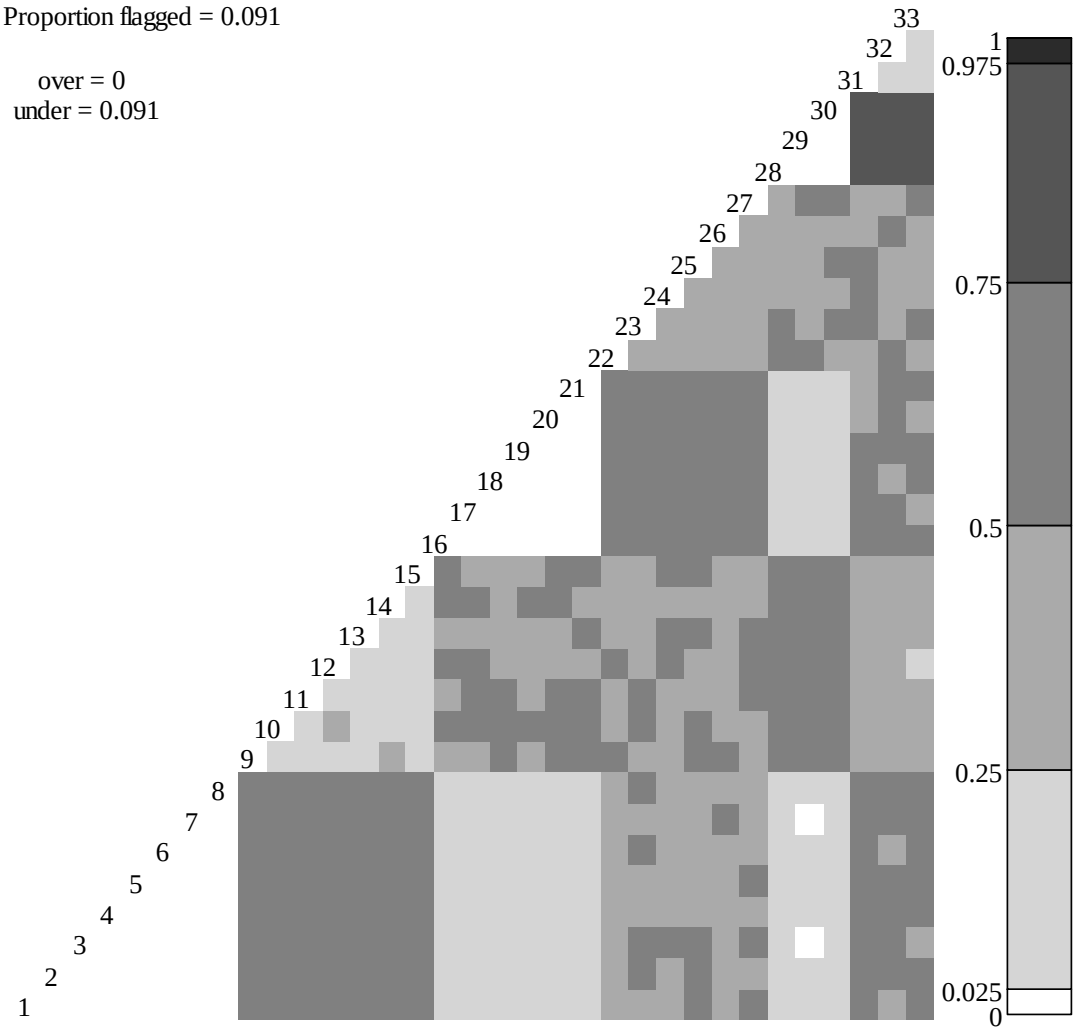



*Figure 28*. Heat map of median PPP-values for bivariate SGDDM or $Q_3$ for Condition 2.1. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for one pair of observables. White (black) squares indicate that the posterior predicted values are systematically lower (higher) than the realized values.

Recall that Model 2 differed from Model 1 only in terms of the addition of a partial mastery class, where the definition of partial mastery consisted of mastery-like performance on a subset of observables that spanned all three latent variables and non-mastery performance on the remaining observables. There were three latent dimensions in the generating model and scoring model, and these latent dimensions were measured by the same observables in both models.

The most important feature in this heat map is the three triangle-shaped clusters of white squares, each representing a local area in the scoring model (Model 1) where the median PPP-values across 100 replications were less than .025. The largest such cluster corresponds to the bivariate associations among Observables 1-8 (i.e. Observable Pairs 1/2, 1/3, 1/4 . . . 7/8), the second cluster corresponds to the bivariate associations among Observables 16-21 (i.e. Observable Pairs 16/17, 16/18, 16/19. . . 20/21), and the third cluster corresponds to the bivariate associations among Observables 28-30 (i.e. Observable Pairs 28/29, 28/30, and 29/30). The characteristic common to these 17 observables is that they were governed by CPT Template 8 in the generating model (Model 2), in which the partial mastery class was as likely as the mastery class to successfully complete the observable. The key difference between the clusters is that each had a different parent: the first cluster depended on $\theta_1$, the second on $\theta_2$, and the third on $\theta_3$. There were residual associations (i.e. positive local dependence) among these observable pairs in the data (generated from Model 2) that were not accounted for by the scoring model (Model 1). In summary, a consequence of fitting a two-class model to data with three classes was that simulees in the partial mastery classes were grouped together with the non-mastery simulees, perhaps because both types diverged from the larger

majority of simulees who performed well across all observables. Poor performance on any cluster of observables therefore would have suggested membership in the only available alternative class. However, for simulees who were partial masters, their tendency to do well on subsets of observables manifested as stronger associations among those observables than the scoring model could account for. This interpretation is supported by evidence at a finer-grained level of detail that will be presented later in the HCI and ICI sections, specifically the proportions of simulees within the latent proficiency profiles and the degree of inter-observable agreement and disagreement within those groups of simulees.

The other two extreme medians in the heat map (corresponding to white squares for Observable Pairs 3/29 and 7/29) are embedded within a rectangular section of observable pairs that had relatively low median PPP-values but did not warrant a flag according to the .025 decision rule. This rectangular group represents the observable pairs that relate Observables 1-8 (the $\theta_1$ group whose intra-cluster observable pairs were all flagged) to Observables 28-30 (the $\theta_3$ group whose intra-cluster observable pairs were all flagged). The analogous observable pairs relating the other flagged triangular clusters exhibited similar levels of positive local dependence (i.e. the rectangular cluster relating Observables 1-8 to Observables 16-21, and the rectangular cluster relating Observables 16-21 to Observables 28-30). The fact that these rectangular clusters relating the white clusters to each other were not white (flagged) themselves tells us that the scoring model did a better job of accounting for the relationships between dimensions than it did within the “partial mastery” clusters of observables within each dimension. This is a clue that the positive local dependence within each white cluster was not due to misspecifications

in the relationships between dimensions. Indeed, the bivariate associations between observables modeled from different dimensions generally contained less local dependence than associations within the same dimension. This accurately reflects the nature of the misspecification, which spanned all the latent variables but was restricted to a subset of observables governed by a particular CPT template. Contemplating the entire pattern of PPP-values, as opposed to focusing only on the flagged values, gives the researcher the best opportunity to distinguish this type of misspecification from other types.

To further detail the observed patterns of median PPP-values in this heat map, it may be useful to categorize the observables according to their latent parentage and their CPT templates. For example, in the generating model for Condition 2.1 (see Figure 4) this categorization yields six categories (or clusters) of observables, each with a different combination of parentage ($\theta_1$, $\theta_2$, or $\theta_3$) and CPT template (8 or 9). Observables 1-8 had $\theta_1$ as their parent and were governed by CPT Template 8. Observables 9-15 shared the same latent parent with the previous observables, but were structured according to CPT Template 9. Observables 9-16 shared CPT Template 8 with Observables 1-8 but the latent parent was $\theta_2$, etc. This categorization is useful because the patterns of median PPP-values in Figure 28 followed the interactions among these clusters.

Consider the first cluster of observables (i.e. Observables 1-8), the cluster parented by $\theta_1$ and governed by CPT Template 8. Observable pairs with both members from this cluster were all flagged due to their extreme positive local dependence, as noted above. By comparison, pairings of an observable from Cluster 1 with an observable from Cluster 2 (shared parentage, different CPT Templates) always yielded modest negative

local dependence, represented in the heat map as grey squares having median PPP-values between .5 and .75. Continuing to move from left to right within the bottom eight rows of the heat map in Figure 28, pairings of an observable from Cluster 1 with an observable from Cluster 3 (different parentage, shared CPT template) consistently yielded positive local dependence, represented in the heat map as the lightest grey squares and having median PPP-values between .025 and .25. Pairings between Cluster 1 and Cluster 4 (different parentage, different CPT templates) did not consistently yield positive or negative local dependence, but were always in the modest range, .25 to .75. Pairings between Cluster 1 and Cluster 5 (different parentage, shared CPT template) yielded positive local dependence, including the only flagged median PPP-values that did not come from intra-cluster pairings of observables (corresponding to white squares for Observable Pairs 3/29 and 7/29). Pairings between Cluster 1 and Cluster 6 (different parentage, different CPT templates) did not consistently yield positive or negative local dependence, but were always in the modest range, .25 to .75.

In summary of Cluster 1, the strongest local dependence was positive local dependence among pairings of observables with the same parent and same CPT template. Next in magnitude was the positive local dependence among pairings of observables with different parents but the same CPT template. Smaller yet in magnitude was the negative local dependence among pairings with the same parent but different CPT templates. Lastly, pairings of observables with different parents and different CPT templates yielded local dependence in the modest range and of varying direction, akin to what was seen due to sampling variability alone in null conditions (see Figure 26).

The second cluster of observables (Observables 9-15) was parented by $\theta_1$ and governed by CPT Template 9. Observable pairs with both members from this cluster tended to exhibit positive local dependence, though two pairings only modestly. Pairings of an observable from Cluster 2 with an observable from Cluster 3 (different parents, different templates) yielded modest local dependence in both directions, as did pairings from Cluster 2 with Cluster 4 (different parents, same template). Pairings between Cluster 2 and Cluster 5 (different parents, different templates) consistently yielded modest negative local dependence (.5 to .75). Pairings between Cluster 2 and Cluster 6 (different parents, same template) yielded modest positive local dependence (.25 to .5), with one pairing below .25. In summary of Cluster 2, the strongest local dependence was positive local dependence among pairings of observables with the same parent and same CPT template, though none of these pairs were flagged and two pairs were modest in magnitude. The remaining pairings were all modest in size, though some were systematic in direction.

The third cluster of observables (Observables 16-21) was parented by $\theta_2$ and governed by CPT Template 8, and was generally consistent in behavior with the patterns observed for Cluster 1. Observable pairs with both members from Cluster 3 were all flagged due to their extreme positive local dependence. Pairings of an observable from Cluster 3 with an observable from Cluster 4 (same parent, different templates) yielded modest negative dependence (.5 to .75). Pairings between Cluster 3 and Cluster 5 (different parents, same template) consistently yielded positive local dependence, represented in the heat map as the lightest grey squares and having median PPP-values between .025 and .25. Pairings between Cluster 3 and Cluster 6 (different parents,

different templates) yielded modest local dependence in both directions (.25 to .75). In summary of Cluster 3, the strongest local dependence was positive local dependence among pairings of observables with the same parent and same CPT template. Next in magnitude was the positive local dependence among pairings of observables with different parents but the same CPT template. Pairings with the same parent but different CPT templates yielded modest negative local dependence. Pairings of observables with different parents and different CPT templates yielded local dependence in the modest range and of varying direction, akin to what was seen due to sampling variability alone in null conditions (see Figure 26).

The fourth cluster of observables (Observables 22-27) was parented by $\theta_2$ and governed by CPT Template 9, and was generally consistent in behavior with the patterns observed for Cluster 2. Observable pairs with both members from Cluster 4 exhibited positive local dependence, though only modestly. Pairings of an observable from Cluster 4 with an observable from Cluster 5 (different parents, different templates) yielded modest local dependence in both directions, as did pairings of Cluster 4 with Cluster 6 (different parents, same template). In summary of Cluster 4, local dependence was always modest in magnitude, and predominantly multidirectional (akin to sampling variability). Systematic positive local dependence did exist among pairings of observables with the same parent and same CPT template, and systematic negative local dependence did exist between pairings of Cluster 4 to Cluster 3 (same parent, different templates).

The fifth cluster of observables (Observables 28-30) was parented by $\theta_3$ and governed by CPT Template 8, and was generally consistent in behavior with the patterns

observed for Clusters 1 and 3. Observable pairs with both members from Cluster 5 were all flagged due to their extreme positive local dependence. Pairings of an observable from Cluster 5 with an observable from Cluster 6 (same parent, different templates) yielded negative local dependence (.75 to .975). This 3x3 cluster of dark grey squares (Observable Pairs (28/31, 28/32, 28/33, 29/31, 29/32, 29/33, 30/31, 30/32, and 30/33) had relatively high median PPP-values, though not extreme enough to be flagged. Recall that the analogous clusters of observable pairs for $\theta_1$ and $\theta_2$ were in the same direction but not as strong in magnitude, which suggests that estimating the parameters for observables dependent upon $\theta_3$ was more difficult in this model than for observables dependent on $\theta_1$ or $\theta_2$. For dimensions with more observables per dimension ($\theta_1$ and $\theta_2$), the scoring model did a better job of identifying the heterogeneity among observables (i.e. those governed by different CPTs), whereas the observables dependent on $\theta_3$ were seen by the scoring model as a more homogenous group than they were generated to be due in part to the lower number of observables dependent on $\theta_3$.

In summary of Cluster 5, the strongest local dependence was positive local dependence among pairings of observables with the same parent and same CPT template. Next in magnitude was the negative local dependence among pairings of observables with different parents but the same CPT template, and the negative local dependence among pairings with the same parent but different CPT templates. Pairings of observables with different parents and different CPT templates yielded modest local dependence which was systematically negative with respect to Cluster 2 and in both directions with respect to Cluster 4.

The sixth cluster of observables (Observables 31-33) was parented by $\theta_3$ and governed by CPT Template 9, and was generally consistent in behavior with the patterns observed for Cluster 2 and Cluster 4. Observable pairs with both members from Cluster 6 exhibited positive local dependence (.025 to .25), though none were flagged. In summary of Cluster 6, the strongest local dependence was positive local dependence among pairings of observables with the same parent and same CPT template, which were similar in magnitude to the negative local dependence observed for pairings with the same parent and different templates. Pairings with different parents and the same template exhibited modest negative local dependence with respect to Cluster 2, but were modest in both directions with respect to Cluster 4. The pairings with different parents and different templates were modest and bidirectional, akin to sampling variability (see Figure 26).

In summary of Figure 28, the overall pattern of the parentage/template effects described above (same/same ≥ different/same ≥ same/different ≥ different/different) provided diagnostic clues as to the characteristic differences between the scoring and generating models. The intra-cluster pairings for any given cluster, which are represented by the triangle-shaped regions bordering the diagonal in Figure 28, had the strongest local dependence within the rows and columns which corresponded to that cluster, but the local dependence was only flagged for clusters governed by the CPT template 8. Collectively these patterns across clusters painted an orderly picture that reflected the impact of the partial mastery misspecification.

Figure 29 depicts a heat map of median PPP-values for the bivariate SGDDM fit function for Condition 3.1. Recall that Model 3 differed from Model 1 only in terms of the addition of seven contextual latent variables.

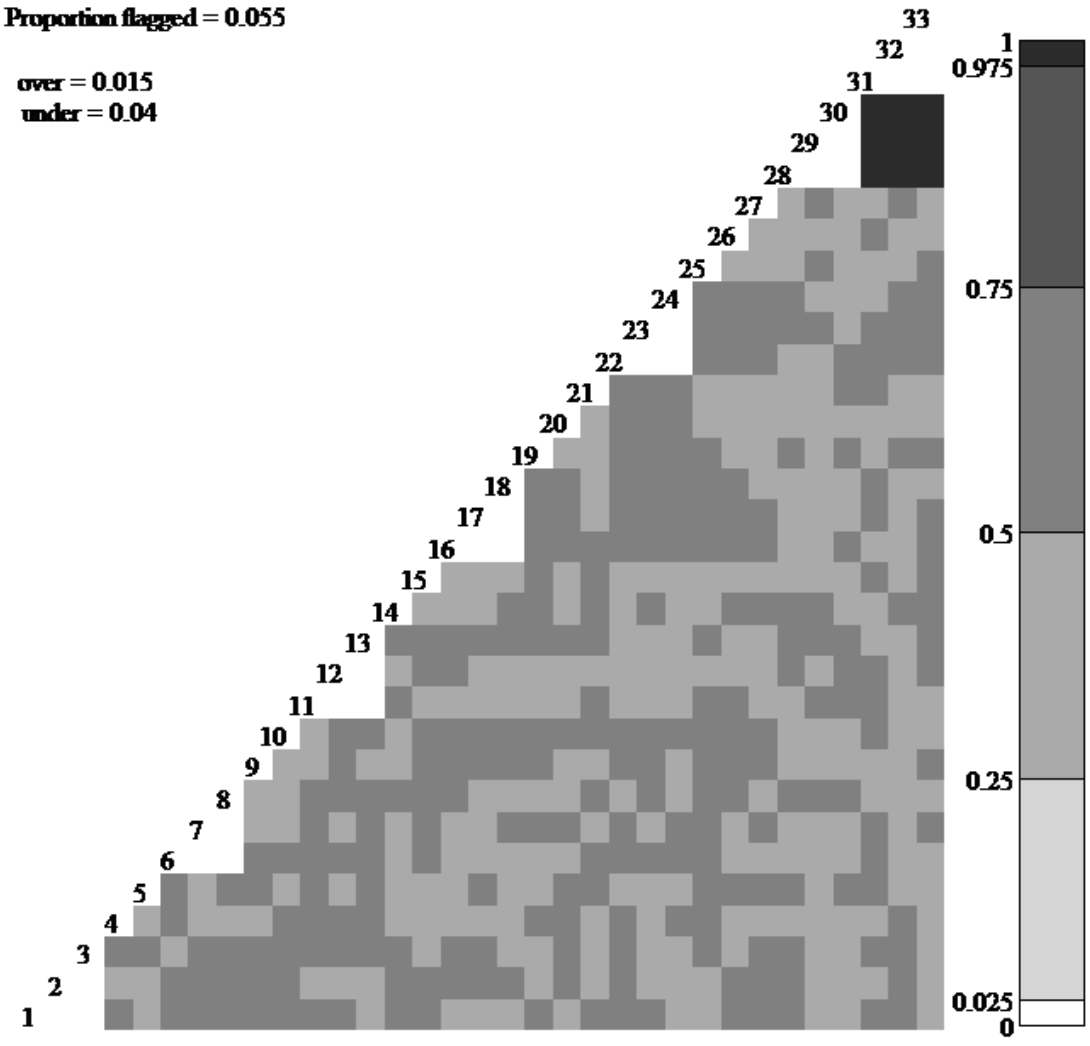



*Figure 29.* Heat map of median PPP-values for bivariate SGDDM or $Q_3$ for Condition 3.1. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for one pair of observables. White (black) squares indicate that the posterior predicted values are systematically lower (higher) than the realized values.

