UNIVERSITY OF CALIFORNIA

Los Angeles

An Integrative Framework of Model Evaluation

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Psychology

by

Wesley Earl Bonifay

2015

ABSTRACT OF THE DISSERTATION

An Integrative Framework of Model Evaluation

by

Wesley Earl Bonifay

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2015

Professor Li Cai, Co-chair

Professor Steven Reise, Co-chair

An important aspect of empirical research is the construction of a model that represents the data. In psychological and educational measurement, models are typically evaluated regarding their ability to fit well to the observed data. Philosophers of science have long recognized that goodness-of-fit to the realized data is an insufficient metric of a model's usefulness; models should also be appraised regarding their generalizability to unseen data. Frequentist statistics, Bayesian inference, and information theory seem to offer philosophically and methodologically dissimilar perspectives on model evaluation. However, this dissertation develops a simple theoretical framework that integrates these three perspectives. Within this framework, the information-theoretic principle of minimum description length is explored in the context of item response

theory modeling. The findings reveal that complexity in item response theory is defined not by the number of freely estimated parameters in a model, but by its functional form. The frequentist, Bayesian, and information-theoretic approaches are then utilized in evaluating the usefulness of a unidimensional 3-parameter logistic model of item response data from the Program for International Student Assessment. Philosophical ramifications, future research directions, and implications for educational and psychological measurement are discussed.

The dissertation of Wesley Earl Bonifay is approved.

James Stigler

Noreen Webb

Li Cai, Committee Co-chair

Steven Reise, Committee Co-chair

University of California, Los Angeles

2015

For Julie and Wren

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would also like to express my great appreciation and admiration for Dr. Steven Reise, my primary advisor at UCLA. Steve took a truly unwarranted chance when he accepted me as his student, and for that I am forever grateful. Over the past five years, Steve has been a constant source of guidance, advice, and hilarity, and above all else, he has instilled in me the confidence to grow as an independent researcher.

This dissertation simply would not exist without Dr. Li Cai. I am indebted to Li for seeing in me potential that I would never have seen on my own. He has been an extraordinary mentor—sharing ideas, counsel, and brilliant insights—despite the fact that I am not officially his graduate student. I have learned a tremendous amount from Li and it has been an honor to work with him.

Thank you also to the other members of my dissertation committee, Dr. Jim Stigler and Dr. Noreen Webb, for their invaluable feedback and criticism, and to Dr. Peter Bentler, for offering his expert opinion about various aspects of this work. I would like to thank Andrew Moskowitz for enduring the countless brainstorming sessions that culminated in many of the ideas presented in this dissertation.

I also wish to acknowledge the U.S. Department of Education, for financial support and training through Grant R305B080016 to the University of California, Los Angeles. The opinions expressed herein are my own and do not reflect the views or

policies of this funding agency.

Finally, I am grateful to my family. I would like to express special gratitude to my parents for fostering my intellectual curiosity and creativity and for providing unconditional love and support. Far above all else, I want to thank my wife, Julie. I absolutely would not have accomplished this without her patience, encouragement, emotional support, patience, advice, understanding, love, patience, faith, trust, reassurance, and patience. (And thanks to Wren for always telling me, "Get to work, Papa!")

VITA

EDUCATION

2004	B.A., Psychology, University of California, Davis
2012	M.A., Quantitative Psychology, University of California, Los Angeles

RESEARCH EXPERIENCE

2010 – 2014	Doctoral Fellow, Advanced Quantitative Methods in Education Research, University of California, Los Angeles
2014	Graduate Student Researcher, University of California, Los Angeles
2014	Summer Research Associate, RAND Corporation, Santa Monica, CA
2013 – 2015	Subcontractor, Educational Testing Service, Princeton, NJ

PUBLICATIONS

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95*(2), 129-140.

Bonifay, W. E. (2015). An illustration of the full-information two-tier item factor analysis model. In Reise, S. P. & Revicki, D. A. (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment.* Routledge: New York.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2015). Bifactor modeling and the evaluation of scale scores. In P. Irwing, T. Booth & D. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing.* London: John Wiley & Sons.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling.*

CHAPTER 1

Introduction

A key component of any scientific undertaking is the construction of a model that explains the data. However, no model is an exact representation of the phenomenon under investigation and, especially in the psychological and educational sciences, useful models are often simplistic approximations of immensely complex mental processes. It is necessary then to evaluate a given model, to judge its characteristics, investigate its nuances, and critique its flaws. Centuries of scientific reasoning have led to ostensibly different schools of thought regarding model evaluation. Yet, despite quarrels about philosophy and technique, each school has the same goal: models should be of use.

1.1 What defines a "useful" model?

Regarding statistical models, Cudeck and Henly (1991) advised, "It is a mistake to ignore either their limitations or their artificiality. The best that one can hope for is that some aspect of a model may be useful for description, prediction, or synthesis" (p. 512). Myung, Pitt, and Kim (2005) elaborated on this advice by proposing several criteria, both qualitative and quantitative, that should be considered when evaluating a psychological model. Among the qualitative criteria are the issues of explanatory

adequacy ("Are the model assumptions plausible and consistent with previous findings?"), interpretability ("Does the model make sense?"), and faithfulness ("Do the theoretical principles embodied in the model enable it to capture the underlying mental process?"). While such concerns are certainly important, they are highly subjective and cannot yet be quantified in a meaningful way. Myung, Pitt, and Kim (2005) therefore present four evaluation criteria that are quantifiable: falsifiability, goodness-of-fit, complexity, and generalizability.

Falsifiability is the possibility of an assertion, hypothesis, or theory to be disproven by some observation or experiment. As Popper (1962), the progenitor of scientific falsification, argued, "A theory which is not refutable by any conceivable event is non-scientific." In model evaluation, falsifiability means that the model under consideration will fail to describe certain patterns of observations. To paraphrase Popper's axiom, a model which is not refutable by any conceivable data is not useful. Despite its importance, however, falsifiability is not a central focus of this dissertation. It is assumed that the models to be discussed in later chapters are technically falsifiable (i.e., for all parameter values in the model, the rank of the Jacobian (partial derivatives) matrix is less than the number of data observations (Bamber and van Santen, 1985)). The following discussion focuses instead on the more pertinent notions of goodness-of-fit, complexity, and generalizability.

1.1.1 Goodness-of-fit

In psychological measurement, the most common method of model evaluation involves computing classical goodness-of-fit statistics. These measures are intended to quantify how "close" the observed values are to the values one would expect under the fitted model. A familiar goodness-of-fit statistic is the R^2 coefficient of determination that is routinely used in regression modeling. This fit measure is found by:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}, \tag{1.1}$$

where the numerator is the sum of squared residuals and the denominator is the total sum of squares. When the discrepancy between the observed data points and the fitted regression curve is minimal, then the numerator in the equation above is small and the R^2 value approaches 1. Such a result would suggest that the regression model has strong goodness-of-fit to the observed data. Although this example may be statistically rudimentary, it clearly demonstrates the concept of goodness-of-fit.

A traditional goodness-of-fit measure for categorical data is Pearson's chi-square test statistic, which is used to determine whether the observed values are consistent with a hypothesized distribution. The chi-square goodness-of-fit test is computed by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(\mathrm{o}_i - \mathrm{e}_i)^2}{\mathrm{e}_i}, \tag{1.2}$$

where o_i is the observed frequency for bin i, and e_i is the expected (i.e., theoretical) frequency for bin i, as asserted by the null hypothesis. The resulting χ^2 value is then

compared to a chi-square distribution in order to assess goodness-of-fit via an asymptotic *p*-value. Many other goodness-of-fit statistics are used in evaluating latent variable models. Some of the most frequently used include the comparative fit index, the Tucker-Lewis index, the Akaike and Bayesian information criteria, the standardized root mean squared residual, and the root mean square error of approximation, among many others (see Hu and Bentler (1999) for an extensive overview of common goodness-of-fit indices).

1.1.2 Complexity

Following the principle of Occam's razor, a model should not only fit the data well, but it should do so in the simplest manner possible. Myung, Pitt, and Kim (2005) define complexity as "a model's inherent flexibility that enables it to fit a wide range of data patterns" (p. 12). There are two factors that are known to influence complexity. The first factor is the number of freely estimated parameters in the model. Simply speaking, the greater the number of parameters, the more complex the model. The second contributor to complexity is the functional form of the model, that is, the way in which the parameters are combined in the model equation. It is important to note that two models with the same number of parameters but different functional forms may differ in complexity. For example, the models $y = x + b$ and $y = e^{xb}$ have the same number of parameters, but they certainly differ in complexity, such that the latter is likely to be

much better at fitting data (Myung, Pitt, & Kim, 2005).

Several methods have been introduced in order to quantify the complexity of a model. Akaike's information criteria (Akaike, 1974) and the Bayesian information criteria (Schwarz, 1978) are two well-known goodness-of-fit indices that are designed to penalize a model for being overly complex. Both of these measures, however, measure complexity by the number of parameters, irrespective of functional form. Less common metrics, such as the geometric complexity criterion (Pitt, Myung, & Zhang, 2002; Rissanen, 1996) and the effective number of parameters (Moody, 1992), take into account the model's functional form as well as the number of parameters.

1.1.3 Generalizability

While goodness-of-fit addresses the closeness of the model to the *observed* data, generalizability is the ability of a model to fit *future* or *unseen* data sampled from the same probability distribution. In psychological modeling, generalizability is a measure of how well a model will fit unseen data samples generated by the same underlying mental processes that produced the observed data. It is important to note that psychological data are necessarily contaminated by random, uncontrollable noise, caused by the cognitive processes underlying the data. Goodness-of-fit statistics, however, are unable to discern the meaningful signal from the intrusive noise, as shown in the following equation (Myung, Pitt, & Kim, 2005):

Goodness-of-fit = Fit to regularity (*generalizability*) + Fit to noise (*overfitting*). (1.3)

Model evaluation based solely on goodness-of-fit will give an overall sense of how close the model fits the data, but it will not provide any information about the separate terms on the right side of Equation 1.3. That is, a goodness-of-fit statistic will not distinguish between how well the model fits the meaningful trend in the data and how well it fits random noise. This issue is further complicated by the complexity of the model, which directly affects the second term on the right-hand side of Equation 1.3; a complex model with many parameters or a more pliable function form will be better able to fit unwanted noise.

The relationship between goodness-of-fit, generalizability, and model complexity is depicted in Figure 1.1. The two curves show that as a model becomes more complex, goodness-of-fit to the observed data and generalizability to future data both increase. At a certain point, however, a model becomes less and less generalizable. When the generalizability begins to wane, the model is said to be *overfitting* the data. Consider the example observed data presented in the bottom-right panel of Figure 1.1. If fit with the simple linear model on the left, generalizability would be adequate (though not stellar), but goodness-of-fit would be unimpressive. The complex model on the right would achieve excellent fit, but it would not generalize well to other data; by overemphasizing the observed data, such a model would serve little use in future samples or research scenarios. The quadratic model in the center panel, however, would provide optimal fit

and maximum generalizability. Note that "optimal fit" is not synonymous with "perfect fit;" rather, "optimal fit" defines the closest fit that can be obtained without sacrificing generalizability. The quadratic model, by deemphasizing exact fit to the observed data and concentrating instead on the underlying trend that one would expect in data yet to be observed, would therefore be appraised as the most useful choice. According to Myung, Pitt, and Kim (2005), generalizability, rather than goodness-of-fit, "should be the guiding principle in model evaluation and selection" (p. 14). Indeed, over 175 years ago, William Whewell (1840) declared, "It is a test of true theories not only to account for but to predict phenomena" (p. 256).

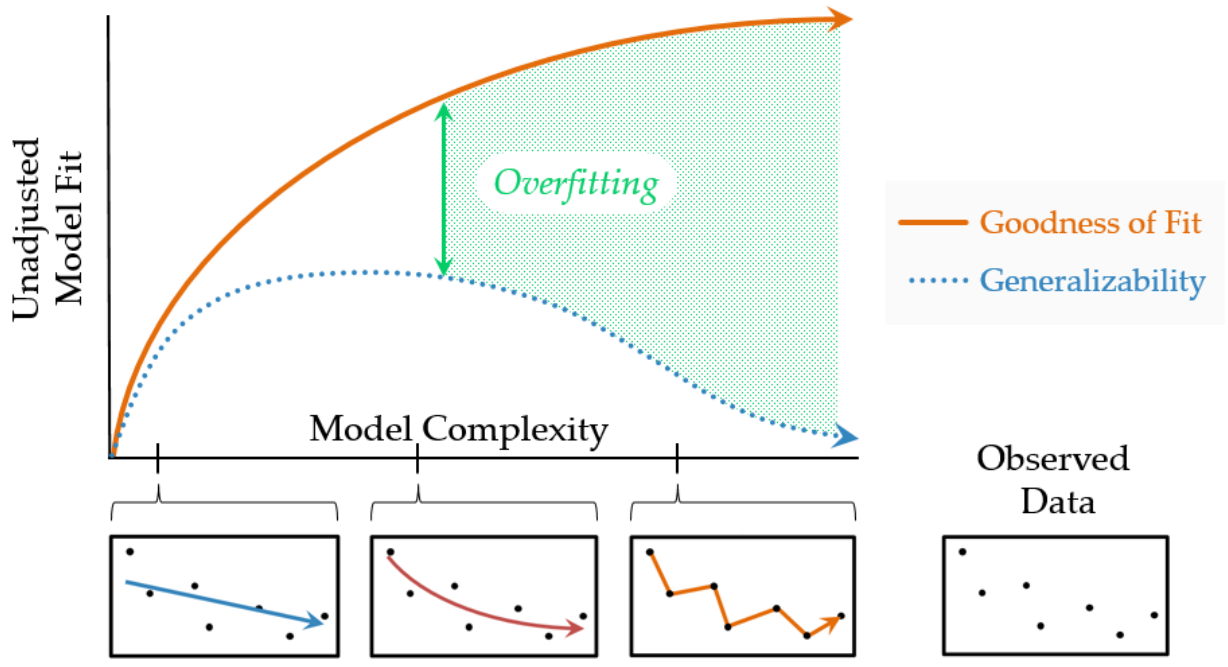


Figure 1.1. Goodness-of-fit and generalizability as a function of model complexity (adapted from Myung & Pitt, 2001; Preacher, 2006)).

1.2 Overview of the Dissertation

In sum, when evaluating a model, the aim should be to optimize usefulness by shifting the focus from goodness-of-fit to generalizability, from fitting models of disproportionate complexity to the observed data, toward fitting models of reasonable complexity to the not-yet-observed data. In this dissertation, I present three competing approaches to model evaluation that differ in their treatment of goodness-of-fit, complexity, and generalizability. In Chapter 2, I discuss the analytic strategies and underlying philosophy of the frequentist perspective (regarding classical goodness-of-fit statistics and the parametric bootstrap procedure), the Bayesian perspective (regarding Bayesian divergence measures and model checking methods), and the information-theoretic perspective (regarding the principle of minimum description length). In Chapter 3, I demonstrate how these three purportedly different approaches can be united in a single integrative framework. Chapters 4, 5, and 6 present a thorough analysis of model complexity in the context of item response theory. Finally, Chapter 7 demonstrates the utility of the framework with regard to empirical data from a large-scale educational assessment and Chapter 8 offers some closing remarks.

CHAPTER 2

Three Perspectives on Model Evaluation

Quantitative researchers have at their disposal a number of methods of model evaluation. These methods differ in the statistical techniques used to appraise the model and in the overall goals of the evaluation. The choice of method is typically guided by the philosophical bent of the researcher. To gain a better understanding of model evaluation, the following provides an overview of the two predominant philosophies of statistical inference, as well as a third perspective that may be less familiar to researchers in the social sciences.

2.1 The Frequentist Perspective

2.1.1 Frequentist Philosophy

Frequentist methodology has long been the predominant statistical approach within psychological measurement. Frequentist inference focuses on $p(D|H)$: the probability of the data D, given the hypothesis H. The data are assumed to be random, meaning that one would expect a replication study to produce a different set of data. The hypothesis is treated as fixed—it is either true or false, but the experimenter does not know which. This perspective is known as "frequentist" because the aim is to determine the *frequency*

with which one should expect to observe the data, given some hypothesis. The frequentist philosophy is perhaps best encapsulated by the concept of the confidence interval: a 95% confidence interval, for example, indicates that in 95% of repeated samples, the "true" value of the parameter under investigation will exist within a given range.

2.1.2 Frequentist Model Evaluation

Goodness-of-fit Statistics

Regarding model evaluation, frequentist researchers often rely on goodness-of-fit, prizing a model for how closely it represents the observed data. However, the use of goodness-of-fit statistics has become controversial. In structural equation modeling (SEM), many goodness-of-fit indices include cutoff values that are interpreted as thresholds for overall model fit (Hu & Bentler, 1999). Hayduk et al. (2007), among others, argue that such cutoff values are sometimes misleading and widely misused; Barrett (2007) even goes so far as to argue that goodness-of-fit indices do not contribute to an SEM analysis in any way.