The most extreme medians are represented by seven small triangle-shaped clusters of white squares and one square-shaped cluster of black squares. The white squares indicate residual dependencies in the data not accounted for by the scoring model (positive local dependence), while the black squares indicate that the model overestimated the

dependencies relative to the observed data (negative local dependence). The seven triangle-shaped clusters of white squares each represent a local area in the scoring model (Model 1) where the median PPP-values across 100 replications were less than .025. Each such cluster corresponds to the bivariate associations among three observables: Observables 1-3 (i.e. Observable Pairs 1/2, 1/3, and 2/3), Observables 6-8 (i.e. Observable Pairs 6/7, 6/8, and 7/8), Observables 11-13 (i.e. Observable Pairs 11/12, 11/13, and 12/13), Observables 16-18 (i.e. Observable Pairs 16/17, 16/18, and 17/18), Observables 22-24 (i.e. Observable Pairs 22/23, 22/24, and 23/24), Observables 28-30 (i.e. Observable Pairs 28/29, 28/30, and 29/30), Observables 31-33 (i.e. Observable Pairs 31/32, 31/33, and 32/33). The characteristic common to these 21 observables is that they were influenced by a contextual latent variable in the generating model (Model 3). There were residual associations among these observable pairs in the data that were not accounted for by the scoring model (Model 1).

Regarding the 3x3 cluster of black squares, the nine implicated observable pairs (28/31, 28/32, 28/33, 29/31, 29/32, 29/33, 30/31, 30/32, and 30/33) represent the inter-cluster observable pairs that relate Observables 28-30 (whose intra-cluster observable pairs were all flagged) to Observables 31-33 (whose intra-cluster observable pairs were all flagged). These black squares indicate that the scoring model overestimated the residual dependencies between observable pairs that had $\theta_3$ as their primary latent parent but had different contextual latent variables as their second latent parent. The analogous clusters of observable pairs for $\theta_1$ and $\theta_2$ did not exhibit this pattern, which suggests that accounting for the variability in responses to observables dependent upon $\theta_3$ was more difficult in this model than for observables dependent on $\theta_1$ or $\theta_2$. Consistent with

conditional covariance theory (Zhang and Stout, 1999), local dependence was more pronounced between observable-pairs reflecting different multiple dimensions in the case of $\theta_3$ due to its higher proportion of multidimensional items relative to $\theta_1$ and $\theta_2$.

In contrast to the pattern discussed previously for the partial mastery misspecification, the pattern of median PPP-values for non-flagged variable pairs in the case of this contextual variable misspecification were all modest in magnitude (.25 to .75) and did not show systematic patterns of directionality. This may be due in part to the fact that none of the contextual latent variables spanned multiple primary latent variables, which was the case for the partial mastery misspecification. If a single "large" contextual latent variable had been defined to coincide with the same 17 observables that defined CPT Template 8), then greater similarities would likely have resulted. Stated differently, if definitions of partial mastery had been operationalized as seven clusters of three observables within the context of individual primary latent variables, then the cross-cluster patterns reported previously may have disappeared. This confound in the study design prohibits a definitive answer. Future research could explore whether alternatively framed misspecifications can produce matching patterns of bivariate data model misfit, potentially even to the extent of model equivalence.

Figure 30 depicts a heat map of median PPP-values for the bivariate SGDDM fit function for Condition 4.1. Recall that Model 4 differed from Model 1 along both manipulated dimensions: contextual latent variables and a third latent class.

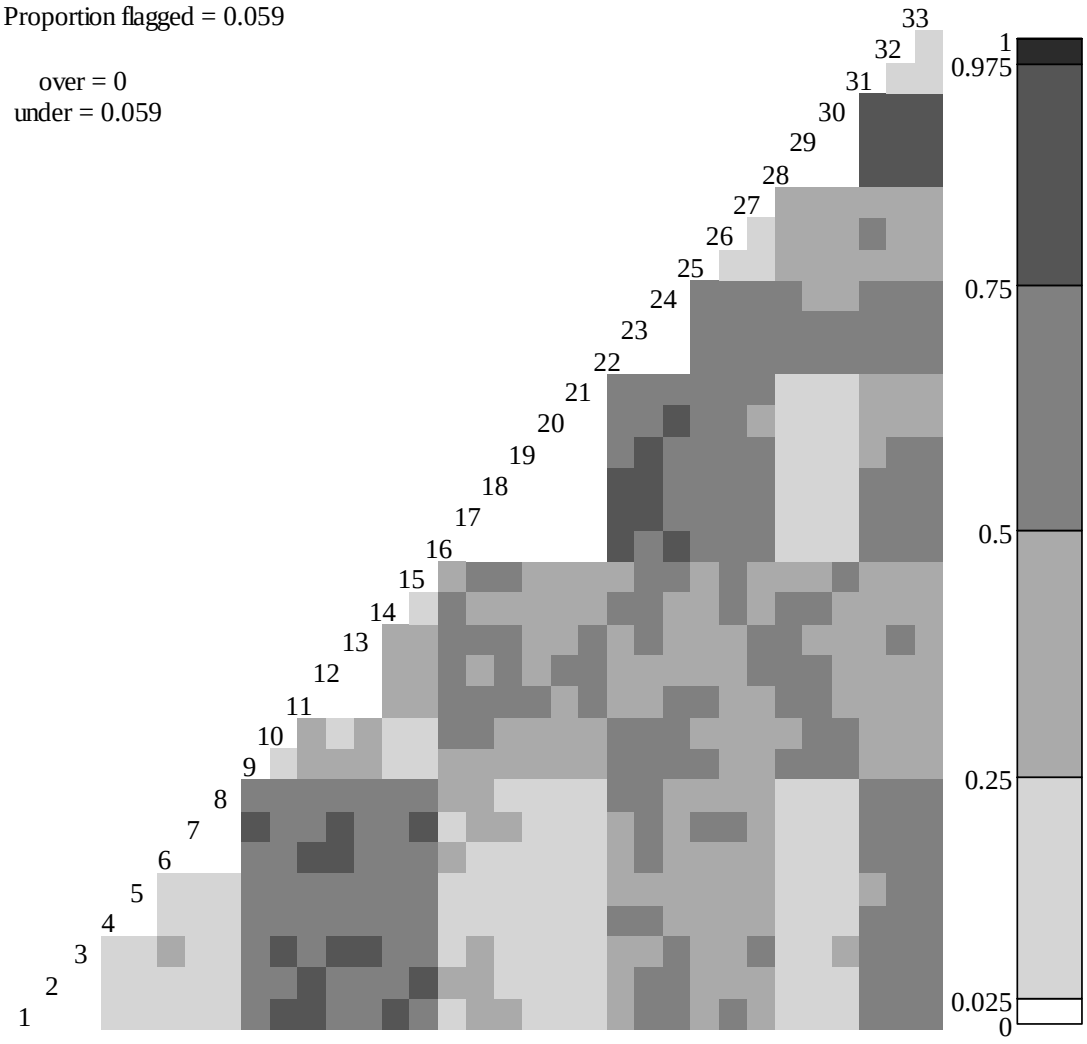



*Figure 30*. Heat map of median PPP-values for bivariate SGDDM or $Q_3$ for Condition 4.1. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for one pair of observables. White (black) squares indicate that the posterior predicted values are systematically lower (higher) than the realized values.

The results depicted in Figure 30 can be generally described as a blending of Figures 28 and 29, as a function of the interaction of the two experimental factors. In Condition 3.1 (Figure 29) all white squares were part of a cluster of three observable pairs corresponding to the intra-member associations of observables with two latent parents.

The differences between Condition 3.1 (Figure 29) and Condition 4.1 (Figure 30) can be described as exceptions to that pattern. Observable pairs 31/32, 31/33, and 32/33 meet that definition but were not flagged. These observables were governed by a CPT table where the partial mastery class acted as the non-mastery class. In Condition 2.1, no such observables were flagged. There were 13 observable pairs that did not meet that definition and were flagged (Observable Pairs 4/5, 16/19, 16/20, 16/21, 17/19, 17/20, 17/21, 18/19, 18/20, 18/21, 19/20, 19/21, 20/21). These observables were generated according to a CPT table where the partial mastery class acted as the mastery class. In Condition 2.1, all such observables were flagged.

Figure 31 depicts a heat map of median PPP-values for the bivariate SGDDM fit function for Condition 5.1. Recall that Model 5 differed from Model 1 only in terms of twelve crossloadings that gave nine observables additional parents.

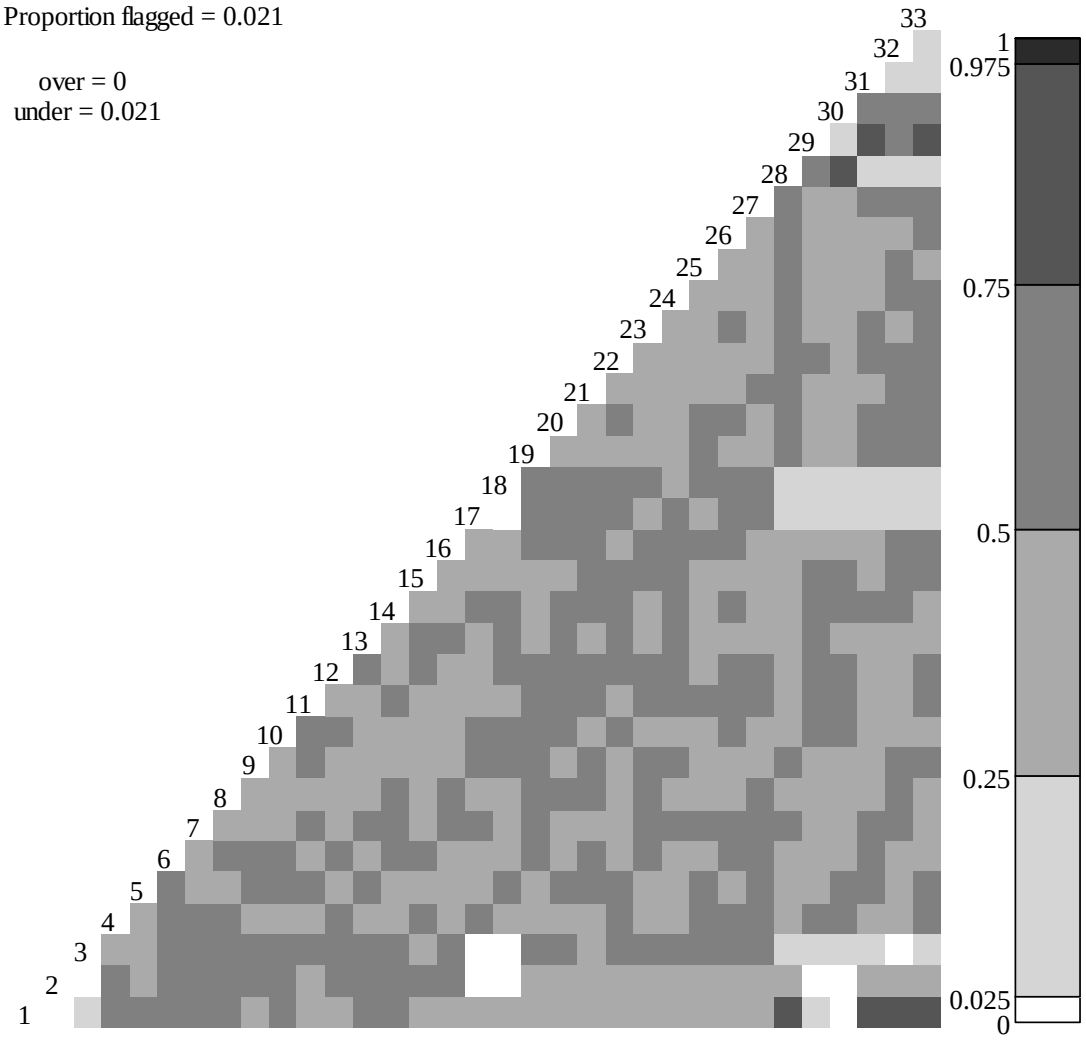



*Figure 31*. Heat map of median PPP-values for bivariate SGDDM or $Q_3$ for Condition 5.1. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for one pair of observables. White (black) squares indicate that the posterior predicted values are systematically lower (higher) than the realized values.

In the heat map there are eleven white squares among 528 total (proportion flagged = .02). Each white square represents a pair of observables with median PPP-values less than .025 (Observable Pairs (1/2, 1/30, 2/3, 2/17, 2/18, 2/29, 2/30, 3/17, 3/18, 3/32, 17/18). There were residual associations among these observable pairs in the data

(generated from Model 5) that were not accounted for by the scoring model (Model 1). These eleven observable pairs were not spatially clustered in the figure, but a meaningful pattern does exist based on the membership of this group: all but one were associations between observables generated with multiple parents (Observables 1-3, 16-18, and 28-30). Of the 36 observables pairs meeting this criterion, 10 were flagged (28%). The remaining flagged pair (Observable Pair 3/32) was anomalous in the sense that it was the only bivariate association flagged in which one member had a single parent and the other member had multiple parents, among 216 such associations. However, this pair can also be thought of as a member of the group of 45 observable pairs in which both member observables were dependent upon $\theta_3$. This group accounted for all but three of the medians outside the central category (.25 to .75) that were not from pairs of multi-parent observables (the remaining three were Observables Pairs 1/31, 1/32, and 1/33, which were relatively high but not flagged). Among these 45 pairs where both observables were a child of $\theta_3$, nine observable pairs were flagged as having extremely low median PPP-values (20%), 24 pairs had relatively low medians but were not flagged (53%), and three pairs had relatively high medians but were not flagged (7%). This finding was part of a larger trend that $\theta_3$ exhibited greater local dependence than $\theta_2$ or $\theta_1$ due to the larger proportion of multidimensionality in $\theta_3$.

Figure 32 depicts a heat map of median PPP-values for the bivariate SGDDM fit function for Condition 6.1. Recall that Model 6 differed from Model 1 along both manipulated factors: the addition of twelve crossloadings, and the addition of a third latent class (partial mastery).