Assessing goodness-of-fit is even more problematic when the data are categorical, as in item response theory (IRT) modeling. Consider the Pearson chi-square goodness-of-fit statistic shown in Equation 1.2. Discrete item response data can be arranged in an n-dimensional contingency table of the observed response pattern

probabilities. It is known that the asymptotic *p*-values of the chi-square statistic are only correct when the expected frequencies in each cell in the contingency table are large (> 5 is a general rule of thumb) (Maydeu-Olivares, 2013). Of course, the probabilities of each response pattern must sum to 1 (i.e., if, given the test taker's ability, the probability of responding correctly is .6, then the probability of responding incorrectly is 1 – .6 = .4) . Thus, as the number of possible response patterns increases, the expected frequencies become quite small and standard *p*-values cannot be used (Bartholomew & Tzamourani, 1999). To illustrate the scope of this problem, Maydeu-Olivares (2013) noted that when an item included 4 or more response categories, the classical chi-square *p*-values became inaccurate for tests of more than 5 items.

As a remedy to this issue, so-called "limited-information" fit statistics have been introduced to IRT modeling. These goodness-of-fit statistics are based on the lower-order margins of the contingency table, usually the univariate and bivariate proportions of correct response/endorsement. Examples of limited-information fit statistics include the M_2 (Maydeu-Olivares & Joe, 2005), R2 (Reiser, 1996), and Y2 (Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006) statistics (the last of which will be discussed in greater detail in Chapter 4). For an overview of limited-information fit assessment, see Maydeu-Olivares (2013).

Clearly, classical goodness-of-fit assessment leaves much to be desired. Fit statistics are misunderstood and misused (Hayduk et al., 2007), the cutoff criteria are

often arbitrary, and many indices are affected by issues such as sample size (Marsh, Balla, & McDonald, 1988), model complexity (Marsh & Balla, 1994), non-normality (Ory & Mokhtarian, 2010), and the number of variables in the model (Kenny & McCoach, 2003). Statistical concerns aside, there is an additional, perhaps more philosophical problem with evaluating a model by its goodness-of-fit: the model is judged according to how closely it fits the *particular* observed data. Even a model with perfect fit is not guaranteed to perform as well in replicated studies. As Jerzy Neyman stated, "Models become plausible by repetition" (cited in Cudeck & Henly, 1991). An alternative to goodness-of-fit-based appraisal is the parametric bootstrap, a model evaluation technique that puts Neyman's frequentist mantra into practice.

The parametric bootstrap

Efron (1979) introduced the bootstrap resampling method as a way to assign measures of accuracy to sample estimates. Rather than relying solely on the observed data, the general (*non-parametric*) bootstrapping procedure simulates new data by sampling from the observed values. This method allows one to investigate the sampling distribution of virtually any statistic. That is, bootstrapping could be used to obtain, for any test statistic, a confidence interval that is empirically derived from the observed data.

A central concept of the bootstrap is the "plug-in principle" (Efron & Tibshirani, 1993). Efron (2003) illustrated the plug-in principle with the diagram shown in Figure

2.1. Here, *P* denotes the unknown probability model that has yielded the observed data vector **x**, which is used to calculate a sample-based estimate $\widehat{\boldsymbol{\theta}}$ of the true parameter $\boldsymbol{\theta}$. In the "real world" of Figure 2.1, the accuracy with which $\widehat{\boldsymbol{\theta}}$ estimates $\boldsymbol{\theta}$ is quantified by confidence intervals, parameter bias, prediction error, and so on. The estimate of the probability model $\hat{P}$ returns the bootstrap data vectors in **x***, which are used to compute the bootstrap samples $\widehat{\boldsymbol{\theta}}^*$. Rather than determining confidence intervals and the like, the variability in $\widehat{\boldsymbol{\theta}}^*$ indicates the accuracy of $\widehat{\boldsymbol{\theta}}$. The large arrow in Figure 2.1 signifies the "plug-in" aspect of bootstrapping: to move from the "real world" to the "bootstrap world" simply requires one to plug in some estimate $\hat{P}$ of the unknown probability model *P* that produced the observed data **x** (Efron, 2003).

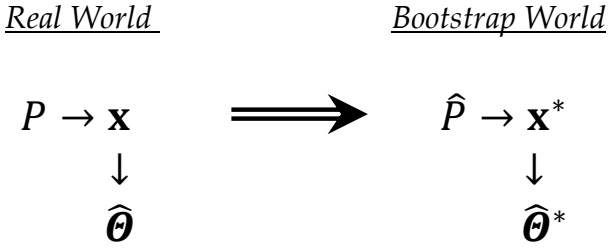


Figure 2.1. A diagram of the plug-in principle (Efron, 2003).

In the "bootstrap world," $\hat{P}$ is a point estimate of *P*, which can be obtained non-parametrically via sampling values from the observed data in **x**, or parametrically via "plugging in" some estimate of the unknown coefficients in *P*. The *parametric* bootstrap, also known as the Monte Carlo bootstrap, is a variant of the traditional non-parametric

bootstrap procedure; rather than resampling from the observed data, the parametric bootstrap involves drawing samples of random values from a fitted model (where each of the resamples is the same size). That is, the resampled data are generated from a parametric estimate of the population and not from the observed data itself. This allows for a different, perhaps more informative type of model appraisal than that provided by classical goodness-of-fit statistics.

The bootstrap method is especially useful regarding the sparse contingency table problem discussed earlier. The parametric bootstrap can be used when standard methods—based on asymptotic results—are not appropriate, as when data are categorical rather than continuous (Aitkin, Anderson, & Hinde, 1981; Collins, Fidlar, Wugalter, & Long, 1993). This process involves the follows steps (Tollenaar & Mooijaart, 2003):

1. The hypothesized model is fit to the data, yielding some traditional goodness-of-fit measure as well as the estimated cell probabilities in the contingency table.
2. A large number *B* of bootstrap samples of the same size as the original data are generated from a multivariate normal distribution with cell probabilities equal to the estimated probabilities from Step 1.
3. The hypothesized model is fit to teach bootstrap sample, resulting in *B* goodness-of-fit measures. The fit measures make up the (unknown) empirical

reference distribution.

4. The original sample goodness-of-fit statistic is then compared to the $(1-\alpha)^{\beta th}$ percentile of the ordered bootstrapped goodness-of-fit statistics. If the observed goodness-of-fit is smaller than the bootstrapped goodness-of-fit measure at this percentile, then the model is accepted; otherwise, the model is rejected.

Although this method of model evaluation has been successfully implemented (see e.g., Bartholomew & Tzamourani, 1999; van der Heijden, Hart, & Dessens, 1997; von Davier, 1997), it is not perfect. Tollenaar and Mooijaart (2003) identified several problems with the parametric bootstrap. First, the resampling process is computationally burdensome, especially with large models that involve many latent variables. Second, there is no guarantee that the estimation process has converged on a global maximum at each re-fitting iteration. Further, the parametric bootstrap displays very weak power in studies with small sample sizes (Tollenaar & Mooijaart, 2003).

2.1.3 Summary of the Frequentist Perspective

In summary, frequentist researchers tend to evaluate their models by citing classical goodness-of-fit statistics or by conducting parametric bootstrapping simulation. The former method is limited by misuse, by arbitrary cutoff values, and most importantly, by reliance on the observed data alone; the latter is limited in terms of computational

strain, estimation concerns, and dependence on the sample size (not to mention that the parametric bootstrap is seldom utilized as it is). Fortunately, the Bayesian perspective offers a number of alternative methods of model appraisal.

2.2 The Bayesian Perspective

2.2.1 Bayesian Philosophy

The Bayesian perspective presents a fundamentally different approach to statistical inference. Rather than focusing on $p(\mathrm{D}|\mathrm{H})$, as in frequentist statistics, Bayesian methodology considers $p(\mathrm{H}|\mathrm{D})$: the probability of the hypothesis H, given the data D. Here, the data are treated as fixed, meaning all inferences must be drawn from the observed values. The hypothesis is treated as random, in that it may or may not be true. This approach is termed "Bayesian" because $p(\mathrm{H}|\mathrm{D})$ is estimated using Bayes' Theorem (Bayes, 1764):

$$P(\mathrm{H}|\mathrm{D}) = \frac{P(\mathrm{D}|\mathrm{H})P(\mathrm{H})}{P(\mathrm{D})}, \tag{2.1}$$

where $p(\mathrm{H}|\mathrm{D})$ is known as the *posterior* probability of the data, given the hypothesis; $p(\mathrm{D}|\mathrm{H})$ is the *likelihood* of the hypothesis, given the data; $p(\mathrm{H})$ is the *prior* probability of the hypothesis; and $p(\mathrm{D})$ is the prior probability of the data. Equation 2.1 states that to make data-based inferences about some hypothesis, one must take into consideration the data itself, as well as any prior knowledge of or expectations about that data. Or, in

Bayesian lingo, the posterior is proportional to the likelihood times the prior.

2.2.2 Bayesian Model Evaluation

The Bayesian approach offers an appealing alternative to frequentist model evaluation. In certain scenarios, such as linear modeling, traditional goodness-of-fit tests may be useful and informative. Further, many fit statistics, such as the chi-square test shown in Equation 1.2, are easy to implement because their distribution is known (or can be approximated). Gelman, Meng, and Stern (1996) argue, however, that reliance on the classical approach is problematic in at least three types of models: models with severe restrictions on the parameters (e.g., positivity constraints); models that are restricted probabilistically due to the presence of a strong prior distribution; and unusual models that do not align to the general linear model. The Bayesian perspective offers several ways to address the shortcomings of frequentist model evaluation, including model divergence measures and model checking techniques.

Model divergence

The Bayesian philosophy, like the frequentist approach, concentrates on evaluation of a model relative to the observed data (though the Bayesian perspectives affords far more flexibility in this regard, as will be discussed later). The Kullback-Leibler (KL) divergence (Kullback & Leibler, 1951), for example, is a measure of the difference

between two probability distributions. In Bayesian statistics, the KL statistic is used as a measure of the discrepancy between the prior distribution and the posterior distribution. As evident in Bayes' Theorem (Equation 2.1), equivalence between the posterior and prior distributions indicates that the data corresponded with the experimenter's previous beliefs about that data. Thus, a model with low KL divergence is one that closely represents the observed data. A similar model evaluation method was developed by Dempster (1974), who formulated a measure of "Bayesian deviance" (Spiegelhalter, Best, Carlin, & van der Linde, 1998) based on the direct use of the posterior distribution of the loglikelihood of the data.

Like the classical goodness-of-fit statistics described earlier, the KL and loglikelihood divergence measures rely solely on the observed data. Further, these discrepancy measures appear to be most useful when some comparison is being made between the posterior densities of two or more models. When a single model is being evaluated, these divergence methods will only return the magnitude of the discrepancy between the posited model and the observed data; such values are difficult to interpret without a reference distribution of some kind. Examining a single posterior density may be useful when some benchmark of the "acceptable" discrepancy magnitude has been established, but this is often not the case. Thus, as Gelman, Meng, and Stern (1996) concluded, "It seems to us that in the context of assessing goodness-of-fit of a model for a given data set, hypothetical replications are inevitable" (p. 802). The "hypothetical

replications" mentioned here are the product of a Bayesian evaluation technique known as *model checking*.

Model checking

Rather than attempting to determine how closely a given model represents the data, model checking allows a researcher to answer the question, "Do the model's flaws have a noticeable effect on the substantive inferences?" (Gelman, Carlin, Stern, and Rubin, 1995). In Bayesian statistics, there are at least three ways to check a model. The first is to examine the sensitivity of inferences to changes in the prior $p(D)$ and the likelihood $p(D|H)$; the second is to check that inferences based on the posterior $p(H|D)$ are reasonable with regard to the substantive context; and the third is to check that the hypothesis (i.e., the model) H fits the data D (Gelman, Meng, & Stern, 1996). In this dissertation, I focus my attention on this third method.

The objective of model checking is to evaluate a model with respect to replicated data that could have been observed (or, in predictive terms, data that could be observed in the future). Model checking allows one to inspect the change that would occur if the experiment that produced today's data were repeated tomorrow with the same model and parameters. Although model checking is certainly a Bayesian technique, it is somewhat related to the frequentist philosophy and the parametric bootstrap in particular. Indeed, Gelman, Meng, and Stern (1996) note that "the posterior predictive

replication appears to be the replication that the classical approach intends to address" (p. 738). Despite this shared goal, Bayesian model checking differs from the parametric bootstrap in several ways, as discussed below.

Posterior predictive model checking (PMC), as introduced by Guttman (1967) and formally defined by Rubin (1984), involves drawing simulated values from the *posterior predictive distribution* of replicated data and comparing the samples with the observed data. The posterior predictive distribution is the probability of the replicated data $\mathrm{D}^{\mathrm{rep}}$, given the model H and the observed data D, as found by

$$p(\mathrm{D}^{\mathrm{rep}}|\mathrm{H},\mathrm{D}) = \int p(\mathrm{D}^{\mathrm{rep}}|\mathrm{H},\boldsymbol{\Theta})p(\boldsymbol{\Theta}|\mathrm{H},\mathrm{D})d\boldsymbol{\Theta}\,, \tag{2.2}$$

where $\boldsymbol{\Theta}$ is the vector of unknown model parameters that produced the observed data[1]. Equation 2.2 serves as the reference distribution by which one may evaluate the model. Having formulated a posterior distribution of $\boldsymbol{\Theta}$ based on the model and the observed data, a predictive distribution of $\boldsymbol{\Theta}$ can be used to "check" the model, as discussed below.

As an alternative to posterior PMC, Box (1980) introduced prior PMC, as expressed by the following:

$$p(\mathrm{D}^{\mathrm{rep}}|\mathrm{H},\mathrm{D}) = \int p(\mathrm{D}^{\mathrm{rep}}|\mathrm{H},\boldsymbol{\Theta})p(\boldsymbol{\Theta})d\boldsymbol{\Theta}\,, \tag{2.3}$$

where the only difference between the *prior predictive distribution* shown here and the

[1] Note that $\boldsymbol{\Theta}$ is presented in capital, bold typeface—this is to differentiate the unknown parameter vector $\boldsymbol{\Theta}$ from the latent trait ability θ that appears repeatedly in subsequent chapters on item response theory modeling.

posterior predictive distribution in Equation 2.2 is the final probability in the integrand ($p(\boldsymbol{\Theta})$ in prior PMC vs. $p(\boldsymbol{\Theta}|\text{H, D})$ in posterior PMC). That is, prior PMC does not stipulate that $\boldsymbol{\Theta}$ reflects the findings derived from the observed sample; indeed, prior PMC does not stipulate that the realized data be analyzed at all.

In posterior PMC, the data are replicated according to the same model H and *the same* vector of unknown parameters $\boldsymbol{\Theta}$ that produced the observed data D; the resulting D^{rep} samples form the (posterior predictive) reference distribution. In prior predictive model checking, the data are replicated according to the same model H but *different* vectors of unknown parameters $\boldsymbol{\theta}_n^{\text{rep}}$ for each replication; the resulting D^{rep} samples form the (prior predictive) reference distribution. Posterior and prior PMC differ in their assumptions. In posterior PMC, it is assumed that responses to future replications of an experiment will be guided by the same underlying phenomena (i.e., $\boldsymbol{\Theta}$) that caused people to respond in a certain way to the present study. In prior PMC, no such assumption is made; the responses to tomorrow's study may be motivated by entirely different phenomena.

Consider, for example, that a researcher plans to administer a test first in Classroom A and later in Classrooms B, C, … Z. The researcher's previously held belief is that test responses will be affected primarily by the temperature in each class. In posterior PMC, the researcher would assume that the test responses of Classrooms B, C, … Z will be influenced by the same predictors that affect Classroom A's scores. Suppose

that, after gathering and analyzing data from Classroom A, the researcher concludes that test responses were not significantly impacted by temperature, and that several additional variables—student ability, teacher competence, test anxiety, response format, and so on—were far more important. The posterior predictive distribution would include this data-based information. In prior PMC, the researcher would assume that previous (data-free) beliefs about the effects of temperature would hold for every classroom (A – Z), despite evidence to the contrary. In fact, the prior predictive distribution would not include *any* information gleaned from the observed data, regardless of whether that data supported or contradicted the prior beliefs.

Clearly, posterior and prior PMC are designed to evaluate different qualities of a model. The posterior predictive distribution can be used to explore the usefulness of a model in analyses of future data that are somewhat similar to the observed data. In that sense, the Bayesian concept of posterior PMC is closely related to the frequentist bootstrapping procedure described earlier. Gelman (2004) acknowledged this similarity:

> For example, if $\boldsymbol{\Theta}$ is estimated by maximum likelihood, it might be convenient to sample D^{rep} from the distribution $p(D|\boldsymbol{\Theta})$, which we would view as an approximate posterior predictive distribution. (p. 759)

The prior predictive distribution, on the other hand, can be used to explore the usefulness of a model in the analysis of any future data, regardless of its convergence with or divergence from the observed data.