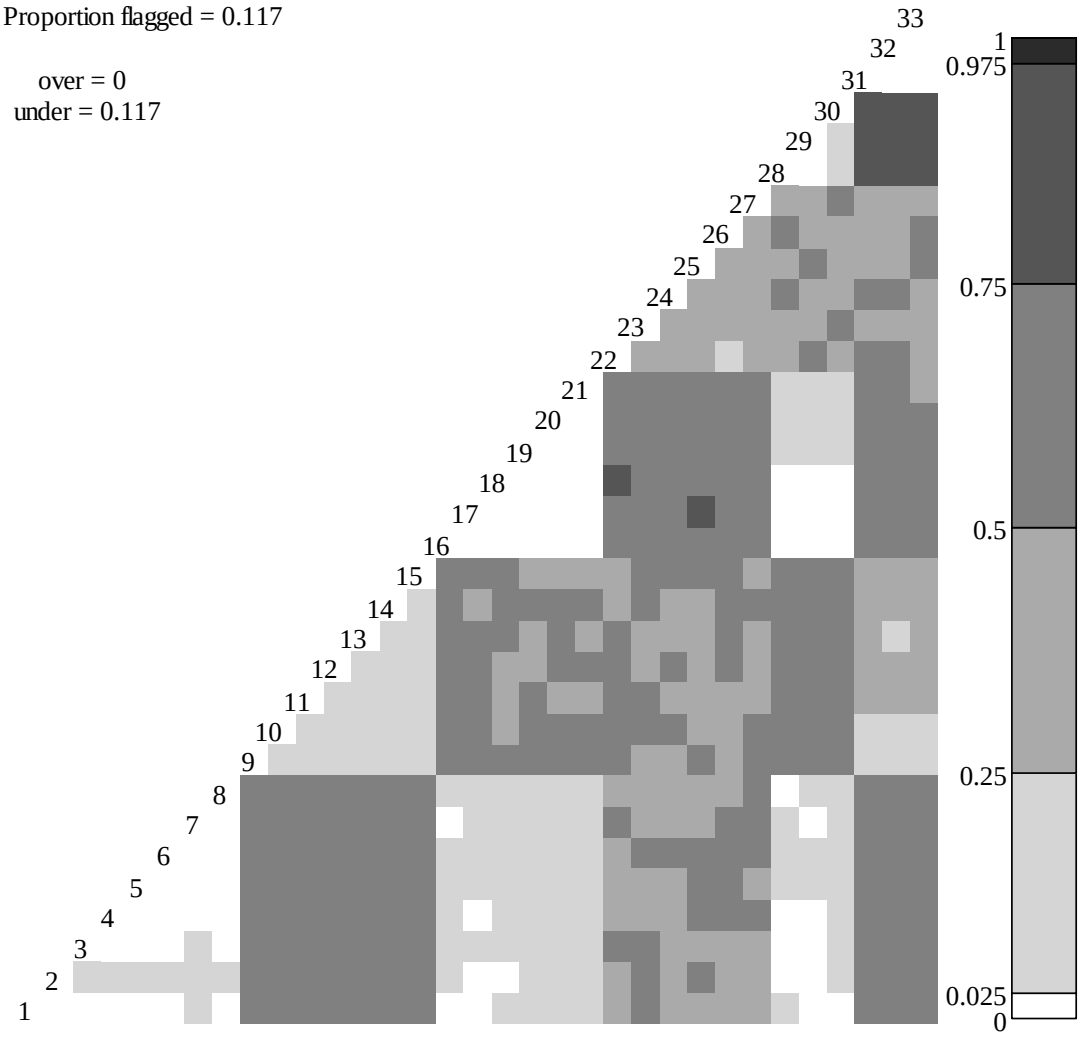



*Figure 32*. Heat map of median PPP-values for bivariate SGDDM or $Q_3$ for Condition 6.1. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for one pair of observables. White (black) squares indicate that the posterior predicted values are systematically lower (higher) than the realized values.

In the heat map there are 62 white squares among 528 total (proportion flagged = .12). Each white square represents a pair of observables with a median PPP-value less than .025. The results depicted for Condition 6.1 (Figure 32) can be generally described as a blending of Condition 2.1 (Figure 28) and Condition 5.1 (Figure 31), as an interaction between the two experimental factors. The pattern in Condition 6.1 (Figure 32) is most

reminiscent of the pattern observed in Figure 28 for Condition 2.1, but with additional white squares in places consistent with the finding from Condition 5.1 (Figure 31) that observable pairs in which both members had multiple parents were sometimes flagged. In Condition 6.1, 25 of the 36 pairings meeting that criterion were flagged (69%), as opposed to 10 out of 36 in Condition 5.1 (28%). The flag rate among observable pairs where both members had multiple parents thus increased in the presence of the additional latent class. The interaction between factors can be viewed from the perspective of the finding from Condition 2.1, namely that the flag rate decreased among observable pairs where both members were governed by a CPT template in which the partial mastery class behaved as the mastery class. In Condition 6.1, 36 of the 46 pairings meeting that criterion were flagged (78%) as opposed to 46 out of 46 in Condition 2.1 (100%). The flag rate among observable pairs with one member having multiple parents was also larger in Condition 6.1 (6 of 216, or 2.8%) relative to Condition 5.1 (1 of 216, or 0.5%). The flagged observable pairs in that category were 4/17, 4/28, 4/29, 7/16, 7/29, and 8/28.

Finally, Observable Pairs 31/32, 31/33, and 32/33 were flagged despite that fact that they did not meet any of the criteria previously associated with flags. These observables shared $\theta_3$ as their only parent and were governed by a CPT where the partial mastery class acted as the non-mastery class. However, this finding can be explained within the context of conditional covariance theory, which states that at higher proportions of multidimensional items multidimensionality can be revealed in terms of item pairs that reflect the primary dimension only. Findings of this description were reported by Levy et al. (2009) in a related study. It appears that the high proportion of multidimensional observables dependent upon $\theta_3$ (three out of six observables, or 50%)

relative to the analogous proportions for $\theta_2$ (three out of twelve observables, or 25%) or $\theta_1$ (three out of fifteen observables, or 20%) was a contributing factor. Note that this factor was present in Condition 5.1, where the same observables had relatively low PPP-values but were not flagged, so the presence of the additional latent class appears to have interacted as well. The additional latent class by itself (i.e. Condition 2.1) resulted in low PPP-values for these same observables, but not extreme enough to be flagged.

### $Q_3$

The results for the $Q_3$ fit function were nearly identical to the bivariate SGDDM results across all conditions. No differences in any of the graphs were perceived, and the only entries in Tables 26-28 that differed between the two functions were the proportion flagged values for Condition 5.5, and the difference there was a single unit in the second decimal place. Consequently the results for the $Q_3$ function were not provided due to their redundancy with the SGDDM bivariate results.

### HCI

As an indicator of person fit, the HCI fit function yielded a PPP-value in each replication of each condition for each person (simulee). However, simulees did not retain their "identities" across replications because new simulees were generated for each replication, so consistency across replications was not a meaningful outcome at the person level (as was consistency across replications at the observable level for the observable-level fit functions). Even though misfitting persons were not explicitly generated into the data, it was of interest to know whether the misfit from model misspecifications would be evident when inspected from a person-fit perspective. Figure 33 shows the HCI PPP-value distributions by condition.

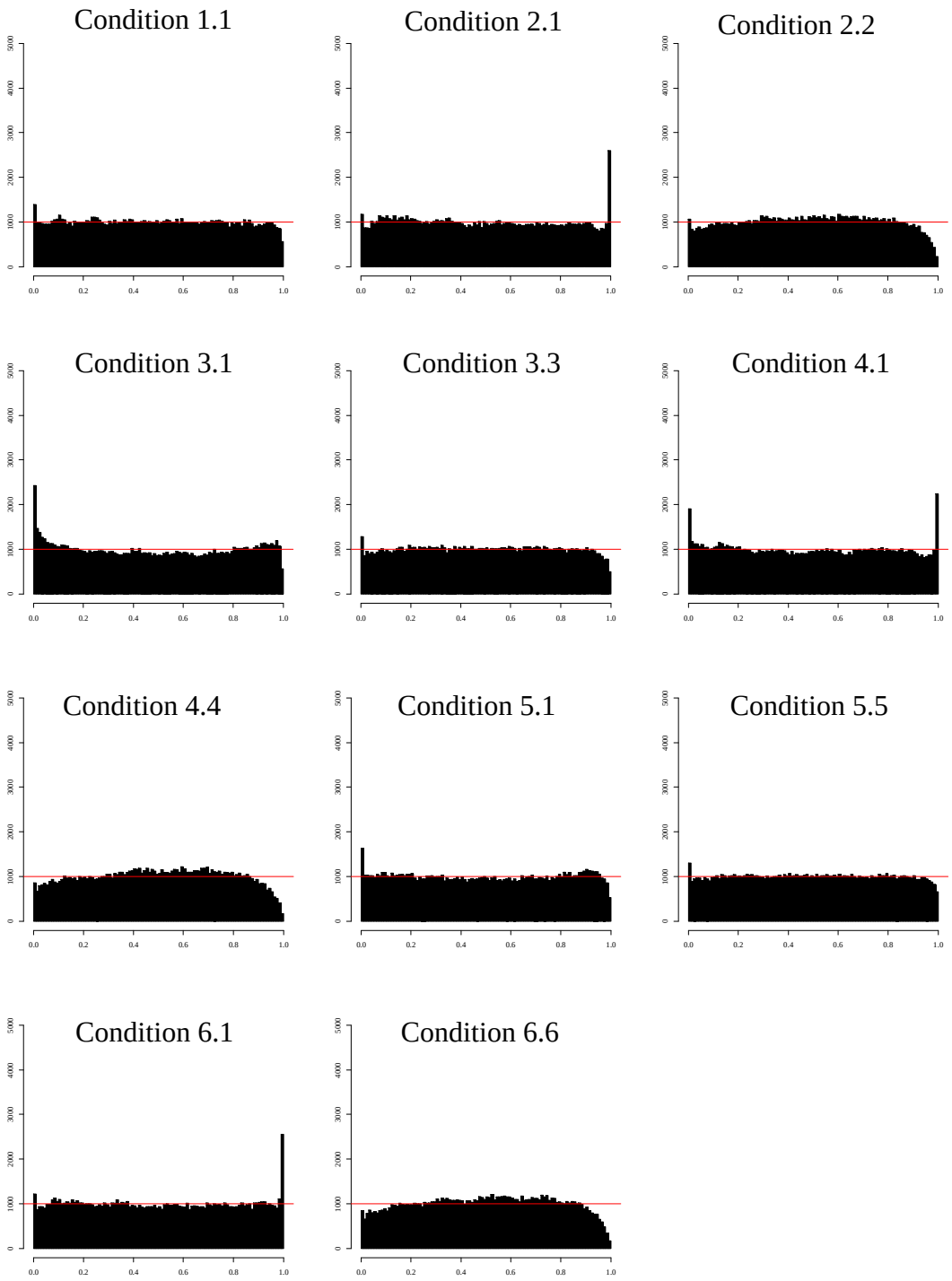



*Figure 33*. Distributions of PPP-values for HCI by condition. The x-axis spans the 0 to 1 range of possible values, with 100 bins at increments of .01. The horizontal line at y=1000 represents uniformity because there are 100,000 values per condition.

Generally speaking, the distributions approached uniformity. Misspecified conditions manifested higher frequencies of extreme values than did null conditions, suggesting that HCI had power to detect misfitting response patterns (simulees). Looking down the last column of Table 26, it can be seen that the proportions of PPP-values flagged as extreme in the null conditions ranged from .03 to .05, while the proportions from the misspecified conditions ranged from .05 to .07, indicating that misspecified conditions on average had approximately 20 additional extreme response patterns relative to null conditions. The medians of the 11 sets of 100,000 PPP-values (1000 persons x 100 replications) were consistently centrally located across conditions, ranging from .48 to .50 (see Table 27). The median effect sizes in the null conditions ranged from .01 to .06 (see Table 28), while the median effect sizes from the misspecified conditions ranged from .02 to .04.

Across all conditions, the empirical sampling distributions of HCI exhibited a negative skew that looked like a mixed modal distribution (see Figure 34), with the smaller mode representing the negative HCI values, i.e. the misfitting response patterns.

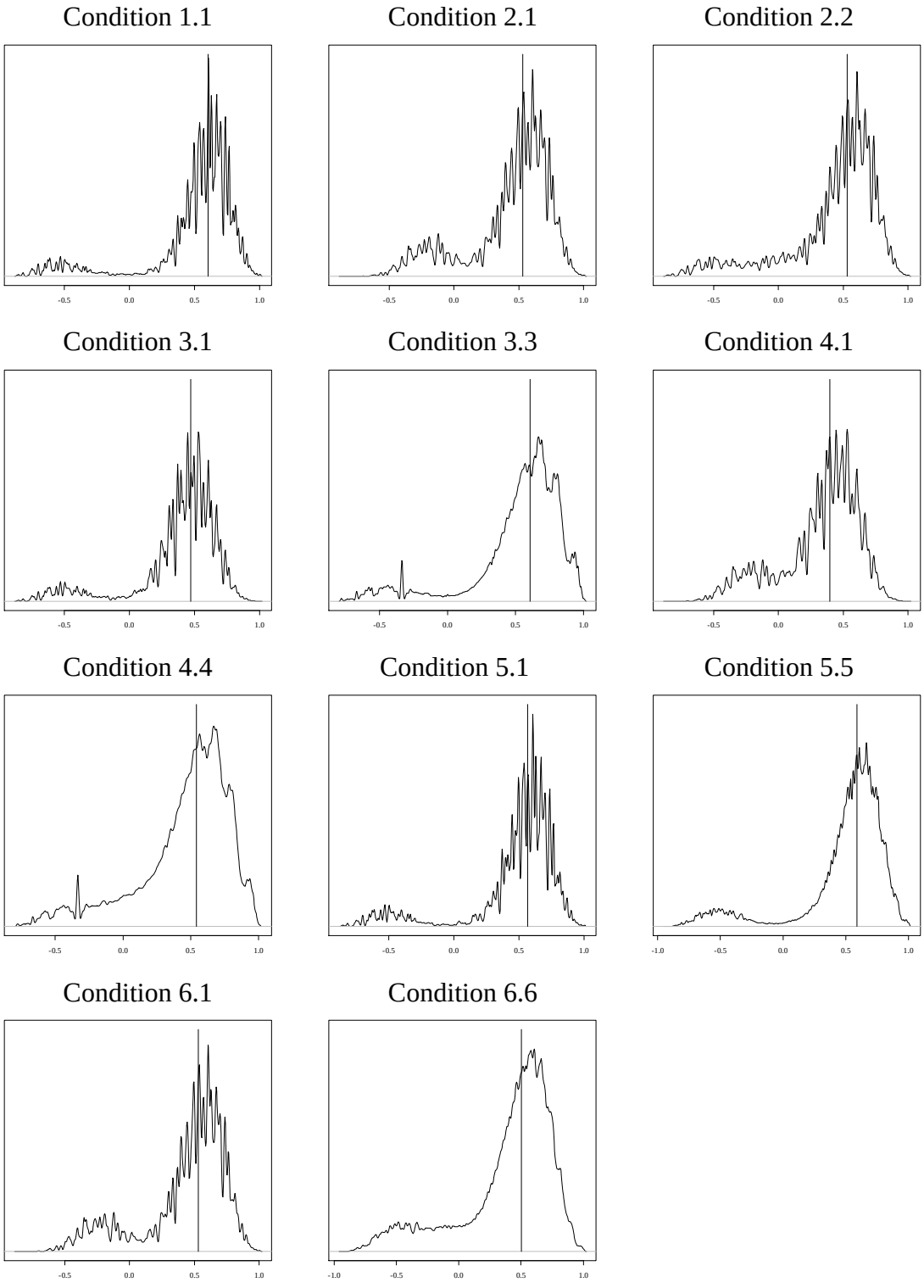



*Figure 34*. Densities of posterior predicted HCI values by condition. The vertical lines represent the means of realized values.

To tease out the characteristics of simulees that were flagged, Tables 29 and 30 report the proportion of simulees from each proficiency profile that were flagged in each condition. Table 29 applies to conditions where the data were generated by a model with two latent classes per primary latent variable (Conditions 1.1, 3.1, 3.3, 5.1, and 5.5), and Table 30 applies to conditions where the data were generated by a model with three latent classes per primary latent variable (Conditions 2.1, 2.2, 4.1, 4.4, 6.1, and 6.6).

Table 29

*Generated primary latent variable proficiencies by condition for conditions with two latent classes per primary latent variable*

| | | Proportion flagged within LV proficiency profile condition | | |
|---|---|---|---|---|
| LV proficiencies $[\theta_1,\theta_2,\theta_3]$ | Proportion within all simulees | Null* | 3.1 | 5.1 |
| [1,1,1] | .08 | .03 | .04 | .04 |
| [1,1,2] | .01 | .02 | .03 | .01 |
| [1,2,1] | .01 | .02 | .04 | .00 |
| [1,2,2] | .00 | .04 | .05 | .04 |
| [2,1,1] | .03 | .03 | .05 | .01 |
| [2,1,2] | .06 | .04 | .05 | .02 |
| [2,2,1] | .08 | .05 | .07 | .08 |
| [2,2,2] | .73 | .05 | .07 | .05 |

*Note*. Flagging refers to simulees with HCI PPP-values less than .025 or greater than .975. *Null conditions with 2 latent classes were 1.1, 3.3, and 5.5.

Table 30

*Generated primary latent variable proficiencies by condition for conditions with three latent classes per primary latent variable*

| | | Proportion flagged within LV proficiency profile | | | |
|---|---|---|---|---|---|
| | | condition | | | |
| LV proficiencies $[\theta_1,\theta_2,\theta_3]$ | Proportion within all simulees | Null* | 2.1 | 4.1 | 6.1 |
| [1,1,1] | .06 | .03 | .54 | .47 | .46 |
| [1,1,2] | .01 | .02 | .23 | .26 | .46 |
| [1,1,3] | .00 | -- | -- | -- | -- |
| [1,2,1] | .02 | .01 | .03 | .06 | .18 |
| [1,2,2] | .01 | .01 | .02 | .03 | .06 |
| [1,2,3] | .00 | -- | -- | -- | -- |
| [1,3,1] | .00 | -- | -- | -- | -- |
| [1,3,2] | .00 | -- | -- | -- | -- |
| [1,3,3] | .00 | -- | -- | -- | -- |
| [2,1,1] | .00 | .01 | .01 | .02 | .05 |
| [2,1,2] | .01 | .00 | .00 | .02 | .02 |
| [2,1,3] | .00 | -- | -- | -- | -- |
| [2,2,1] | .01 | .00 | .00 | .00 | .00 |
| [2,2,2] | .05 | .01 | .00 | .00 | .00 |
| [2,2,3] | .01 | .00 | .00 | .00 | .01 |
| [2,3,1] | .00 | -- | -- | -- | -- |
| [2,3,2] | .03 | .00 | .01 | .01 | .01 |
| [2,3,3] | .00 | .00 | .01 | .02 | .02 |
| [3,1,1] | .00 | -- | -- | -- | -- |
| [3,1,2] | .00 | -- | -- | -- | -- |
| [3,1,3] | .00 | -- | -- | -- | -- |
| [3,2,1] | .00 | -- | -- | -- | -- |
| [3,2,2] | .05 | .01 | .01 | .03 | .01 |
| [3,2,3] | .19 | .02 | .02 | .03 | .02 |
| [3,3,1] | .00 | -- | -- | -- | -- |
| [3,3,2] | .06 | .03 | .03 | .05 | .03 |
| [3,3,3] | .51 | .04 | .04 | .06 | .04 |

*Note*. Flagging refers to simulees with HCI PPP-values less than .025 or greater than .975. *Null conditions with 2 latent classes were 1.1, 3.3, and 5.5.