Test quantities

In both types of model checking, failings of the model are indicated by the presence of systematic differences between the simulated and observed data. To assess such differences, Bayesian researchers select statistical measures that represent certain aspects of the data that are deemed relevant to the topic under investigation. Such measures are referred to as *test quantities*. Specifically, in Bayesian inference, an observed test quantity $T(\mathrm{D})$ is tested against the same test quantity in the reference distribution $T(\mathrm{D}^{\mathrm{rep}})$. A posterior predictive *p*-value (*PPP* value) is then computed to quantify the likelihood of $T(\mathrm{D})$ in the reference distribution:

$$PPP = p\big(T(\mathrm{D}^{\mathrm{rep}}) \geq T(\mathrm{D})\big). \tag{2.4}$$

PPP values denote the similarity between the realized and predicted data, relative to the chosen test quantity. The presence of systematic differences between the observed and predictive values is indicated by $PPP \leq .05$ or $\geq .95$; *PPP* values near .50 indicate that there are no such differences (Stone & Zhu, 2015).

Test quantities can certainly be used in an omnibus sense, by evaluating the general goodness-of-fit of the whole model (Gelman, Carlin, Stern, & Rubin, 1995); however, one of the main advantages of model checking is the capability to study specific features of the data, rather than the overall goodness-of-fit. As Gelman, Meng, and Stern (1996) stated, "We know that virtually all models are wrong, and thus a more relevant focus is how the model fits in aspects that are important for our problems at

hand" (p. 757). Thus, test quantities are typically chosen by the researcher in order to assess some characteristic of the data that is not directly addressed by the probability model. For example, rather than considering the mean or variance, one may wish to explore the degree to which the rank ordering of the observed sample is the same in the replicated distribution as in the observed data. Similarly, it may be informative to know whether the minimum value in the observed data is an outlier relative to the simulated distribution. The potential to examine any feature of the data highlights the versatility of the model checking method.

2.2.3 Summary of the Bayesian Perspective

The Bayesian philosophy presents several useful tools for in-depth model evaluation. The flexibility of model checking methods and the use of tests quantities are extolled by Gelman, Meng, and Stern (1996): "Indeed, Bayesian inference is a powerful tool for learning about model defects, because we have the ability to examine, as a discrepancy measure, any function of data and parameters" (p. 758). However, the Bayesian approach is not ideal. The main drawback of Bayesian inference is the subjectivity of specifying the prior $p(H)$. Bayes' theory does not place any constraints on how one should set the prior, so the choice of prior may differ from person to person, depending on each person's previous beliefs about the phenomenon under investigation.

Consequently, the posterior $p(H|D)$, which is a function of the prior $p(H)$, can also differ from one researcher to the next. Thus, the model checking techniques discussed above are based on the subjective opinions that characterize the prior and posterior distributions. As Gelman, Meng, and Stern (1996) warn, "Predictions obtained under strong incorrect prior specifications may be quite far from the observed data" (p. 757). Thus, Bayesian model checking methods, like the classical goodness-of-fit and parametric bootstrap methods, are somewhat problematic.

2.3 The Information-theoretic Perspective

2.3.1 Information-theoretic Philosophy

A third, less well-known method of model evaluation is grounded in information theory. The goal in the information-theoretic approach to modeling is to *compress* the data as much as possible by identifying *regularities* (i.e., patterns or trends) in the data. The concept of data compression is perhaps best understood via the following example (borrowed from Grünwald, 2007). Consider a 10,000-digit binary sequence that follows the pattern

$$\{1000100010001000100010001000...1000100010001000100010001000\}. \quad (2.4)$$

Obviously, the simple regularity (i.e., pattern) in this data is the repetition of the sequence `1000`. Thus, there exists a straightforward and simple "rule" or "law" that describes the data pattern in (2.4). Now consider a 10,000-digit binary sequence in

which the data are purely random:

{00101101011110101000010101111...10111000101010110100111001110}. (2.5)

Here, the 1s and 0s are truly random, meaning no regularities exist in the data. There is no simple law that precisely explains the sequence of digits in (2.5). Finally, consider a sequence of 10,000 binary digits that does not appear to possess any simple regularities, but in fact, we know that there are four times as many 1s as 0s:

{01110110111111101111100111101...11011101110101110011111011111}. (2.6)

In this case, the regularity may not be discernible, because this sequence adheres to a statistical rule rather than a deterministic one.

The random pattern example in (2.5) includes no regularities, so it would be impossible to accurately predict whether a 1 or 0 would come next in the sequence. The values in (2.4) and (2.6), however, each contain a regularity that can be identified and used to predict subsequent values in the pattern. In example (2.4), it is certain that the next cluster of digits in the sequence will be "1000." In (2.6), the next digit is not known with absolute precision, but it is four times more likely to be a 1 than a 0. Thus, detection of the regularity in a data set leads to a better prediction of future data that is expected to behave according to the same rule.

Every regularity that exists in the data can be used to compress the data into a symbolic statement or description that is shorter than the data itself. That is, the laws that govern the data can be used to reproduce, to the very digit, the full data pattern

(without relying on any input). The first example discussed above can be described in simple English as, "The pattern "`1000`" 2,500 times." This description can be written more succinctly in a computer programming language. If one were to type "`rep(c(1,0,0,0),2500)`" and then press ENTER in the R statistical software program, for example, the 10,000-digit data set shown in example (2.4) would be fully reproduced without any errors. However, a random sequence such as that generated by coin tosses is considered *incompressible*—these data are truly random and thus cannot be represented by a short(er) description that explains the laws governing the data pattern. In fact, to reproduce in R the exact 10,000-digit pattern shown in example (2.5) would require writing out all 10,000 numerals verbatim; there are no regularities, and therefore no programming shortcuts to help reproduce the data (and, in fact, a reproduction in R would require the syntax "`print(c(`" before the sequence, a comma between every entry in the sequence, and two close parentheses at the end; thus, the code would be far longer than the data itself).

Sequence (2.6) presented above differs from the others in that it does not possess a simple deterministic regularity to simplify matters; yet it is in fact compressible. Instructing R to create a data vector with "four times as many `1`s as `0`s" will not recreate the full 10,000-digit sequence, but the known ratio of `1`s to `0`s is certainly a regularity that can be used to compress the data down to more manageable dimensions. Direct computation shows that of all 10,000-digit patterns, there are fewer than $2^{7,213}$ sequences

that satisfy the criterion of "four times as many `1`s as `0`s" (Grünwald, 2007). By comparing this with the total number of all possible binary sequences ($2^{10,000}$), we can compute the ratio between (a) the number of sequences wherein there are four times as many `1`s as `0`s, and (b) the total number of all possible sequences. This ratio reveals that extremely few patterns actually meet the criterion; specifically, the proportion is smaller than:

$$\frac{2^{7,213}}{2^{10,000}} = 2^{-2,787}. \tag{2.7}$$

In other words, of all possible data patterns, astronomically few follow the rule that describes the pattern of data in (2.6). A verbal description of the exact sequence that characterizes these data would read something like, "This sequence is one of the $2^{7,213}$ sequences of 10,000 digits in which there are four times as many `1`s as `0`s. If all of these sequences are listed in order, it is sequence number *i*." Identifying the regularity of "four times as many `1`s as `0`s" facilitated the compression of the data to just those patterns that obey this simple law.

Data compression (i.e., concise descriptions, terse R code) is at the heart of information theory. The length of the shortest programming code that prints the desired data sequence D and then halts is defined as the *Kolmogorov complexity* of the sequence (Kolmogorov, 1965). The lower the Kolmogorov complexity, the more regular/less random/simpler the sequence. Unfortunately, the Kolmogorov complexity cannot be directly computed—there is no program that can automatically provide the shortest

code that describes a given data sequence. Fortunately, 13 years after Kolmogorov posited his theory, Rissanen (1978) quantified the information-theoretic approach to model evaluation by introducing the principle of *minimum description length*.

2.3.2 The Minimum Description Length Principle

The minimum description length principle (MDL; Rissanen, 1978; Grünwald, 2007) is a method of inductive inference, based on the idea that data can be represented by a set of symbols—or *code*—that is shorter than the literal length of the data itself. That is, the data can be compressed. MDL states that the more regularities that exist in the data, the more it can be compressed. Conversely, the more one is able to compress the data, the more one can learn about the data (i.e., by understanding the regularities in the data).

The philosophy underlying the MDL principle can be divided into two main tenets. The first tenet is that the goal of inductive inference should be to "squeeze out as much regularity as possible" from the data. The main task is to separate structure (i.e., meaningful information) from noise (i.e., accidental information). The structural part of data sequence (2.6) is the presence of four times as many `1`s as `0`s; the noise is represented by the $2^{2,787}$ sequences that fail to meet that criterion. To correctly model the data, one must identify the structure and minimize the noise. Of course, noise is defined relative to the specific model under consideration. In information-theoretic terms, noise is represented as the residual number of bits (`1`s and `0`s) needed to encode the data after

the best model has been selected. In that sense, noise is not a random variable; it is a function of the selected model and the observed data.

The second tenet of MDL relates to the notion of a "true" distribution. According to Rissanen (1978), most methods of inductive inference are flawed because they assume that the true state of nature is represented by the selected model. Often, however, that is simply not the case. Thus, methods that presume to capture the "truth" are only clearly interpretable under assumptions that, in practice, are usually violated. MDL, on the other hand, relies solely on the data and does not make an assumption of some underlying "true" data-generating mechanism. As Grünwald (2007) noted, "The MDL philosophy is agnostic about whether any of the models under consideration is "true," or whether something like a "true distribution" even exists" (p. 29). Instead of capturing the truth, the MDL principle aims to find the best model that represents the data. As Rissanen (1989) wrote,

> We never want to make the false assumption that the observed data actually were generated by a distribution of some kind, say Gaussian, and then go on to analyze the consequences and make further deductions. Our deductions may be entertaining but quite irrelevant to the task at hand, namely, to learn useful properties from the data. (p. 14)

This echoes the sentiments of Bandler and Grinder (1979), who declared, "We have no idea about the 'real' nature of things … The function of modeling is to arrive at descriptions which are useful" (p. 7).

MDL is especially useful when choosing between competing models. If the

choice between candidate models is based simply on goodness-of-fit to the observed data, then there is a risk that the better fitting model will overfit the data. Consider the two models presented in Figure 2.2. In terms of the number of parameters, M_b is a more complex model than M_a. Although M_a may do a better job of identifying the underlying trend in the data, M_b will achieve better fit by capturing more of the random noise. In order to select the best model, a tradeoff is needed between goodness-of-fit and model complexity.

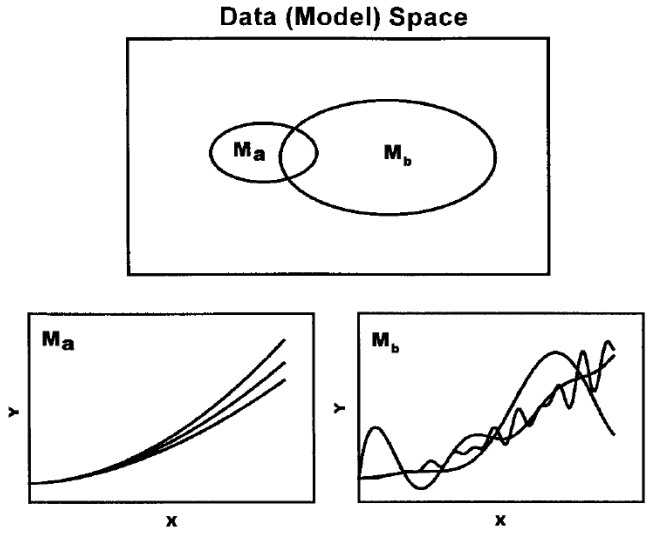


Figure 2.2. Regions in the data space occupied by two models, M_a (simple) and M_b (complex), and the range of data patterns that can be generated by each model (Pitt, Myung, & Zhang, 2002).

The two-part version of the MDL principle addresses this tradeoff directly. Let H_1, H_2, ..., H_n be a list of candidate models that each represent a different hypothesis about the data. In information-theoretic terms (where a *bit* is a binary digit—either a `0` or a `1`), the best hypothesis H to explain the data D is the one that minimizes the sum of

two parts: $L(\mathrm{H}) + L(\mathrm{D}|\mathrm{H})$, where $L(\mathrm{H})$ is the length, in bits, of the description of the hypothesis H, and $L(\mathrm{D}|\mathrm{H})$ is the length, in bits, of the description of the data once it has been encoded according to the hypothesis. In more common terminology, $L(\mathrm{H})$ represents the model and $L(\mathrm{D}|\mathrm{H})$ represents the goodness-of-fit of the model to the data. One can usually find a very complicated model (i.e., a model with large $L(\mathrm{H})$) to explain the data, and it may have excellent fit (i.e., small $L(\mathrm{D}|\mathrm{H})$). Alternately, one can find a simplistic model (small $L(\mathrm{H})$) that has very poor fit (large $L(\mathrm{D}|\mathrm{H})$. Under the MDL principle, the sum of these two parts will be minimized to arrive at a hypothesis/model that is relatively (but not overly) simple and has good (but not perfect) fit. In the early articles on MDL, Rissanen (1978, 1983) advocated choosing a minimax code that minimizes the shortest total description length $L(\mathrm{H}) + L(\mathrm{D}|\mathrm{H})$ over all possible data sequences.

Several expressions of the MDL principle have been developed. A formulation that perhaps best embodies the theory of MDL is given by the normalized maximum likelihood (NML; Rissanen, 2001). Let $\mathbb{D}$ be the complete data space. The NML is then given by:

$$\mathrm{NML} = \frac{p(\mathrm{D}|\widehat{\boldsymbol{\theta}}^*(D))}{\sum_{\mathbb{D}} p(\mathrm{D}|\widehat{\boldsymbol{\theta}}^*(D))}, \tag{2.8}$$

where D is the observed data, $\mathbb{D}$ is any possible data, and $\widehat{\boldsymbol{\theta}}^*(\cdot)$ denotes the maximum likelihood parameter values for a given data set. NML is an indicator of how well a model fits the particular observed data, relative to how well that model would fit any

possible data. This logic of this expression will become very important when discussing the design of the simulation study in Chapter 4.

In sum, the MDL principle, as expressed by the NML, enables one to evaluate a model by considering both goodness-of-fit and model complexity. Following the logic of the NML criterion, the complexity of model M_b in Figure 2.2 has imbued it with the potential to fit well to a greater range of data patterns, relative to the simpler model M_a. In Preacher's (2006) wording, M_b therefore demonstrates higher *fitting propensity*—a term that encapsulates much of the remainder of this dissertation.

2.4 Summary

This chapter has explored three perspectives on model evaluation. The contrasts between these approaches can be simplified by introducing a bit of notation. As in Equation 2.8, let $\mathbb{D}$ represent the complete data space of any and all possible data. The mathematical symbol $\in$ means "is an element of the set," such that $D \in \mathbb{D}$ means "the observed data is an element of the complete data space." In other words, the observed data can be viewed as just one particular instantiation among all possible data sets. The frequentist focuses then on $p(D \in \mathbb{D} | H)$, the probability of the *particular* observed data given the hypothesis; the Bayesian focuses on $p(H | D \in \mathbb{D})$, the probability of the hypothesis given the *particular* observed data; and the information theorist focuses on $p(H | \mathbb{D})$, the probability of the hypothesis given the *complete* data space. With regard to

model evaluation, the three approaches differ in their treatment of the data (whether observed, replicated, or representative of the complete data space) and their method of evaluation (whether refitting the model or computing a test quantity). The following chapter, however, presents an integrative framework that unites all three perspectives.

CHAPTER 3

An Integrative Framework

The model evaluation methods discussed in Chapter 2 appear to be dissimilar. The division between frequentist and Bayesian statistics has existed for many decades, at least since the work of Laplace and Poisson in the 18th century, and the information theoretic principle of minimum description length appears to be wholly foreign to social science research. Grünwald (2007) affirmed that MDL represents "a radical philosophy of learning and statistical inference that is considerably different from the ideas underlying mainstream statistics, both frequentist and Bayesian" (p. 29).

The current research proposes that these three approaches are not, in fact, so dissimilar. Table 3.1 presents a simple framework that unites the frequentist, Bayesian, and information-theoretic approaches to model evaluation, wherein the columns represent different treatments of the data and the rows address how the model is evaluated. More specifically, the columns indicate gradations of departure from the observed data, ranging from reliance solely on the observed values to the use of all possible data. The rows describe the method of appraisal, either by fitting (or re-fitting) the model or by computing a test quantity.