Generally speaking, the patterns evident in Table 29 for conditions with two latent classes per primary latent variable were relatively weak compared to the patterns evident in Table 30 for the conditions with three latent classes per primary latent variable, suggesting that HCI was less sensitive to the contextual and crossloadings misspecifications than to the partial mastery misspecification. The "null" column within Table 29 shows that simulees in conditions 1.1, 3.3, and 5.5 tended to get flagged at approximately the nominal rate of 5% if they belonged to the profile with the highest proficiencies (i.e. profile [2,2,2]). This profile was by far the largest, comprising about 73% of all simulees in these conditions. Simulees with other proficiency profiles were flagged at slightly conservative rates, with proportions ranging from .02 to .04.

For Condition 3.1, the proportion of flagged simulees from each proficiency profile was elevated by one or two points relative to the null conditions. This indicates that more simulees were flagged marginally in Condition 3.1 than in the null conditions, but that no profile in particular was more susceptible to misfit than the others. For Condition 5.1, differences between the flag rates for each proficiency profile relative to the null conditions were directionally inconsistent and small in magnitude, ranging from -.01 to .03.

Shifting attention to the conditions with three latent classes per primary latent variable in Table 30, the findings were more striking. In Condition 2.1, 54% of the simulees with non-mastery status on all three latent variables (i.e. profile [1,1,1]) were flagged. This translates to about 30 additional flagged simulees per replication relative to the null conditions. Further inspection of these simulees indicated that their realized HCI values tended to be around two standard deviations lower than their posterior predicted

HCI values, which means that the degree of misfit attributed by HCI to these simulees was much stronger in their realized response patterns than in their replicated response patterns. Simulees from proficiency profile [1,1,2] were also flagged at a disproportionately high rate of 23%, while other profiles were flagged at rates similar to the null conditions (differences ≤ .02). Results for Condition 4.1 were generally similar to Condition 2.1, with profiles [1,1,1] and [1,1,2] exhibiting proportions of .47 and .26 respectively. The proportion flagged from profile [1,2,1] was .06 (a difference of .05 relative to the null conditions), while the differences relative to the null conditions for all other profiles were ≤ .02. In Condition 6.1 profiles [1,1,1] and [1,1,2] each had proportions of .46, and the proportion flagged from profile [1,2,1] was .18. Profiles [1,2,2] and [2,1,1] had proportions of .06 and .05 respectively (corresponding to differences of .05 and .04 relative to the null conditions), and the differences relative to the null conditions for all other profiles were ≤ .02. Figure 22 can be used to explain these findings by focusing on Condition 2.1 as an illustrative example. Each row in the figure provides the conditional probability of a correct response to each observable for a given latent proficiency profile. In the first panel, values correspond to the parameters as generated by Model 2, which had three latent classes for each of three latent variables for a total of 27 latent profiles. When data generated from Model 2 were fit to Model 1, which had two latent classes per latent variable for a total of eight latent profiles, the resultant estimates of conditional probability parameters shown in Panel 2 reflect the coerced consolidation of 27 categories into 8 categories. Such a process necessitates changes to the definitions of at least some categories, or changes to the aggregated characteristics of members within a given category, or both. For example, consider the

first row in Figure 22: the "non-mastery" class as generated from Model 2 (i.e. profile [1,1,1] ) had a 20% probability of correctly responding to each observable, but the [1,1,1] profile as estimated by Model 1 when fit to the same data had a substantially larger chance of correctly responding to Observables 1-8, 16-21, and 28-30. Inspection of individual response patterns confirmed that simulees from Condition 2.1 who were flagged for having extreme HCI PPP-values were as a rule assigned to appropriate classes (e.g. response patterns generated from profile [1,1,1] were assigned to profile [1,1,1] by the scoring model). The reason for their extremely low realized HCI values relative to their posterior predicted HCI values was directly tied to the categorical definitions in the scoring model relative to the generating model. Recall that posterior predicted response patterns were generated from estimated model parameters consistent with the values in Panel 2, but realized response patterns were generated from the parameters in Panel 1. It is clear from row by row comparisons that certain profiles in Panel 1 were less likely to generate response patterns that would be consonant with the consolidated categorization of simulees as expressed in Panel 2.

It is important to note that response patterns of individual simulees from throughout the response space could be provided as examples of the underlying principle that simulees were flagged by HCI not because response patterns were necessarily extreme (i.e. the realized HCI value could have been anywhere in the spectrum), but because the disparity between the realized HCI value and posterior predicted HCI values was systematically large. Profile [1,1,1] was the profile most affected by this principle in the investigated conditions, which is why it is being used as an illustrative example, but this principle applies to the other profiles and other conditions as well. The definition of

profile [1,1,1] in the scoring model allowed for a larger degree of success on the assessment than did the definition of profile [1,1,1] in the generating model. HCI detected the fact that simulees who performed low in the realized dataset were often much more successful in the posterior predicted data because they benefitted from the relaxed definition of the lowest category. This is ironic because the lowest performing examinees would serve as prototypes of "non-mastery" according to an intuitive taxonomy, but the scoring model's [1,1,1] profile class was more inclusive of response patterns generated from what were conceived of as the "partial mastery" categories. The simulees with the lowest levels of mastery were thus disproportionately flagged by HCI because their extreme levels of non-mastery were less consistent with the homogenized definition of non-mastery constructed by the categorically restricted scoring model.

HCI provided its strongest evidence of misfit for the latent class misspecification, while the contextual variable misspecification and the crossloadings misspecification displayed weaker evidence. The overall pattern of results suggests that HCI can provide useful fine-grained feedback within a PPMC framework. By inspecting differences in the realized response patterns of flagged persons relative to their posterior predicted response patterns, researchers can identify weaknesses associated with the estimated (scoring) model. While it may be too soon to speculate on the generalizability of these findings, HCI appears promising as a PPMC fit function even when person misfit is not the focal interest.

**ICI**

As noted previously when presenting the distributions of PPP-values aggregated across null and misspecified conditions, the performance of ICI differed somewhat across

misspecified conditions and across observables within some conditions. To help detail these differences, Figures 35-37 depict heat maps for the ICI fit function by condition and observable, with each figure mapping one of the three outcomes used in this study: proportions of extreme PPP-values (Figure 35), median PPP-values (Figure 36), and median effect size (Figure 37).

| | Observable | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| 1.1 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 2.1 | .01 | .01 | .00 | .00 | .01 | .01 | .01 | .01 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .48 | .47 | .47 | .45 | .45 | .52 | .00 | .00 | .00 | .00 | .00 | .00 | .01 | .01 | .01 | .01 | .01 | .01 |
| 2.2 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 3.1 | .09 | .09 | .09 | .00 | .00 | .03 | .04 | .07 | .00 | .00 | .12 | .05 | .05 | .00 | .00 | .00 | .02 | .02 | .00 | .00 | .00 | .04 | .02 | .01 | .00 | .00 | .00 | .04 | .05 | .05 | .03 | .01 | .04 |
| 3.3 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 4.1 | .39 | .36 | .35 | .02 | .00 | .41 | .36 | .40 | .00 | .00 | .03 | .01 | .00 | .00 | .00 | 1.00 | 1.00 | 1.00 | .27 | .31 | .36 | .00 | .00 | .00 | .00 | .00 | .00 | .96 | .96 | .98 | .00 | .00 | .00 |
| 4.4 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 5.1 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 5.5 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |
| 6.1 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .37 | .32 | .37 | .20 | .28 | .27 | .01 | .01 | .01 | .01 | .01 | .01 | .00 | .00 | .00 | .02 | .02 | .01 |
| 6.6 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 | .00 |

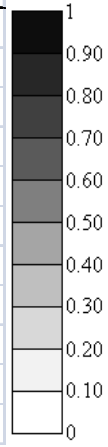



*Figure 35*. Heat map of proportions of extreme PPP-values across all replications for ICI by condition and observable. Each square in the matrix represents a proportion of 100 PPP-values flagged as extreme (less than .025 or greater than .975). Increasingly dark shading is used for larger proportions.

Focusing first on Figure 35, each square in the heat map represents a proportion of 100 PPP-values flagged as extreme (less than .025 or greater than .975). For the null conditions, these proportions represent the observed Type-I error rates, while in the

misspecified conditions they represent observed power. Note that these power rates would be different for alternative alpha levels. In the figure, increasingly dark shading is used for larger proportions. For all null conditions, and additionally for Condition 5.1, the proportion of extreme PPP-values was .00 across all observables (as it was for all conditions for the other observable-level fit functions: PC, $\chi^2$, RPS, and GLS). This is consistent with what was depicted in Figure 11, where it was seen that the distributions of PPP-values in these conditions never included the values defined as extreme. Of greater complexity is the differential performance across observables for the remaining four misspecified conditions (2.1, 3.1, 4.1, and 6.1).

For Condition 2.1, the proportion of flagged PPP-values ranged from .00 to .01 for all observables except Observables 16-21, which ranged from .45 to .52. The six observables with higher flag rates had $\theta_2$ as their parent and were governed by CPT Template 8. Observables with a different CPT structure and a different parent (i.e. Observables 31-33), observables with the same parent but a different CPT structure (i.e. Observables 22-27), or observables with the same CPT structure but a different parent (i.e. Observables 1-8 and 28-30) were flagged at near-zero rates.

The pattern of results for Condition 6.1 was similar to Condition 2.1. The proportion of flagged PPP-values ranged from .00 to .02 for all observables except Observables 16-21, which ranged from .20 to .37. The characteristic common to these six observables with higher flag rates was that they had $\theta_2$ as a parent, and they were governed by a CPT template in which the partial mastery class was as likely as the mastery class to be successful. Observables with different parentage and/or governed by CPT structures in which the partial mastery class was equal to the non-mastery class were flagged at near-zero rates.

For Condition 3.1, the observables with a single latent parent (those corresponding to CPT Template 4: Observables 4-5, 9-10, 14-15, 19-21, and 25-27) had proportions of .00, while proportions for observables with two latent parents ranged from .00 to .12. In Condition 4.1 observables were governed by one of four CPT templates, with somewhat differing results according to template. Observables governed by CPT Template 9 (Observables 9-10, 14-15, and 25-27) had proportions of .00, while proportions for observables governed by CPT Template 13 ranged from .00 to .03. Results within CPT Template 8 were strikingly divergent: Observables 4-5 had proportions of .02 and .00, while Observables 19-21 ranged from .27 to .36. Observables governed by CPT Template 12 all had non-trivial proportions, but rates varied widely: Observables 1-3 and 6-8 ranged from .35 to .41, while Observables 16-18 and 28-30 ranged from .96 to 1.00.

When looking across the columns of Figure 35, the most striking feature is that Observables 16-21 were flagged in at least 20% of the replications for Conditions 2.1, 4.1, and 6.1, but were flagged in 0% of the replications for the remaining conditions (except two observables were flagged in 2% of replications in Condition 3.1). The characteristic common to the three conditions with higher flag rates was the latent class misspecification (presence of the partial mastery class in the generating models but not the scoring models). However, within these three conditions the observables with the same CPT template but different parents (i.e. $\theta_1$ or $\theta_3$) exhibited minimal power. This inconsistency will be discussed in further detail later.

Moving on to the second outcome, Figure 36 depicts a heat map of the median PPP-values across all replications for the ICI fit function by condition and observable.

| | Observable | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| 1.1 | .45 | .45 | .45 | .47 | .44 | .45 | .45 | .45 | .44 | .44 | .42 | .43 | .46 | .45 | .46 | .44 | .45 | .44 | .45 | .47 | .45 | .45 | .45 | .47 | .44 | .46 | .47 | .44 | .43 | .41 | .44 | .45 | .44 |
| 2.1 | .12 | .12 | .12 | .12 | .14 | .12 | .13 | .13 | .41 | .44 | .41 | .40 | .41 | .44 | .42 | .03 | .03 | .03 | .03 | .03 | .02 | .50 | .50 | .48 | .51 | .50 | .53 | .32 | .31 | .29 | .37 | .37 | .38 |
| 2.2 | .45 | .44 | .46 | .45 | .47 | .46 | .47 | .45 | .46 | .47 | .46 | .46 | .45 | .46 | .46 | .43 | .46 | .44 | .45 | .46 | .44 | .44 | .44 | .44 | .45 | .45 | .45 | .49 | .52 | .50 | .46 | .49 | .46 |
| 3.1 | .08 | .07 | .08 | .43 | .42 | .08 | .08 | .09 | .45 | .45 | .08 | .07 | .08 | .46 | .46 | .09 | .11 | .09 | .47 | .46 | .43 | .10 | .11 | .09 | .44 | .45 | .47 | .24 | .22 | .21 | .28 | .26 | .26 |
| 3.3 | .40 | .39 | .40 | .43 | .42 | .39 | .37 | .39 | .43 | .42 | .40 | .41 | .41 | .45 | .44 | .41 | .38 | .39 | .45 | .45 | .42 | .40 | .42 | .40 | .43 | .43 | .44 | .39 | .40 | .40 | .40 | .40 | .40 |
| 4.1 | .03 | .04 | .03 | .15 | .17 | .03 | .03 | .03 | .43 | .43 | .15 | .15 | .14 | .45 | .45 | .00 | .00 | .00 | .04 | .04 | .04 | .38 | .41 | .38 | .52 | .53 | .50 | .00 | .00 | .00 | .70 | .67 | .70 |
| 4.4 | .40 | .41 | .41 | .44 | .49 | .38 | .40 | .39 | .43 | .44 | .42 | .40 | .42 | .45 | .45 | .38 | .39 | .37 | .46 | .45 | .45 | .40 | .44 | .41 | .45 | .44 | .46 | .38 | .38 | .36 | .42 | .43 | .42 |
| 5.1 | .34 | .28 | .32 | .45 | .45 | .47 | .46 | .45 | .42 | .46 | .45 | .46 | .44 | .47 | .46 | .47 | .34 | .32 | .43 | .44 | .44 | .44 | .45 | .44 | .44 | .43 | .43 | .36 | .58 | .57 | .29 | .26 | .27 |
| 5.5 | .40 | .35 | .39 | .44 | .44 | .46 | .45 | .43 | .43 | .47 | .45 | .46 | .45 | .47 | .44 | .40 | .34 | .37 | .44 | .46 | .46 | .44 | .44 | .42 | .44 | .44 | .43 | .38 | .36 | .39 | .42 | .41 | .40 |
| 6.1 | .16 | .19 | .18 | .16 | .16 | .15 | .15 | .16 | .41 | .40 | .42 | .37 | .38 | .37 | .37 | .04 | .04 | .04 | .04 | .04 | .04 | .49 | .50 | .48 | .50 | .49 | .50 | .43 | .43 | .44 | .21 | .20 | .22 |
| 6.6 | .39 | .33 | .42 | .48 | .45 | .46 | .45 | .46 | .47 | .45 | .46 | .45 | .45 | .44 | .46 | .42 | .35 | .44 | .43 | .43 | .44 | .45 | .44 | .44 | .46 | .44 | .45 | .44 | .34 | .41 | .47 | .47 | .48 |

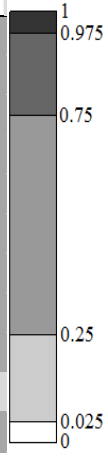





*Figure 36.* Heat map of median PPP-values across all replications for ICI by condition and observable. Each square in the matrix represents the median of 100 PPP-values (1 per replication) for each observable across conditions.

Each square in the matrix represents the median of 100 PPP-values (one per replication). The shading rules were chosen for comparability to a hypothesis-testing framework. Black (white) shading indicates a median in the upper (lower) 2.5% of possible values, yielding 5% of the distribution as "extreme" enough warrant a flag of misfit. Conditions 2.1 and 4.1 were the only conditions to exhibit median values in the flagged ranges. For Condition 2.1, Observable 21 was flagged (median = .02), and for Condition 4.1, Observables 16-18 and 28-30 were flagged (all medians = .00).

Note that different decision criteria would yield different visual patterns and interpretations, but the underlying results (i.e. the distributions of PPP-values) would remain the same. For example, in Condition 3.1 the observables influenced by a contextual latent variable exhibited median PPP-values that were far from centralized but were not below the .025 threshold. This is an example of how PPP-value distributions can yield different interpretations depending on how they are summarized.

Shifting attention to the third outcome, Figure 37 provides a heat map of the median effect sizes across all replications for the ICI fit function by condition and observable. Each square in the matrix represents a median of 100 effect sizes (one per replication). Increasingly darker shading indicates larger median effects.

| | Observable | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
| 1.1 | .1 | .1 | .1 | .1 | .1 | .1 | .2 | .2 | .1 | .2 | .2 | .2 | .1 | .1 | .1 | .1 | .1 | .2 | .1 | .1 | .1 | .1 | .1 | .1 | .2 | .1 | .1 | .1 | .2 | .2 | .2 | .1 | .2 |
| 2.1 | 1.2 | 1.1 | 1.2 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | .2 | .1 | .3 | .3 | .2 | .2 | .2 | 1.9 | 1.9 | 1.9 | 1.9 | 1.9 | 2.0 | .0 | .0 | .1 | .0 | .0 | -.1 | .5 | .5 | .6 | .4 | .4 | .3 |
| 2.2 | .1 | .2 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .2 | .1 | .2 | .1 | .1 | .2 | .2 | .2 | .1 | .1 | .1 | .1 | .0 | -.1 | .0 | .1 | .0 | .1 |
| 3.1 | 1.4 | 1.4 | 1.4 | .2 | .2 | 1.4 | 1.4 | 1.4 | .1 | .1 | 1.4 | 1.4 | 1.4 | .1 | .1 | 1.3 | 1.3 | 1.4 | .1 | .1 | .2 | 1.3 | 1.2 | 1.3 | .2 | .1 | .1 | .7 | .8 | .8 | .6 | .6 | .6 |
| 3.3 | .3 | .3 | .3 | .1 | .2 | .3 | .3 | .3 | .2 | .2 | .3 | .3 | .2 | .2 | .1 | .2 | .3 | .3 | .1 | .1 | .2 | .3 | .2 | .3 | .2 | .2 | .2 | .3 | .3 | .3 | .3 | .3 | .3 |
| 4.1 | 1.8 | 1.8 | 1.8 | 1.1 | .9 | 1.8 | 1.8 | 1.9 | .2 | .2 | 1.1 | 1.1 | 1.1 | .2 | .1 | 3.0 | 3.0 | 3.0 | 1.7 | 1.8 | 1.8 | .3 | .2 | .3 | .0 | -.1 | .0 | 3.0 | 2.9 | 3.0 | -.5 | -.5 | -.5 |
| 4.4 | .3 | .2 | .2 | .1 | .0 | .3 | .3 | .3 | .2 | .2 | .2 | .3 | .2 | .1 | .1 | .3 | .3 | .3 | .1 | .1 | .2 | .2 | .1 | .2 | .1 | .2 | .1 | .3 | .3 | .3 | .2 | .2 | .2 |
| 5.1 | .4 | .6 | .5 | .2 | .2 | .1 | .2 | .1 | .2 | .1 | .1 | .1 | .2 | .1 | .1 | .1 | .4 | .5 | .2 | .2 | .2 | .2 | .2 | .2 | .1 | .2 | .2 | .3 | -.2 | -.2 | .5 | .6 | .6 |
| 5.5 | .3 | .4 | .3 | .2 | .2 | .1 | .1 | .2 | .2 | .1 | .1 | .1 | .1 | .1 | .2 | .3 | .4 | .4 | .2 | .1 | .1 | .2 | .2 | .2 | .1 | .2 | .2 | .3 | .4 | .2 | .2 | .2 | .3 |
| 6.1 | 1.0 | .9 | .9 | 1.0 | 1.0 | 1.1 | 1.0 | 1.0 | .3 | .3 | .3 | .3 | .3 | .3 | .3 | 1.7 | 1.8 | 1.8 | 1.7 | 1.7 | 1.7 | .1 | .0 | .1 | .0 | .0 | .0 | .2 | .2 | .2 | .8 | .8 | .8 |
| 6.6 | .3 | .4 | .2 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .1 | .2 | .2 | .1 | .2 | .4 | .2 | .2 | .2 | .2 | .1 | .2 | .1 | .1 | .2 | .2 | .2 | .4 | .2 | .1 | .1 | .1 |

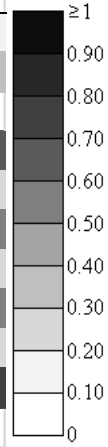



*Figure 37*. Heat map of median effect size values across all replications for ICI by condition and observable. Each square in the matrix represents the median of 100 effect sizes (1 per replication) for each observable across conditions.

The median effects in null conditions ranged from -0.05 to 0.43. Some differentiation of effect size was evident across observables according to CPT templates, though not as strongly as in misspecified conditions. For example, in Condition 6.6 the observables with three latent parents (governed by CPT Template 16) each had median effect sizes around 0 .4, while median effect sizes for all other observables in that condition were less than 0.3. In Conditions 3.3 and 4.4, observables influenced by a contextual latent variable (CPT Templates 11-13) tended to have slightly larger median effect sizes than observables with a single latent parent. These findings suggest that the effect size metric was somewhat sensitive to the fact that some conditional probabilities were more difficult to estimate than others, and that sampling variability alone caused noticeable differences in effect sizes across some CPT templates.

In the misspecified conditions, the differentiation of values across observables according to CPT template tended to be much stronger. For example, in the second row of Figure 37, it can be seen that for Condition 2.1, median effects for observables governed by CPT Template 8 (Observables 1-8, 16-21, and 28-30) ranged from about 0.3 to 2.0, while the observables governed by CPT Template 9 (Observables 9-15, 22-27, and 31-33) ranged from about -0.1 to 0.3. A pattern was also evident within the observables associated with CPT 8, with larger effects for observables dependent upon $\theta_2$ (1.9 to 2.0) versus $\theta_1$ (1.1 to 1.2), both of which were much larger than those for $\theta_3$ (0.3 to 0.6).

For Condition 3.1, median effects for observables with a single latent parent (CPT Template 4: Observables 4-5, 9-10, 14-15, 19-21, and 25-27) ranged from about 0.1 to 0.2, while observables with the additional influence of a contextual latent variable (CPT Template 11: Observables 1-3, 6-8, 11-13, 16-18, 22-24, and 28-33) ranged from about

0.6 to 1.4. A pattern was also evident within the observables associated with CPT 11, with larger effects for observables associated with $\theta_1$ (0.1 to 0.4) or $\theta_2$ (1.2 to 1.4) than $\theta_3$ (0.6 to 0.8).

Condition 4.1 had median effects ranging from 1.8 to 3.0 for the twelve observables with two latent parents and which were mastered by the partial mastery class (those corresponding to CPT Template 12: Observables 1-3, 6-8, 16-18, and 28-30), while the observables corresponding to CPT Template 8 (Observables 4-5 and 19-21) ranged from 0.9 to 1.8, and all other observables ranged from -0.1 to 0.3. Patterns within the CPT templates exhibiting larger effects were again evident. Observables associated with CPT 12 showed larger effects for observables associated with $\theta_2$ (3.0) or $\theta_3$ (2.9 to 3.0) than $\theta_1$ (1.8 to 1.9). Observables associated with CPT 8 showed larger effects for observables associated with $\theta_2$ (1.7 to 1.8) than $\theta_1$ (0.9 to 1.1). Observables associated with CPT 13 (Observables 11-13, 22-24, and 31-33) showed larger effects for observables associated with $\theta_1$ (about 1.1) than $\theta_3$ (-0.5) or $\theta_2$ (0.2 to 0.3). Note that some of the large median effect sizes in this condition were associated with observables that were not flagged according to the proportion of extreme PPP-values. This illustrates the importance of not relying on the effect sizes exclusively, but rather considering them in the context of the PPP-values.

Condition 5.1 exhibited less clean patterns across observables than the other misspecified conditions. Most of the single-parent observables in this condition had effect sizes in the 0.1 to 0.2 range, but strong exceptions occurred for Observables 31-33, which had median values of approximately 0.5 to 0.6. The observables with three latent parents (CPT Template 14: Observables 2, 17, and 29) had median effects of similar magnitude

(0.4 to 0.6), while observables with two latent parents (CPT Template 11: Observables 1, 3, 16, 18, 28, and 30) ranged from about -0.2 to 0.5.

For Condition 6.1, the median effects for observables from CPT Template 9 (Observables 9-15, 22-27, and 31-33) ranged from 0.0 to 0.8, while those from CPT Template 8 (Observables 4-8 and 19-21) ranged from 1.0 to 1.7. Patterns within the CPT templates exhibiting larger effects were again evident. Observables associated with CPT 8 showed larger effects for observables associated with $\theta_2$ (1.7) than $\theta_1$ (1.0 to 1.1). Observables associated with CPT 13 (Observables 11-13, 22-24, and 31-33) showed larger effects for observables associated with $\theta_3$ (0.8) than $\theta_1$ (0.3) or $\theta_2$ (0.0 to 0.1). Among observables with two or three latent parents (Templates 15 and 16), the observables associated with $\theta_2$ (Observables 16-18) showed larger median effects (1.7 to 1.8) than did those of $\theta_1$ (0.9 to 1.0) or $\theta_3$ (0.2).

To clarify the mechanisms underlying the performance of ICI, Condition 2.1 is used as an illustrative example. Due to similarities between ICI and HCI, the principles discussed previously in the context of HCI results (see Figure 22) are relevant for understanding the performance of ICI. The results for any given observable using ICI can be thought of as a weighted average of the effects described for particular proficiency profiles in the HCI results. As with HCI, the essence of ICI boils down to comparisons between observed (or posterior predicted) response patterns and the response patterns implied by the *Q*-matrix for the scoring model. The proportion of mismatches when comparing the observed and implied responses (also referred to as “mismatches per comparison” or MPC) is rescaled to form the index value: ICI = 1 – 2*MPC. To understand why ICI PPP-values tended to be more extreme for observables parented by

$\theta_2$ than by $\theta_1$ or $\theta_3$, consider Figure 38, which reports the mismatches per comparison (MPC) for simulees within each proficiency level of each latent variable.

| | | Mean mismatches per comparison (realized data) | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Proportion | Observable ($x_j$) | | | | | | | | | | | | | | |
| $\theta_1$ | of simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 1 | .10 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 |
| 2 | .10 | .50 | .50 | .50 | .50 | .50 | .50 | .50 | .50 | .53 | .53 | .52 | .53 | .53 | .53 | .53 |
| 3 | .80 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 |
| | | | | | | | | | | | | | | | | |
| $\theta_2$ | | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | | | |
| 1 | .08 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | | | |
| 2 | .33 | .52 | .52 | .52 | .52 | .52 | .52 | .52 | .52 | .52 | .52 | .52 | .52 | | | |
| 3 | .59 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | | | |
| | | | | | | | | | | | | | | | | |
| $\theta_3$ | | 28 | 29 | 30 | 31 | 32 | 33 | | | | | | | | | |
| 1 | .10 | .32 | .32 | .32 | .32 | .32 | .32 | | | | | | | | | |
| 2 | .20 | .54 | .54 | .54 | .54 | .54 | .53 | | | | | | | | | |
| 3 | .70 | .32 | .32 | .32 | .32 | .32 | .32 | | | | | | | | | |
| | | | | | | | | | | | | | | | | |
| | | Mean mismatches per comparison (posterior predicted data) | | | | | | | | | | | | | | |
| | Proportion | Observable ($x_j$) | | | | | | | | | | | | | | |
| $\theta_1$ | of simulees | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 1 | .17 | .48 | .48 | .48 | .48 | .48 | .48 | .48 | .48 | .40 | .40 | .40 | .40 | .40 | .40 | .40 |
| 2 | .83 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .32 | .33 | .33 | .33 | .33 | .33 | .33 | .33 |
| | | | | | | | | | | | | | | | | |
| $\theta_2$ | | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | | | |
| 1 | .38 | .53 | .53 | .53 | .53 | .53 | .53 | .47 | .47 | .47 | .47 | .47 | .47 | | | |
| 2 | .62 | .32 | .32 | .32 | .32 | .32 | .32 | .33 | .33 | .33 | .33 | .33 | .33 | | | |
| | | | | | | | | | | | | | | | | |
| $\theta_3$ | | 28 | 29 | 30 | 31 | 32 | 33 | | | | | | | | | |
| 1 | .25 | .50 | .50 | .50 | .42 | .42 | .42 | | | | | | | | | |
| 2 | .75 | .33 | .33 | .33 | .35 | .35 | .35 | | | | | | | | | |

*Figure 38*. Mean MPC by latent proficiency and observable for Condition 2.1. Upper panel represents realized data generated from Model 2. Lower panel represents posterior predicted data generated by the estimated parameters when Model 1 was fit to data generated from Model 2.

The upper panel represents the realized data and the lower panel represents the posterior predicted data. Simply put, there was a greater proportion of simulees with a partial mastery proficiency level for $\theta_2$ than for $\theta_1$ or $\theta_3$, and it was these partial-mastery simulees whose response patterns were more likely to produce mismatches relative to the response patterns implied by the *Q*-matrix for the scoring model. The partial mastery

class was by definition in conflict with an assumption underlying ICI: that the *Q*-matrix is sufficient to predict response patterns. In the *Q*-matrix, proficiency was a dichotomous prospect, with success or failure implied by the presence or absence of the latent trait. For simulees with partial mastery status on a latent trait, the aforementioned assumption did not hold. The inability of the *Q*-matrix to account for the response behavior of partial-mastery simulees was borne out by the increased rate of mismatches for partial-mastery simulees in the realized data, but it was the differing proportions of partial-mastery simulees across latent variables that impacted the estimation of conditional probability parameters for the scoring model, the subsequent generation of posterior predicted data, and the relative value of the resultant posterior predicted ICI values to the observed ICI values (i.e. the PPP-values).

From the viewpoint offered by Figure 38 it is clear that in the realized data the typical level of mismatch was relatively consistent across observables for simulees within a given proficiency level. Proficiency values of 1 or 3 (non-mastery or mastery) corresponded to approximately 1 mismatch in every 3 comparisons, while proficiency values of 2 (partial mastery) corresponded to approximately 1 mismatch for every 2 comparisons. The key point of the upper panel is that the three latent variables differed with respect to the proportion of simulees having the higher mismatch rate (i.e. simulees in the partial mastery class). In the realized data, 33% of simulees were partially proficient on $\theta_2$, while 20% were partially proficient on $\theta_3$ and 10% were partially proficient on $\theta_1$.

The lower panel of Figure 38 tells the analogous if slightly more complicated story about the absolute fit of each observable in the posterior predicted data. While

simulees with proficiency values of 2 (mastery) had relatively consistent levels of mismatch across all observables (approximately 1 mismatch per 3 comparisons as was the case in the realized data), the level of mismatch for simulees with proficiency values of 1 (non-mastery) aligned with the blocks of observables corresponding to combinations of CPT template and latent parent. MPC values for simulees with proficiency values of 1 was approximately 1 mismatch per 2 comparisons for Observables 1-8, 16-21, and 28-30, which were the observables where the partial mastery class responded like the mastery class (CPT Template 8), while for the remaining observables (CPT Template 9) MPC values were somewhere between the two levels reported thus far: Observables 9-15(MPC = .40, Observables 22-27(MPC = .47), and Observables 31-33(MPC = .42). The divergence of MPC values within proficiency level 1 for each latent variable was reflective of the fact that this less-proficient class in the scoring model was a coerced homogenization of the heterogeneity that existed in the realized data (see Figure 22). However, the consequences for the posterior predicted data were more detectable with respect to $\theta_2$ due to the larger proportion of partial-mastery simulees. Specifically, the estimated conditional probability parameters for members of the non-mastery classes in the scoring model (which allowed for only two classes) represented a composite of the partial- mastery and non-mastery classes that existed in the realized data. For $\theta_2$, the estimated conditional probabilities of the relevant observables was pulled higher (toward the level of the mastery class) by the mastery-level performance of the partial mastery class on those observables, while the analogous parameters for the observables underlying $\theta_1$ and $\theta_3$ were impacted to a lesser degree commensurate with their smaller proportions of partial-mastery simulees.

Table 31 depicts a simplified example of how conditional probability patterns impact the ICI computations. The purpose of this table is to further illustrate the process by which patterns of conditional probabilities such as those provided in Figure 22 translate into MPC values like those provided in Figure 38.

Table 31

*Simplified example of the impact of conditional probability patterns on ICI outcomes*

| Realized CP | | Postpred CP | | Realized MPC | Postpred MPC | Realized ICI | Postpred ICI | PPP-value |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | .5 |
| 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | .5 |
| 0 | 1 | 0 | 1 | 1 | 1 | -1 | -1 | .5 |
| .80 | .80 | .80 | .80 | .32 | .32 | 0.36 | 0.36 | .5 |
| .20 | .20 | .20 | .20 | .32 | .32 | 0.36 | 0.36 | .5 |
| .20 | .80 | .20 | .80 | .68 | .68 | -0.36 | -0.36 | .5 |
| .20 | .80 | .20 | .67 | .68 | .60 | -0.36 | -0.20 | > .5 |
| .20 | .80 | .20 | .52 | .68 | .51 | -0.36 | -0.02 | >> .5 |
| .20 | .80 | .20 | .45 | .68 | .47 | -0.36 | 0.06 | >>> .5 |
| .20 | .80 | .80 | .80 | .68 | .32 | -0.36 | 0.36 | >>>> .5 |
| 0 | 1 | 1 | 1 | 1 | 0 | -1 | 1 | >>>>> .5 |
| .80 | .80 | .20 | .45 | .32 | .47 | 0.36 | 0.06 | < .5 |
| .80 | .80 | .20 | .52 | .32 | .51 | 0.36 | -0.02 | << .5 |
| .80 | .80 | .20 | .67 | .32 | .60 | 0.36 | -0.20 | <<< .5 |
| .80 | .80 | .20 | .80 | .32 | .68 | 0.36 | -0.36 | <<<< .5 |
| 1 | 1 | 0 | 1 | 0 | 1 | 1 | -1 | <<<<< .5 |

For the purposes of this example consider only two observables, where the scoring model implies that both observables reflect a single latent ability, and that success on both observables requires this latent ability according to the associated *Q*-matrix. The first two columns in Table 31 represent conditional probabilities for the two theoretical observables in realized data. The next two columns represent conditional probabilities for the same two observables in posterior predicted data. The remaining columns are

computations based upon the first four columns: realized and posterior predicted MPC and ICI values, and PPP-values. MPC is the ratio of mismatches to comparisons. ICI is a linear transformation of MPC: ICI = 1 – 2*MPC. PPP-values are the proportion of posterior predicted ICI values that meet or exceed the realized ICI value. Note that because there is only a single inter-observable comparison (per simulee) in this simplified example, the MPC, ICI, and PPP-values for both observables are equal. Therefore a single column was used to represent values that apply to both observables.