The first column, labeled "*Observed data*," includes methods wherein a model is

	Observed data	*Replicated data*	*All possible data*
(Re-)fit the model	Goodness-of-fit, Parametric bootstrap		Minimum description length
Compute a test quantity	Model divergence	Predictive model checking	

Table 3.1. *An integrative framework of model evaluation.*

evaluated with respect to the observed data only; that is, no data are generated beyond the values that have been provided by the sample. In frequentist inference, the most common method of model evaluation entails fitting the model and computing classical goodness-of-fit statistics. This method extends directly to the parametric bootstrap via the "plug-in principle" discussed Section 2.1.2. Both methods rely entirely on the observed values: goodness-of-fit assessment involves fitting the model to the observed data, while the parametric bootstrap involves re-fitting the model to many bootstrap samples derived directly from the observed data. These traditional model appraisal methods occupy the upper-left cell of Table 3.1. A Bayesian analog to these methods involves fitting a test quantity, such as the Kullback-Leibler divergence measure, to the prior and posterior distributions derived from the observed data; this technique is presented in the lower-left cell of the table.

The second column in Table 3.1 comprises methods wherein the model is evaluated with respect to replicated data. Posterior predictive model checking, as discussed earlier, examines a model by sampling from the Bayesian posterior and then computing a test quantity of some kind. Prior PMC involves sampling from the prior and computing a test quantity. Further, there exist several related techniques that have not been discussed, such as conditional posterior checking or "poor person's" posterior checking wherein the posterior distribution shown in Equation 2.2 is replaced with a multivariate normal approximation (Lee, Cai, & Kuhfeld, in press). All of these methods of model evaluation operate in the lower cell of the center column. The upper cell, however, is vacant—model checking procedures that rely on re-fitting the model rather than computing a test quantity are heretofore unexplored techniques.

The final column, labeled "*All possible data*," comprises the information-theoretic approach to model evaluation, as represented by the minimum description length principle. This column marks the furthest distance from the observed data. In fact, no data need to be observed at all, and the model can be evaluated with respect to many data sets that have been sampled randomly and uniformly from the complete data space (as discussed in great detail in the following chapters). MDL involves re-fitting a model to each of the randomly generated data sets, and thus exists in the upper cell of Column 5. The lower cell is empty because, to my knowledge, there is not yet an information-theoretic method that evaluates a model by generating random data and

then computing a test quantity.

Chapter 7 demonstrates the usefulness of this framework in the context of an IRT analysis of empirical data. However, for that discussion to be informative, the MDL principle—which is entirely novel in the IRT literature—requires formal investigation. Hence, to get a better understanding of the information-theoretic perspective, I conducted an extensive simulation study on the complexity of IRT models. The study specifications and methods are described in Chapter 4, the results are presented in Chapter 5, and the findings are discussed in Chapter 6.

CHAPTER 4

On the Complexity of IRT Models

In psychometric research, it has become common practice to evaluate a number of structurally different measurement models in order to determine the optimal model. In psychological and education research, models are frequently evaluated according to traditional frequentist criteria such as goodness-of-fit to the observed data. Bayesian model checking techniques are seldom seen in the IRT literature, though recent advances in computing power and software availability (e.g., WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000); Stan (Stan Development Team, 2014)) have led to a burgeoning interest in the Bayesian modeling of psychometric data. The principle of minimum description length is wholly foreign to research in psychological and educational measurement. The present work attempts to rectify this oversight by pioneering the use of MDL in the context of psychometric model evaluation.

This study was inspired by the work of Preacher (2006), who explored the MDL principle in the context of structural equation modeling (SEM). Specifically, he examined the concept of *fitting propensity*—a structural model's ability to fit diverse patterns of data, all else being equal. He found that models with the same number of free parameters, but different structures, may exhibit different fitting propensities. That

is, the arrangement of the latent variables in the model may result in an inbuilt tendency to fit any possible data. While this line of research is quite promising, it has yet to be embraced by SEM scholars, due in part to various complications related to generation and estimation of continuous data, and perhaps also because of the curious reasoning underlying the principle of MDL.

Although the philosophical and logical elements of MDL are still alien to many psychometricians, IRT appears to be more accommodating than SEM with regard to various technical aspects of MDL analysis. IRT, unlike SEM, was developed exclusively for modeling categorical data; this greatly simplifies the data generation process, as discussed below. Further, there are a number of statistics that can be derived from an IRT analysis and evaluated in accordance with the MDL principle. These statistics include item-fit measures, local dependence indices, and other aspects of item-level analysis that are uncommon in SEM research. Thus, while Preacher (2006) invoked the MDL principle to provide valuable insight regarding the global fit (via the standardized root mean square) of competing structural models, the IRT analysis presented herein explores not only global fit, but also several metrics that are specialized for IRT model evaluation.

Further, in Preacher's (2006) work on SEM, the structural models under investigation were arbitrary arrangements of causal paths between a few latent variables. Although his findings about fitting propensity in SEM were quite profound,

the unsystematic nature of the models made it difficult to generalize MDL for use in other research scenarios. Common IRT models, however, are given labels that identify their item trace lines (e.g., Rasch, normal ogive, 1PL, 2PL, 3PL, graded response, etc.) and/or their multidimensional factor structure (e.g., bifactor, 2nd-order, correlated traits, two-tier, etc.). This enables one to draw important MDL-based conclusions about certain named models that are popular in IRT research.

The MDL analyses presented in this chapter concern five dichotomous IRT models: an exploratory factor analytic model, a bifactor model, two cognitive diagnostic models, and a unidimensional model.

4.1 IRT Models for Dichotomous Data

4.1.1 Exploratory Factor Analytic Model

The first model under consideration was an exploratory factor analysis (EFA; Spearman, 1904, 1927) model (see Gorsuch (1997) regarding the role of EFA in item analysis). Factor analysis is a statistical technique that aims to model the covariance between variables/items by identifying underlying latent dimensions. In utilizing an exploratory (rather than confirmatory) model, the researcher does not fix *a priori* any of the paths between the latent and observed variables; rather, the model is free to "explore" the combination of latent factors that best represent the manifest variables (i.e., with optimal interpretability and parsimony). For example, an EFA model of a