In the first row of Table 31, the conditional probability of success in the realized data on both observables in the simplified example was 1, meaning that all simulees always completed both observables correctly. The corresponding realized MPC value of 0 reflects the fact that there were no mismatches between the observed response patterns and the response patterns implied by the scoring model *Q*-matrix (i.e. there was zero disagreement between observed responses that were expected to agree according to the *Q*-matrix). The corresponding realized ICI value of 1 indicates perfect fit between the observed and model-implied responses to the observable(s). The posterior predicted values in this row mirror the realized values. The PPP-value of 0.5 reflects the expectation that upon many replications of a PPMC process, the realized ICI (or MPC) value for each observable should be centered with respect to the posterior predicted ICI (or MPC) values.

The next five rows of Table 31 illustrate that whenever the pattern of conditional probabilities is the same for realized and posterior predicted data, the PPP-values should be centered (indicating good fit) regardless of the absolute fit as indicated by the MPC and ICI values. The absolute fit of the observables as measured by MPC or ICI varies

independently of the relative fit measured by the PPP-values. As in Row 1, Row 2 exhibits perfect agreement between the two observables in the realized and posterior predicted data, so the indicators of fit are identical between Row 1 and Row 2 despite the values of the conditional probabilities taking on the opposite extreme of 0 instead of 1. In Row 3 the observables are in perfect disagreement, which is reflected by MPC and ICI values taking on the extreme opposite values with respect to Rows 1 and 2 yet maintaining perfect fit with respect to the PPP-values due to the match between realized and posterior predicted response patterns. Rows 4-6 follow the patterns of Rows 1-3 but use conditional probability values that governed data generation within the current study.

Rows 7-11 of Table 31 illustrate that PPP-values will be greater than .5 to the extent that posterior predicted ICI values exceed realized ICI values, which is to say that there is greater inter-observable agreement in the posterior predicted data than in the realized data (i.e. smaller discrepancy between the conditional probabilities of success). Conversely, Rows 12-16 illustrate that PPP-values will be less than .5 to the extent that realized ICI values exceed posterior predicted ICI values, which occurs in this example when there is greater inter-observable agreement in the realized data than in the posterior predicted data.

Figure 39 illustrates how inter-observable agreement (match) and disagreement (mismatch) vary as a function of the conditional probabilities of a correct response for two given observables.

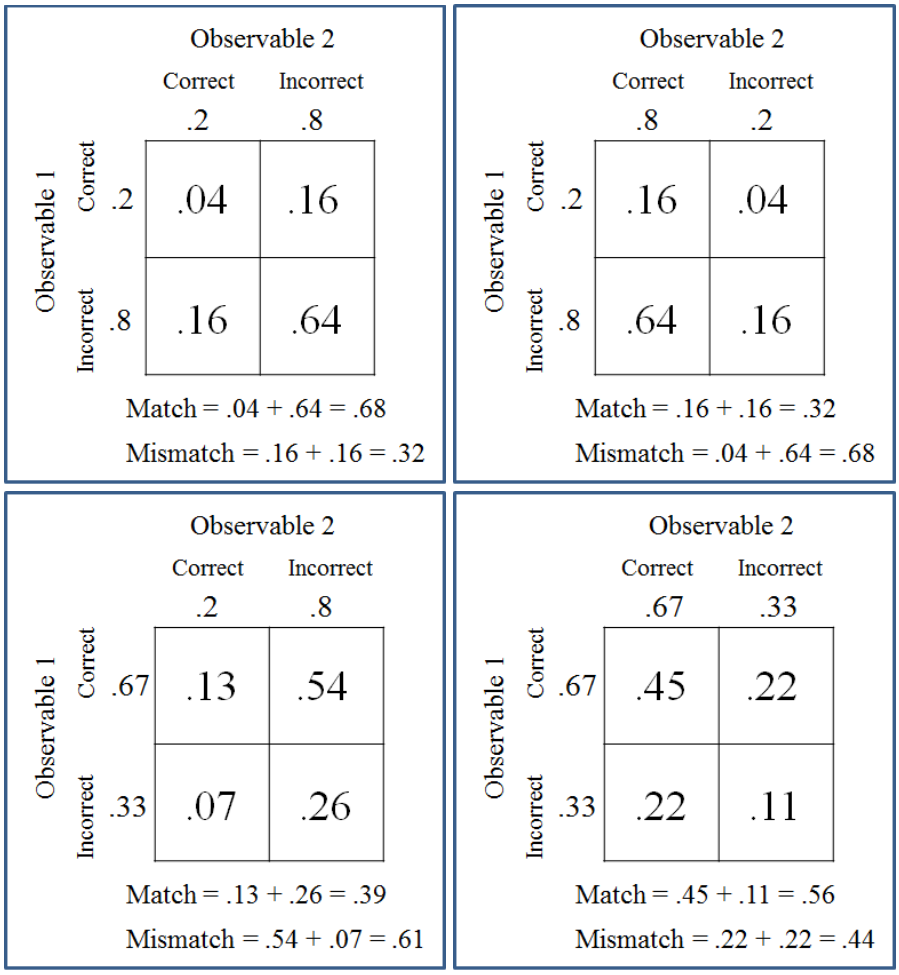



*Figure 39.* Examples of inter-observable agreement (match) and disagreement (mismatch) as a function of the conditional probabilities of a correct response.

Within the context of this simplified example, the computations illustrated in this figure are equivalent to the MPC computations shown previously in Table 31 because there is only one inter-observable comparison (per simulee). For example, consider the upper left panel of Figure 39, which illustrates the inter-observable agreement and disagreement that would be expected between two observables with conditional probabilities of success of .2. This corresponds also to the realized data in the fifth row of the simplified example shown in Table 31. Assuming the responses to each observable to be independent within the context of the model (i.e. after conditioning upon the latent variable that parents both

observables), a probabilistic representation of the four possible outcomes is presented. In approximately 4% of the outcomes, simulees would respond correctly to both observables. In approximately 16% of the outcomes, simulees would respond correctly to Observable 1 but incorrectly to Observable 2, and in another 16% percent of the outcomes simulees would respond incorrectly to Observable 1 but correctly to Observable 2. In approximately 64% of the outcomes, simulees would respond incorrectly to both observables. In total, 68% of the responses are matches (agreement between observables) and 32% are mismatches (disagreement between observables). By comparison, the proportions of agreement and disagreement in the upper right panel of Figure 39 switch with respect to the upper left panel because the conditional probabilities of success and failure for Observable 2 are inverted. In this case, there is a 16% chance that any given simulee will answer both observables correct, a 4% chance that Observable 1 will be answered correctly but Observable 2 incorrectly, a 64% chance that Observable 1 will be answered incorrectly but Observable 2 correctly, and a 16% chance that both observables will be answered incorrectly. The lower panels provide additional examples using alternative conditional probability values. The values in all four panels were selected for continuity with values in Figure 22, which becomes useful when applying the principles illustrated here back to the more complex case of the present study. MPC, ICI, and HCI computations in the present study can be thought of as aggregations of computations like those illustrated in Figure 39.

In the more complex case of the present study, decisions about which inter-observable comparisons are included in a given computation are based upon logical rules implied by the relationship between a given *Q*-matrix and scoring model. One reason for

using Condition 2.1 as the illustrative example for understanding the mechanisms underlying the performance of HCI and ICI is that the comparison-inclusion rules implied by the *Q*-matrix are more easily represented in a figure than they would be for conditions like 4.1 or 6.1 where the comparison rules are complicated by multiple parents for some observables.

Figure 38, Table 31, and Figure 40 were provided to help reconcile the apparent inconsistency of ICI across conditions and observables (seen in Figures 12 and 35-37) using Condition 2.1 as an illustrative example. The main point of these was to emphasize that ICI values were aggregations across different types of simulees having different degrees of misfit and who were disproportionately represented. Each proficiency profile represented in Figure 22 was itself an aggregation across individual simulees with varying degrees of misfit, but each profile had a typical level of misfit that was implied by the pattern of conditional probabilities of its members. Decomposing such high-level aggregations into constituent parts made it possible to see that when underlying factors were accounted for, ICI performed consistently after all.

## Discussion

### Discrepancy Measures

Consistent with previous PPMC research, all investigated fit functions tended to perform conservatively, but SGDDM, $Q_3$, and HCI only mildly so. Adequate power to detect at least some types of misfit was demonstrated by SGDDM, $Q_3$, HCI, ICI, and to a lesser extent Deviance, while PC, $\chi^2$, RPS, and GLS were powerless across all investigated factors. Bivariate SGDDM and $Q_3$ were extremely similar in their effectiveness and in terms of computation time. This study therefore offered no basis for

choosing one over the other. Each could be recommended as a useful member of the PPMC toolkit. However, their apparent redundancy suggests that using either is preferable to using both.

The observed power of SGDDM as a global measure was 1.00 in all misspecified conditions except Condition 5.1, the crossloadings misspecification, where observed power was .93. This finding is consistent with findings in Levy and Svetina (2011), which found that GDDM performed a bit better at detecting what amount to extra variables than it did at crossloadings, holding other things roughly constant. In the present study, part of the reason for this relative underperformance may have been due to design elements within this specific crossloadings misspecification. It is possible that alternative crossloadings misspecifications would have yielded more or less observed power in terms of extreme global SGDDM PPP-values.

Building upon findings from Levy, Mislevy, and Sinharay (2009) that bivariate fit functions were useful for detecting multidimensionality, the bivariate fit functions in this study (SGDDM and $Q_3$) were found to provide powerful and detailed feedback for all investigated types of misfit. The differential effects by CPT table highlighted in the results section demonstrated the effectiveness of the bivariate fit functions for detecting systematic differences in the conditional probabilities of successfully completing observables between observed and model-implied data across different combinations of complex multidimensional BNs.

The heat maps of median PPP-values for the bivariate SGDDM for the five misspecified conditions in this study each reflected different patterns of positive and negative local dependence, while holding constant the scoring model. This finding is

useful for understanding that both positive and negative local dependence can be caused by a single underlying misspecification, and that different patterns of such dependencies may suggest clues as to the identity of the misspecification. Observed patterns of positive and negative local dependence mimicked those in similar studies in IRT (Levy et al., 2009) including those with conjunctive effects (Levy, 2011).

For unidimensional IRT models, Habing and Roussos (2003) proved that positive and negative dependencies are always balanced because the data constitute a closed system. Recent work has suggested that the same principles would apply for multidimensional IRT models (Levy & Svetina, 2011) and BNs (Levy et. al, 2011), but in these contexts the speculation is yet unproven. The present study argues for the position that positive dependencies in one locality indicate the existence of negative dependencies somewhere else. In practice, one seeks to identify an interpretation that is consistent with the entire pattern of positive and negative local dependence. Given the complexity of such patterns, and the limitations associated with categorized representations, it may not be clear how a coherent cause could manifest both types of local dependence. Sometimes a theoretically grounded explanation may only be apparent for part of the observed pattern. In the author's previous experience, resolving one type of misfit (over or under predictions) tends to resolve both. Positive local dependence among some observables and negative local dependence among other observables can be jointly caused by the same source of misfit, so implementing model modifications consistent with theoretical moorings for the most prominent pattern of misfit may resolve less clearly understood local dependencies as a byproduct. The reported results suggest that specific

interpretations could potentially be identified in practice by diagnosing observed patterns of positive and negative local dependence relative to simulated results.

The breadth of effectiveness of the bivariate fit functions to detect a variety of misspecifications could also create ambiguity when generalized to the variety and complexity of misspecifications that exist with real data. It is likely to be much more difficult in practice to diagnose a misspecification based solely on the patterns of positive and negative local dependence provided by bivariate fit functions. The bivariate heat maps provided examples of differential patterns across the small number of misspecifications investigated here, but it is unknown whether such examples will become more or less ambiguous with future research. For example, in this study the contextual variable misspecification affected three observables per contextual variable, and produced flags for each intra-cluster pairing within the contextual grouping. Meanwhile, the partial mastery misspecification produced flags for each intra-cluster pairing of observables that were relevant to the definition of partial mastery, which spanned all three primary latent variables. If the contextual latent variables had represented the same observables as the definitions of partial mastery represented, then distinguishing between these two types of misspecifications may or may not have been possible. Therefore, observing a cluster of flags representing all the intra-cluster pairings of a set of observables in practice could represent either type of misspecification (or potentially other types of non-investigated types of misspecifications). The cross-loadings misspecification produced a similar but weaker pattern, in that only a subset of the intra-cluster pairings of misspecified observables were flagged. In practice this type of

misspecification could also become entangled with the others when attempting to interpret complex patterns of results under varying conditions.

A strategy to help reduce the ambiguity that may enshroud interpretation of bivariate heat maps is to include fit functions whose effectiveness is more limited with regard to misspecifications, creating roles for them as diagnostic specialists (see comments regarding ICI below). While the explication of such roles will require future research, the present study suggests that simulation studies devoted to this purpose could assemble a group of discrepancy measures to guide the process of attributing specific misspecifications to observed patterns of misfit.

The finding by Williamson, Mislevy, and Almond (2000) that GLS was useful for detecting errors associated with the number of latent classes was not replicated under the investigated conditions. In the present study, ICI was better suited for detecting latent class misspecifications than was GLS, but ICI showed reduced power for detecting the other types of investigated misspecifications. The narrower utility of ICI relative to SGDDM may enhance diagnostic potential when both functions are used in conjunction for model modification purposes. For example, when SGDDM flags a cluster of variable pairs, many alternative causal misspecifications may be possible. If ICI does not flag a variable that is implicated by SGDDM, then an additional latent class may not be the best modification to make, whereas it might be if both functions do implicate the variable in question.

**Effect Size**

An effect size measure for PPMC was introduced for the primary purpose of making distinctions between the fit of equal (or nearly equal) PPP-values. This purpose

applies to PPP-values within a replication, which corresponds to the results that an applied researcher would yield with real data, as well as to PPP-values aggregated across replications as reported in a simulation study such as this one. For example, in Figure 12 the panels representing Conditions 3.1 and 4.1 are both summarized by a PPP-value of .00. Comparison of the two scatterplots makes it evident that the differences between realized and posterior predicted SGDDM global values tended to be larger in Condition 4.1 than in Condition 3.1, but this information is not contained in the PPP-values, and even in graphical form interpreting these differences could become subjective owing to perceptual differences among people. The median effect size for Condition 4.1 was 15.06, while for Condition 3.1 it was 9.41. In this example, relying on the PPP-values alone would be to essentially equate the degree of misfit between the conditions by omitting information that distinguishes them. On the other hand, using an effect size alone would also omit information, as it is possible to obtain equal effect sizes even when PPP-values are opposites. The effect size is thus meant to supplement not to supplant the PPP-value.

The effect size measure also showed some utility for comparing the aggregated misfit of conditions with similar (or censored) values on the other outcome measures. The "proportion flagged" and "median PPP-value" outcomes were more susceptible to floor and ceiling effects due to their metrics. The effect size measure made it possible to differentiate results that were artificially equated due to the boundaries of those outcomes. An example of this was described in the results for SGDDM global.

The fit functions that were effective according to the PPP-values outcomes tended to exhibit larger effect sizes in misspecified conditions than in null conditions, while fit

functions that were ineffective in terms of PPP-values tended to exhibit minimal effect sizes across misspecified and null conditions. In summary, the effect sizes were largely consistent with the other outcomes, but helped to provide additional evidence for distinguishing the effectiveness across fit functions. Effect size is an alternative numerical summary to the PPP-value, both of which may be useful abbreviations of the complex patterns expressed more fully in graphical form. Neither numerical summary is an end in itself.

The ranges of observed effect sizes varied widely across fit functions in this study, which suggests that some fit functions may be much more sensitive than others to minor misspecifications. For example, the most modest of the investigated misspecifications was represented in Condition 5.1, which consisted of 12 crossloadings in the generating model that were not present in the scoring model. The median effect size across replications for deviance in this condition was 0.11, while for SGDDM global it was 3.60. Both of these fit functions operated at the global level, and the disparity between their effect sizes was not due just to the crossloadings misspecification, as evidenced by the fact that the disparities between these functions for the other misspecified conditions were even larger. In addition, the disparities between the median effect sizes of these fit functions were much smaller across null conditions (the disparity was as small as 0.01 in Condition 3.3). The magnitudes of the observed SGGDM effect sizes suggest that SGDDM could potentially be sensitive enough to detect misspecifications that consisted of fewer crossloadings, while the effect sizes for deviance in Condition 5.1 were barely larger than the effect sizes from null conditions, suggesting that less severe misspecifications may not be detectable. Future research is

needed to determine whether these speculations are accurate. The largest median effect size for a null condition in this study was -0.28 (Condition 5.5 for SGDDM subscale $\theta_1$), which suggests that modest effect sizes can be achieved by sampling variability alone and should therefore not be interpreted as theoretically meaningful. Much more research is needed to better understand the properties of the introduced effect measure, and to consider alternative effect measures.

**Computing Time**

Thus far the fit functions have been discussed mostly in terms of their effectiveness at detecting misfit, irrespective of their efficiency in terms of computation time. In applied studies, none of the investigated fit functions would likely be prohibitively time consuming to include because they operated on the order of minutes. In the present study, computation was conducted on a number of machines simultaneously, with machines varying according to their computational power. On the fastest machine, which was approximately four times as fast as the slower machines, the following fit functions took about one minute each to conduct PPMC per replication: Deviance, SGDDM, $Q_3$, PC, and $\chi^2$. The other functions took longer to compute (GLS ≈ 6 min, RPS ≈ 8 min, HCI ≈ 30 min, and ICI ≈ 35 min), due partly to the looping over simulees that was required for each of these, and for HCI and ICI due also to multiple conditioning statements within each loop over simulees. Presumably the computation times for these functions would decrease relative to the faster fit functions with smaller samples. It is also possible that more efficient programming could reduce these computation times. It should be emphasized that these times varied considerably even within this study, as they reflect a number of underlying influences, including differences

across conditions, the structure of the user-created R code, number of subjects, sample sizes, numbers and sizes of files read and written, number of MCMC chains, iterations, replications, etc. The issue of computation time is raised simply to illustrate the point that at present it is a legitimate practical consideration for many users or would-be users of PPMC. In this study, hundreds of computer hours were devoted to the simulation component. As a helpful tip, it was found that running multiple R sessions simultaneously on a given machine greatly improved the rate of completion, particularly on machines with multi-core processors. If a single R session had been used for this study, it would have taken about 50 days for the fastest available machine to complete just the simulation component (or about 200 days for the slowest), assuming uninterrupted 24-hour days. PPMC in WinBUGS and R may be overly time consuming when a researcher's goal is to select the best-fitting model among a number of competing alternatives, such as when a variety of modifications are possible based on PPMC feedback from an initial model. Programming, estimating, and analyzing phases can each take a number of hours or days depending on the circumstances. One possible approach in such situations is to use a graphical BN program such Netica or GeNie to more rapidly select among competing models based on loglikelihood values, then to critique the chosen model(s) in greater detail using PPMC procedures. Future research is needed to establish whether such a strategy would be effective.

**Recommendations**

For use in practice to critique the data model fit of multidimensional BNs using PPMC, the following recommendations are given regarding discrepancy measures. SGDDM (or $Q_3$) should be utilized at global and bivariate levels, and additionally at

subscale levels when applicable. At the global level, the measure is primarily useful for summarizing or ranking the misfit of comparable models. At the bivariate level valuable diagnostic feedback can be harvested but is potentially ambiguous, particularly without content expertise to help identify theoretically relevant patterns. HCI is recommended as a measure of person fit even in applications where person fit is not of central interest due to the alternative perspective that aggregation at the person level provides. Comparisons of realized response patterns to posterior predicted response patterns for flagged examinees can point to specific model inadequacies, and are recommended when fine-grained feedback is desired. HCI may be more useful for latent class misspecifications than for other types of misspecifications. ICI is recommended as a measure of observable (item) fit, and like HCI is also best suited for detecting latent class misspecifications, but it may also be useful for diagnosing other misspecifications when used in conjunction with SGDDM. If SGDDM indicates misfit but ICI does not, a latent class misspecification may be a less likely cause. HCI and ICI were designed for use in conjunctive models only, and are expected to perform poorly in fully compensatory structures. Models investigated in the present study had some conjunctive approximations and some compensatory elements. Alternative item-level and person-level discrepancy measures will likely need to be found for compensatory models. Deviance is not recommended per se, but is relatively easy to implement and may prove more capable of detecting types of misfit not investigated in this study. PC is recommended as a procedural check due to its computational ease and interpretational transparency. It is a convenient tool for verifying that PPMC computer code is functioning properly.

Regarding the examination and summarization of results, it is recommended that researchers use varying kinds of aggregation and presentation. Plots of realized versus posterior predicted values contain rich information that is not contained in PPP-values (or effect sizes) alone. In situations where graphical displays are impractical due to volume, PPP-values and effect sizes can be used together with graphical samples to summarize information. PPP-values are not recommended for strictly dichotomous decision rules akin to hypothesis testing. In situations where decision rules are implemented for convenience (e.g. heat maps), alternative decision criteria should be explored to see how interpretations might change.

**Limitations**

The present study helped to answer ongoing questions about the usefulness of PPMC for detecting data model misfit in BNs, but many questions were raised as well. While some useful discrepancy measures have been identified, there is no limit to the number that could be investigated due to the flexibility of PPMC. Similarly, the models investigated in this study mimicked models implemented in an applied research program, but limitless opportunities exist for alternative model structures and misspecifications. Features held constant in the present study, such as the strength of the contextual effects across latent variables, could be manipulated systematically within a separate investigation. Similarly, features that varied in the present study, such as the number of observed variables per primary latent variable, could be held constant in alternative studies to better isolate other factors of interest. A question raised in the results section for the bivariate fit functions is whether alternatively constructed misspecifications could produce matching patterns of bivariate data model misfit. It was beyond the scope of this

study to systematically investigate the partial label switching phenomenon encountered herein, but the options for handling this type of label switching could be explored in simulation studies devoted to the purpose of comparing alternative methodologies. The effect size introduced in this study was essentially a standardized difference score between realized and posterior predicted values, akin to a Cohen's *d*, but many alternatives are possible. In addition to the future research needed to better understand the performance of the introduced effect measure, alternative effect measures could be compared.

The present study began with the intention of comparing the effectiveness of the PPMC framework to the alternative frameworks discussed in the appendix. The scope of that initial design was reduced (thankfully) to a focus within the PPMC framework exclusively, but questions remain about when alternative frameworks might offer advantages over PPMC. These advantages are presumed primarily to consist of computational advantages (i.e. heuristic techniques may offer results that approximate PPMC results in less time), though other advantages are possible as well. Future research is needed to clarify the advantages and disadvantages of using statistics heuristically versus committing to a framework that estimates the reference distributions empirically. Within frameworks that estimate empirical reference distributions (i.e. PPMC vs. PB), future research is needed to compare the similarity of results between these conceptually similar but philosophically divergent methods.

References

Agresti, A. (2002). *Categorical data analysis, 2nd ed.* New York: Wiley.

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

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*, 341-359.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23,* 223-238.

Almond, R. G., Mislevy, R. J., Williamson, D., & Yan, D. (2012). Bayesian networks in educational assessment tutorial (Unpublished work). Retrieved from http://ecd.ralmond.net/tutorial/bninea-handout.pdf

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*, 491-521.

Armstrong, R. D., & Shi, M. (2009). Model-free CUSUM methods for person fit. *Journal of Educational Measurement, 46*, 408-428.

Bayes, T. (1763). A letter from the late reverend Mr. Thomas Bayes, F. R. S. to John Canton, M. and F. R. S. Philosophical Transactions, 53, 269-271. Retrieved from http://rstl.royalsocietypublishing.org/content/53/269.full.pdf+html

Berkhof, J., van Mechelen, I., & Gelman, A. (2004). Enhancing the performance of a posterior predictive check (Tech. Report No. 0350). Louvain-la-Neuve, Belgium: IAP Statistics Network.

Bottcher, S. G., & Dethlefsen, C. (2012). DEAL: Learning Bayesian networks with mixed variables (Version 1.2-35) [R package]. Retrieved from http://www.r-project.org/

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*, 791-799.

Briggs, D. C., Alonzo, A. C., Schwab, S., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33-63.

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

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Chung, H., Loken, E., & Schafer, J. L. (2004). Difficulties in drawing inferences with finite-mixture models. *The American Statistician, 58,* 152-158.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: with applications in the social, behavioral, and health sciences.* Hoboken, NJ: John Wiley & Sons.

Congdon, P. (2003). *Applied Bayesian modeling*. New York: John Wiley.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46,* 429-449.

Decision Systems Laboratory. (2012). GeNie graphical network interface (version 2.0) [computer software]. University of Pittsburgh. http://genie.sis.pitt.edu/about.html#genie

De Morgan, A. (1837). Review of Laplace's theorie analytique des probabilites. (3rd Edition). *Dublin Review, 2, 3,* 338-354, 237-248.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380-396.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods, 10,* 101-119.

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

Fienberg, S. E. (2006). When did Bayesian inference become “Bayesian”?. *Bayesian Analysis, 1,* 1-40.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York: Chapman & Hall.

Gelman, A., Goegebeur, Y., Tuerlinckx, & van Mechelen, I. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Applied Statistics, 49,* 247-268.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6,* 733-807.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66,* 8-38.

Glas, C. A. W., & Meijer, R. R. (2003). A bayesian approach to person fit analysis in item response theory models. *Applied Psychological Measurement, 27,* 217-233.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, B, 14,* 104-114.

Gorin, J. S. (2009). Diagnostic classification models: are they necessary? Commentary on Rupp and Templin (2008). *Measurement, 7,* 31-33.

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

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika, 68,* 435-451.

Hjort, N., Dahl, F. A., & Steinbakk, G. H. (2006). Post-processing posterior predictive *p* values. *Journal of the American Statistical Association, 101,* 1157-1174.

Holland, P. W. (1985). On the study of differential item performance without IRT. *Proceedings of the 27th annual conference of the Military Testing Association* (Vol. I, pp. 282-287), San Diego.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika, 66,* 109-132.

Kahneman, D. (2011). *Thinking Fast and Slow.* New York: Farrar, Straus, & Giroux.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277-298.

Kline, R. B. (2005). Principles and practice of structural equation modeling (second edition). New York: The Guilford Press.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49,* 59-81.

Lai, H., Gierl, M. J., & Cui, Y. (April, 2012). *Item consistency index: An item-fit index for cognitive diagnostic assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC, Canada.

Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods & Research, 24,* 492-516.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41,* 205-237.

Levy, R. (2006). *Posterior predictive model checking for multidimensionality in Item Response Theory and Bayesian networks* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Levy, R. (2009). Evidentiary reasoning in diagnostic classification models. *Measurement, 7,* 36-41.

Levy, R. (2011). Posterior predictive model checking for conjunctive multidimensionality in item response theory. *Journal of Educational and Behavioral Statistics, 36,* 672-694.

Levy, R., Crawford, A. V., Fay, D. M., & Poole, K. L. (2011, April). Data-model fit assessment for Bayesian networks for simulation-based assessment. In R. J. Mislevy (Chair), *Modeling strategies in a complex automated performance assessment environment*. Symposium conducted at the annual meeting of the American Educational Research Association, New Orleans, LA.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4,* 333-369.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement, 33,* 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*, 208-232.

Levy, R., Xu, Y., Yel, N., & Svetina, D. (2012). A standardized generalized dimensionality discrepancy measure and a standardized model-based covariance for dimensionality assessment for multidimensional item response models. Unpublished manuscript.

Li, Y., Bolt, D. M., & Fu, J. (2006). A Comparison of Alternative Models for Testlets. *Applied Psychological Measurement, 30*, 3-21.

Li, F., Cohen, A. S., Kim, S. H., & Cho, S. J. (2009). Model Selection Methods for Mixture Dichotomous IRT Models. *Applied Psychological Measurement, 33*, 353-373.

Loken, E. (2004). Using latent class analysis to model temperament types. *Multivariate Behavioral Research, 39,* 625-652.

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49-57.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *AppliedPsychological Measurement, 25*, 107-135.

Meng, X. L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22,* 1142-1160.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), Cognitively diagnostic assessment (pp. 43-71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). Bayes nets in educational assessment: Where the numbers come from. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 437-446.

Mislevy, R. J., & Bock, R. D. (1986). *Bilog: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary research and perspective, 1*(1), 3-62.

Muthén, L.K. and Muthén, B.O. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nelder, J. A., & Wedderburn, W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society A, 135*, 370-384.

Norsys Software Corporation. (1995-2014). Netica (4.08) [Computer software]. Vancouver, BC, Canada: Norsys Software Corporation. Retrieved from http://www.norsys.com/index.html

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement. 24*, 50-64.

Pearl, J. (1988). *Probabilistic reasoning in intelligence systems: Networks of plausible inference.* San Mateo, CA: Morgan-Kaufmann.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer-Verlag.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361-372.

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

Rupp, A. A., Levy, R., DiCerbo, K. E., Sweet, S., Crawford, A. V., Calico, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4(1), 49-110*.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219-262.

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

SAS Institute Inc. (2002-2013). SAS/STAT software (version 9.1.3) [computer software]. Cary, NC. http://sas.com

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Shute, V. J., & Almond, R. G. (2008). You can't fatten a hog by weighing it – or can you? Evaluating an assessment for learning system called ACED. *International Journal of Artificial Intelligence in Education, 18*, 289-316.

Sijtsma, K. & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149-157.

Sinharay, S. (2003). *Practical applications of posterior predictive model checking for assessing fit of the common item response theory models* (Research Report RR-03-33). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-03-33-Sinharay.pdf

Sinharay, S. (2004). *Model diagnostics for Bayesian networks* (Research Report RR-04-17). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-04-17.pdf

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375-394.

Sinharay, S. (2006a). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*, 429-449.

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

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

Sinharay, S., Almond, R. G., & Yan, D. (2004). *Model checking for models with discrete proficiency variables in educational assessment* (Research Report RR-04-04). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-04-07.pdf

Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research and Perspectives, 7*, 46-49.

Sinharay, S., & Johnson, M. S. (2003). *Simulation studies applying posterior predictive*

*model checking for assessing fit of the common item response theory models* (Research Report RR-03-28). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-03-28-Sinharay.pdf

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior Predictive Assessment of Item Response Theory Models. *Applied Psychological Measurement, 30,* 298-321.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research, 45,* 553-573.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Lunn, D. (2007). *WinBUGS user manual: version 1.4.3*. Cambridge: MRC Biostatistics Unit. Retrieved from: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

Steedle, J. T. (2008). *Latent class analysis of diagnostic science assessment data using Bayesian networks* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

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

Tatsuoka, C. (2009). Diagnostic models as partially ordered sets. *Measurement, 7,* 49-53.

Tatsuoka, K. K. (2009). Cognitive assessment: An introduction to the rule space method. New York, NY: Routledge.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a monte carlo study. *Methods of Psychological Research Online, 2,* 30-48.

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

Weiss, R. E. (1996). *Bayesian model checking with applications to hierarchical models.* (Technical Report). Department of Biostatistics. University of California. Los Angeles.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., DiCerbo, K. E., Crawford, A. V., Choi, Y., Chapple, K., & Behrens, J. T. (2012). A Bayesian network approach to modeling learning environments. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science* (pp. 257-292). Boston, MA: Sense Publishers.

Williamson, D. M. (2000). Utility of model criticism indices for Bayesian inference networks in cognitive assessment (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.

Williamson, D. M., Almond, R. G., & Mislevy, R. J. (2000). Model criticism of Bayesian networks with latent variables. *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence,* 634-643.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105,* 276-289.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (Research Report RR-03-32). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-03-32-Yan.pdf

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

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

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Yuan, K. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research, 40,* 115-148.

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items.

# APPENDIX

## DESIGN SIMPLIFICATION AND ALTERNATIVE FRAMEWORKS

Pursuant to recommendations obtained from the dissertation committee during the proposal defense meeting, it was determined that the scope of the proposed project would be narrowed. Specifically, only the PPMC framework was to be investigated in the present study, leaving comparisons to NRD and HT frameworks for future investigation. This design simplification focused and prioritized the purpose of the study around understanding the utility of the fit functions within PPMC, as opposed to comparing the utility of different frameworks. The main purpose of the simplified design was therefore to describe the performance of the discrepancy measures under the proposed conditions. The principal outcome measure of the study remained unchanged: the proportion of replications in which misfit was indicated by extreme PPP-values for each fit function. Removal of the NRD and HT frameworks did not decrease the computational burden appreciably because most of the computational burden of the original design was due to PPMC. The following discussion of alternative frameworks may still be of interest for a reader who is considering this study within a much broader model-checking context.

**Alternative Frameworks**

Model criticism is accomplished by mathematical functions that highlight particular features of the data-model relationship. The output from a particular fit function can be interpreted in a number of different ways, even holding constant a given model and dataset. These different ways to provide contextual meaning to the output of fit functions are labeled here as alternative model-checking frameworks, and are organized into four categories based on what the outputted values from fit functions are referenced against: no reference distribution (NRD), hypothesis testing (HT), parametric bootstrapping (PB; a.k.a. resampling), and posterior predictive model checking (PPMC).

In the sections below, each model criticism framework is discussed in terms of how fit functions are contextualized. Briefly summarized, the NRD framework does not appeal to reference distributions, but rather directly evaluates the fit function values relative to values obtained from competing models or to values recommended from experts (based on some theoretical and/or empirical grounding). An HT framework appeals to an analytically derived sampling distribution of the fit function, which is the distribution of values the fit function would be expected to take if the same model were to be fit again and again under replicated conditions (i.e. repeated independent samples from the same population). PB appeals to a reference distribution of fit-function values empirically generated from a point estimate (frequentist solution) of the model parameters. PB and HT share frequentist philosophies but differ in the ways replications are defined. PPMC appeals to a reference distribution of fit-function values from a Bayesian posterior distribution, most often empirically generated, unless conjugacy allows the posterior predictive distribution to be obtained analytically. PPMC and PB are related conceptually but differ in their philosophical underpinnings and computational implementation.

Given a particular model and dataset, a fit function (e.g. $Q_3$) highlights some feature of the data-model relationship. In our running example, $Q_3$ serves as a check of the local independence assumption. A $Q_3$ value is computed for each pair of variables in the model, and those values can range from -1 to 1. The values of the $Q_3$ fit function are interpreted within the context of the given model and dataset. Are the observed $Q_3$ values consistent with what would be expected given this particular model? As will be discussed in greater detail in sections to follow, the four alternative model checking

frameworks each appeal to different reference points for interpretation of fit functions. The functions themselves are to some degree interchangeable, though there are some functions which are suited better or worse to certain frameworks.

**No reference distribution.** It is possible to interpret the value of a fit function solely within the context of the fit function scale itself, by comparing an observed value directly to another value of interest. In the NRD framework, reference values are often obtained from competing models (i.e. relative fit comparisons). Alternatively, reference values are obtained by an appeal to authoritative sources in the literature, whether they are methodological studies yielding recommended values, or applied studies similar to the study providing the observed values.

***Running Example using $Q_3$.*** The fit function $Q_3$ can be evaluated differently according to one's model-checking framework. In each framework, the mathematical function contained in Equation 4 is used to assess the residual associations among pairs of modeled variables. The question becomes how to form an evaluation of the magnitude of the $Q_3$ function with respect to some frame of reference.

When no distribution of reference values is used, the observed $Q_3$ value for each pair of variables is compared directly to an analogous value from a competing model, or to an *a priori* cutoff value, perhaps recommended by previous researchers in the domain or by methodologists. In IRT models, a cutoff value of .2 has been used for $Q_3$ (Chen & Thissen, 1997), meaning that values between .2 and -.2 indicate an acceptably low level of residual dependence, while values between .2 and 1 (or between -.2 and -1) indicate levels of residual dependence that are large enough to warrant concern about LI violation. For example, consider a researcher who observes a $Q_3$ value of .17 for a pair of variables.

Within the NRD framework, the observed value of $Q_3$ is less than the cutoff value of .2, indicating that the model exhibited an adequate level of fit in terms of the residual dependencies between this pair of observed variables.

Note that this is the only framework in which the observed $Q_3$ values are not measured against a reference distribution. The purpose of a reference distribution, which will be exemplified in the other frameworks, is to gauge the frequency with which an observed value would be expected, typically expressed as an interval consisting of values equal to and greater than the observed value under an assumed (null) model. The judgment one typically makes in such a framework is thus a normative (norm-referenced) judgment. Values are labeled as significant on the basis of their lower frequencies of occurrence in the population. Alpha-level values for judging statistical significance are relativistic; they are not anchored on the scale of the fit function per se, but are ranges of values that occupy a predetermined portion of the distribution. The lack of a reference distribution in the NRD framework is simpler in the sense that an observed value is directly compared to an existing value from another model, or from some authoritative source. The comparison value is the criterion of good fit, and the subsequent criterion-referenced judgment is an easy one: the lower $Q_3$ is better in an absolute sense.

**Hypothesis testing.** An HT framework compares the observed statistic (using a sample of data) to the distribution of the same statistic that would be observed upon repeated sampling of equally-sized data sets from the same population (i.e. the sampling distribution). The location of the observed statistic can then be expressed in terms of a *p*-value, which represents the proportion of the sampling distribution with values of the test

statistic that are more extreme than the observed value. An $\alpha$-value of .05 is the most conventional criterion of statistical significance used in psychological domains.

***Running Example using $Q_3$.*** In the context of HT, the observed $Q_3$ value for each pair of variables is measured against an analytically derived reference distribution. For a given pair of variables, the observed $Q_3$ value is interpreted as a member of a population of values that would be observed if the study were to be replicated an infinite number of times. The relative magnitude of the observed value in relation to this analytic derivation of population values (sampling distribution) provides the researcher with the context to judge the significance of the observation. Yen (1984) proposed that the mean of the sampling distribution for the $Q_3$ statistic in IRT models should be $1/(n\text{-}1)$, and the variance of a Fisher *r*- to *z*- transformation should be $1/(n\text{-}3)$. Chen and Thissen (1997) argued that those normal-theory assumptions only hold when the residuals being correlated by $Q_3$ follow a bivariate Gaussian distribution (which may not be the case for IRT or BN models). A preferred sampling distribution for $Q_3$ has yet to be established, which is problematic when working within this framework. The two frameworks discussed below circumvent this problem of needing analytic reference distributions for fit functions by generating appropriate reference distributions empirically.

**Parametric bootstrapping.** A technique related philosophically to HT---both frameworks stem from a frequentist origin---is PB, also called resampling (Efron, 1979; Efron & Tibshirani, 1993; Langeheine, Pannekoek, & Van de Pol, 1996; von Davier, 1997). Within the PB framework, reference distributions are built empirically using generated data. The generated data consist of multiple replications generated from the same set of model parameters (i.e. the "solution" from whatever estimation routine was

used). Depending on the fit functions being used, the generated data may be compared to the observed data at this stage, or the model may be re-estimated using each of the newly generated datasets to obtain resultant model parameters from each solution that can then be used to calculate fit functions (e.g. Templin & Henson, 2006). In either case, the statistics that comprise the reference distribution stem from the replicated datasets and serve as the empirical sampling distribution for the model fit statistics from the original dataset.

Stated more formally, let P represent the vector of proportions in the population which constitute the probabilities of all possible response patterns, and let p represent a sample from P. The sum of P (or of any p) for any BN is 1, but the number of possible response patterns for typical BNs is so large that the probability of individual response patterns is often infinitesimal. Let $\Theta$ represent the population model parameters, such that P is a function (F) of those parameters, $P = F(\Theta)$. The parametric bootstrap procedure begins with an estimate of the population model parameters ($\widehat{\Theta}_{obs}$) derived from an observed sample ($p_{obs}$). Let $\hat{p}$ represent the bootstrapped datasets (samples) which are then generated from the model:

$$F(\widehat{\Theta}_{obs}) \xrightarrow{\text{yields}} \hat{p}_1, \hat{p}_2, \dots \hat{p}_n \qquad \text{(A1)},$$

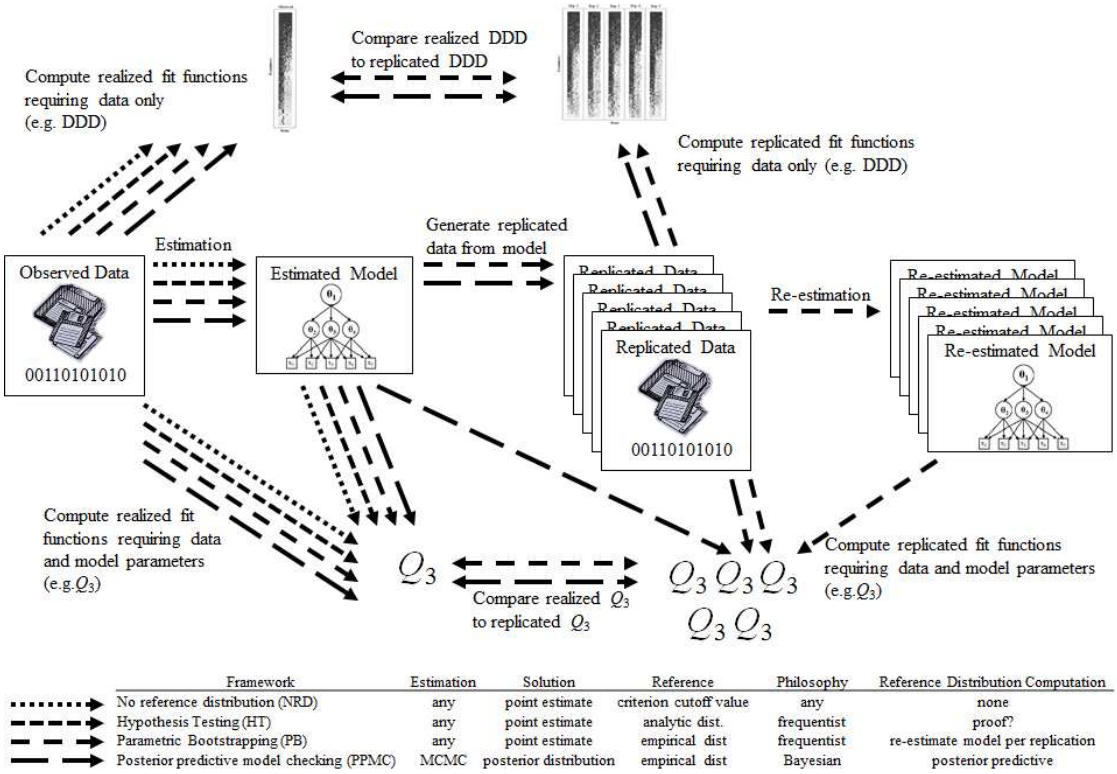
where *n* is the number of bootstrapped datasets. Features of the observed data ($p_{obs}$) can be compared to the bootstrapped data ($\hat{p}$) using functions that do not require model parameters, i.e. $T(p_{obs})$ compared to $T(\hat{p}_1), T(\hat{p}_2), \dots, T(\hat{p}_n)$, where *T* is a test statistic capturing a feature of the data. Or, for functions requiring model parameters, the model is re-estimated to yield $\widehat{\Theta}_1, \widehat{\Theta}_2, \dots, \widehat{\Theta}_n$, which can then be compared to $\widehat{\Theta}_{obs}$ , i.e.,

$D\left(\mathrm{p}_{\mathrm{obs}}, \widehat{\Theta}_{\mathrm{obs}}\right)$ compared to $D\left(\hat{\mathrm{p}}_1, \widehat{\Theta}_1\right), T(\hat{\mathrm{p}}_2, \widehat{\Theta}_2), \ldots, T(\hat{\mathrm{p}}_n, \widehat{\Theta}_n)$, where *D* denotes a function capturing data-model fit in terms of the discrepancy between the data and the parameters.

***Running Example using $Q_3$.*** In the context of PB, the observed $Q_3$ value for each pair of variables is measured against an empirically generated reference distribution. Using the model parameters estimated from re-fitting the model to the bootstrapped data, a $Q_3$ value for each pair of variables is calculated using each bootstrapped data set. If 500 bootstrapped data sets are used, then each set contains a $Q_3$ value for each pair of observed variables. For a given pair of observed variables, the 500 $Q_3$ values coming from 500 different bootstrapped datasets form the reference distribution for the single observed $Q_3$ value. The observed $Q_3$ value and the 500 replicated values are posited as members of the same population of values. The question is whether it is appropriate to consider the observed value as having come from the same population as the others. The relative magnitude of the observed value in relation to the distribution of empirically generated values provides the researcher with the context to judge the significance of the observation.

**Summary of alternative frameworks.** The purpose of a fit function is to highlight some feature of the data-model relationship. The output from a particular fit function can be interpreted in a number of different ways, even holding constant a given model and dataset. These different ways of providing contextual meaning to the outputted values from fit functions are labeled here as alternative model-checking frameworks, and are organized into four categories based on what the outputted values of the fit functions are referenced against: Posterior predictive model checking (PPMC), parametric

bootstrapping (PB; a.k.a. resampling), hypothesis testing (HT), and no reference distribution (NRD). Figure A1 summarizes the different characteristics and general procedures of these four frameworks.




| Framework | Estimation | Solution | Reference | Philosophy | Reference Distribution Computation |
|---|---|---|---|---|---|
| No reference distribution (NRD) | any | point estimate | criterion cutoff value | any | none |
| Hypothesis Testing (HT) | any | point estimate | analytic dist. | frequentist | proof? |
| Parametric Bootstrapping (PB) | any | point estimate | empirical dist | frequentist | re-estimate model per replication |
| Posterior predictive model checking (PPMC) | MCMC | posterior distribution | empirical dist | Bayesian | posterior predictive |

*Figure A1*. Comparison of four model-checking frameworks using Direct Data Display (DDD) and $Q_3$ as example fit functions.

The alternative model checking frameworks each appeal to different references for interpretation of fit functions. Briefly summarized, PPMC appeals to a reference distribution of fit-function values empirically generated from a Bayesian posterior distribution. PB appeals to a reference distribution of fit-function values empirically generated from a point estimate (frequentist solution) of the model parameters. HT appeals to an analytically derived sampling distribution of the fit function, which is the distribution of values the fit function would be expected to take if the same model were to

be fit again and again under replicated conditions. NRD makes direct comparisons without appeal to a distribution of reference values.

It was beyond the scope of the present study to extensively compare the various tradeoffs associated with these four frameworks. For example, one dimension the frameworks could be compared along is computational speed. Consider the running example using $Q_3$, a fit function that can be computed within any of the four frameworks. What are the various computational requirements associated with the frameworks as they each employ the same fit function? As described in the preceding paragraphs, each framework shares the computations in Equation 4 for each pair of observed variables in the data set of interest. The number of variable pairs is given by

$$J(J-1)/2 \qquad \text{(A2)},$$

where $J$ is the number of observed variables. In a dataset with 33 observed variables (the number of observed variables in the present study), there are 528 computations to perform in order to obtain the observed $Q_3$ values. This set of 528 computations would be executed under any of the frameworks. The NRD framework requires *only* these calculations, while the other three frameworks require additional calculations.

The HT framework requires the reference distribution to be analytically derived. In the case of $Q_3$, sampling distributions have been proposed for IRT models (Yen, 1984; Yen, 1993) with some debate regarding their accuracy (e.g. Chen & Thissen, 1997), but have not been thoroughly investigated in the context of BNs (Rupp, Templin, & Henson, 2010, Ch. 12). Setting aside the serious and often prohibitively difficult issue of obtaining a trusted sampling distribution, the computations required under this framework when a sampling distribution has been obtained are only slightly more than what is required

under the framework with no reference distribution. Each observed value is located within the sampling distribution, and typically is assigned a *p*-value corresponding to the area of the distribution occupied by values more extreme than the observed value. A set of 528 *p*-value calculations would thus represent the additional computations needed under the HT framework relative to the NRD framework.

The PPMC framework is considerably more intensive computationally than the HT or NRD frameworks because a reference distribution is built using replicated (generated) datasets. In addition to the computations required to generate the replicated datasets, the calculations required for the observed data (see Equation 4) are repeated using each replicated dataset as a substitute for the observed dataset. If 500 replicated datasets are generated, there are 528 * 500 = 264,000 $Q_3$ computations. The step of locating the observed values in relation to the reference values, which was carried out under the HT framework, can be applied to the PPMC framework as well, although these PPP-values should not be equated with a formal hypothesis test (Levy, 2011; Sinharay, 2006b).

The PB framework is the most computationally demanding in this $Q_3$ example. Setting aside any differences in the computational demands for estimation of a Bayesian solution (posterior distribution) relative to a frequentist solution for the same model (if the model can be estimated using frequentist techniques), the two frameworks differ when model parameters are required as inputs for the fit function (which is the case in this $Q_3$ example). Recall that for PPMC each generated dataset comes from a different set of model parameters, each representing a unique draw from the posterior distribution, while for PB the generated datasets all come from the same set of model parameters.

When bootstrapped model parameters are required for a fit function, each bootstrapped dataset must be re-estimated to obtain those bootstrapped model parameters. This additional estimation time is prohibitive for some applications. Then, for each bootstrapped dataset (typically numbering in the hundreds) and its associated model parameters, the calculations performed using the observed data are replicated (see Equation 4). If 500 bootstrapped datasets are generated, there will be 528 * 500 = 264,000 $Q_3$ computations. In the PB framework, each of the 528 observed $Q_3$ values would belong to a population of $Q_3$ values represented by the set of 500 bootstrapped values. A final step, as in the HT framework, is to calculate *p*-values to summarize the location of the observed values with respect to the reference values. Note that in situations where model parameters are *not* required as inputs for a fit function, the PB and PPMC frameworks converge in their methodology after the generated datasets are complete. The procedures for comparing observed and generated data when model parameters are irrelevant are quite similar for the two frameworks, and would be essentially equivalent in terms of post-estimation computational demands.

In summary, to compare the four frameworks on the dimension of computational burden using the $Q_3$ function, one must first consider the time needed for model fitting, in which it is often the case that PPMC takes longer due to the need to reach the full posterior distribution. Regarding the computational time after model fitting, the frameworks are loosely ranked, from fastest to slowest, in the following order: NRD, HT, PPMC, and PB. This ordering would be expected to hold for any fit function that relies on model parameters for its computation, as opposed to fit functions that require data only. For fit functions that do not require model parameters as input, the order of

computational difficulty could change with respect to PPMC and PB, depending on the amount of time required for model estimation. In PB, the bootstrapped parameters come from re-estimating the model using each of the bootstrapped datasets. In PB, the bootstrapped parameters come from re-estimating the model using each of the bootstrapped datasets. For relatively complex models, the time required to re-estimate a model solution for each bootstrapped dataset could exceed the time required to conduct PPMC, which relies on a single (albeit often slower) estimation routine to obtain the distributions of all model parameters included under the posterior distribution umbrella. Computational comparisons between PPMC and PB for different models and fit functions are of interest for future research.