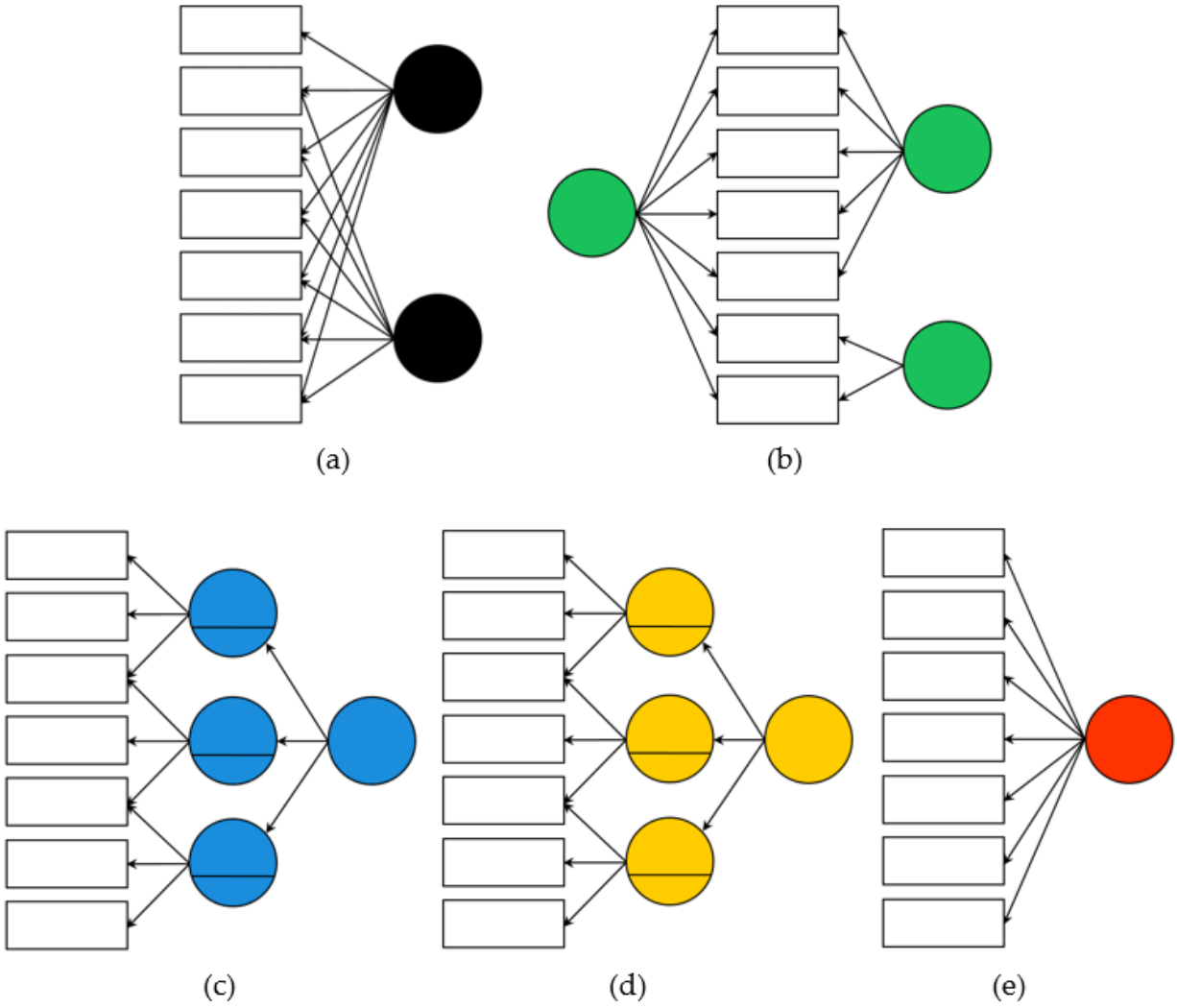


Figure 4.1. Path diagrams of the (a) exploratory factor analysis model; (b) bifactor model; (c) deterministic input, noisy and-gate (DINA) model; (d) deterministic input, noisy or-gate (DINO) model; and (e) unidimensional 3PL model.

mathematics assessment may uncover latent algebra, geometry, and calculus factors that explain the statistical commonalities between particular clusters of items.

Figure 4.1 uses the graphical practices common to structural equation modeling: rectangles represent manifest variables (i.e., items), circles represent latent variables, and the arrows from the latent variables to the manifest variables represent structural

coefficients, or factor loadings. Figure 4.1(a) provides a visual representation of the EFA model under analysis in this chapter. In this case, two factors were selected to represent the seven items. As this was an exploratory model, all of the items were free to load on both of the factors (save the path from Factor 2 to Item 1, which was constrained to zero for model identification purposes).

4.1.2 Bifactor Model

The bifactor model (Holzinger & Swineford, 1937) is a factor structure wherein the covariance among a set of items is explained by a single primary dimension (or "general factor") and multiple specific dimensions (or "group factors"). The primary dimension in a bifactor model represents the overall construct that the test was designed to assess, while the specific dimensions represent narrow subconstructs among groups of items. A psychiatric screening questionnaire, for instance, might measure overall depression (the primary dimension) by including small clusters of questions about mood, sleeping habits, diet, and so forth (the specific dimensions).

The bifactor model has enjoyed a resurgence of late (Reise, 2012), partly because of its strong performance in a number of model comparison studies. Rodriguez, Reise, and Haviland (in press), for example, examined 50 recent psychological research articles in which the bifactor model was selected as the best choice among several competing models (e.g., Ackerman, Donnellan, & Robbins, 2012; Gibbons, Rush, & Immekus, 2009;

Immekus & Imbrie, 2008; Irwin et al., 2012; Osman et al., 2012; Patrick, Hicks, Nichols, & Kreuger, 2007; Reise, Morizot, & Hays, 2007; Simms, Grös, Watson, & O'Hara, 2008; and Yang, Tommet, & Jones, 2009, among many others). All too frequently, this decision was based primarily on the superior goodness of fit of the bifactor model, with minimal regard for it complexity or ability to generalize to future data. In some studies, goodness-of-fit alone was offered as support for a posited theory or hypothesis. For example, Longley, Calamari, Wu, and Wade (2010), developed competing models of anxiety symptoms and concluded, "The better fit of the bifactor model indicates congruence with the integrative model and our hypotheses about hypochondriasis and [obsessive-compulsive disorder] and panic attack symptoms" (p. 461).

The tendency of the bifactor model to exhibit superior goodness-of-fit may be due to its inherent ability to capture random noise in the data. That is, the functional form of the bifactor model may result in overfitting the sample data, thereby causing researchers to draw conclusions that do not generalize to other scenarios. As Thomas (2012) cautioned, "Indiscriminate use of the bifactor model without proper regard for theory is highly questionable . . . Simply put, the bifactor model's added benefit may not excuse its complexity" (p. 108). Indeed, one of the primary motivations for this study was the need to formally evaluate the bifactor model, and by doing so, to lessen its "indiscriminate use."

The particular bifactor structure that was analyzed in the present study is shown

in Figure 4.1(b). This design included a primary dimension, upon which all seven items loaded, and two specific factors. The first specific factor explained variance among items 1 through 5 and the second factor explained variance among items 6 and 7. For model identification purposes, the item factor loadings associated with the second specific factor were constrained to be equal.

4.1.3 Diagnostic Classification Models

Another type of latent variable modeling involves classifying individuals with regard to the attributes underlying the items on a test. For example, presence of symptoms (in psychological assessment) or mastery of content areas (in educational assessment) can be specified as discrete latent variables that are characterized by the various attributes of the items. Models that include such variables yield attribute profiles (i.e., latent classes)—patterns of presence/absence of psychological symptoms or mastery/non-mastery of educational content areas—that can be used to diagnose psychological disorders or ascertain academic shortcomings. These sorts of models are referred to as diagnostic classification models, cognitive diagnostic models, cognitive assessment models, or restricted latent class models, among other labels (Rupp, Templin, & Henson, 2010). The present study included two popular diagnostic classification models: the deterministic input noisy and-gate (DINA) model and the deterministic input noisy or-gate (DINO) model.

The DINA model (Haertel, 1989; Junker & Sijstma, 2001) is non-compensatory, or conjunctive, meaning that presence/mastery of one attribute will not compensate for absence/non-mastery on other attributes. The "and-gate" portion of the DINA acronym indicates that all item attributes must be present/mastered in order to endorse an item/produce the correct response. For example, part A of the DSM-5 diagnostic criteria for autism spectrum disorder requires the presence of deficits in social-emotional reciprocity *and* deficits in nonverbal communicative behaviors *and* deficits in developing, maintaining, and understanding relationships (American Psychiatric Association, 2013). The presence of just one or two of these attributes would not suffice for clinical diagnosis.

Of course, not all items or scales will be non-compensatory. Rather, it may be that presence/mastery of any one of the attributes associated with an item will compensate for the absence/non-mastery of the other attributes. To model items that are compensatory in nature, one can employ the DINO model (Templin & Henson, 2006). Here, the acronym includes an "or-gate," which permits an endorsement/correct response when a respondent demonstrates the presence/mastery of one or more attributes. Part B of the DSM-5 diagnostic criteria for autism spectrum disorder, for instance, requires the presence of at least two of the following: repetitive motor movements *or* inflexible adherence to routines *or* intensely fixated interests *or* hyperreactivity to sensory input (American Psychiatric Association, 2013). The presence

of all four symptoms is not necessary for diagnosis; rather, the presence of any pair of these symptoms will be sufficient.

An informative property of both the DINA and DINO models is the ability to model the attribute space. That is, the pattern of symptomatology or mastery that makes up an attribute profile can itself be measured with a standard IRT model utilizing either the logistic or normal-ogive parameterization (de la Torre & Douglas, 2004). In this so-called "structured tetrachoric model" (Rupp, Templin, & Henson, 2010), each of the discrete latent attribute variables loads on one or more continuous higher-order latent factors. High loadings in the attribute space would indicate a strong relationship between the latent factor(s) and the diagnostic or classification criteria represented by the attributes.

The DINA and DINO models that were analyzed in this study are illustrated in Figure 4.1(c) and (d). Several diagrammatic conventions have been introduced to represent the distinct characteristics of diagnostic classification models. First, the latent attribute variables are divided by a chord, which serves as a visual reminder that these are discrete variables. Second, each diagram includes a pair of cross-loadings, or "interaction effects," which showcase the key difference between these two models. Consider Item 3 for example. In the DINA model, a correct response to Item 3 would require mastery of both Attributes 1 *and* 2. In the DINO model, a correct response to Item 3 would require mastery of either Attributes 1 *or* 2. The remaining paths (denoted

as λs) from the attributes to Items 1, 2, 4, 6, and 7 are termed "main effects" and they indicate items that are associated with a single attribute. Finally, the higher-order latent variable does not include a horizontal chord because it represents a continuous dimension.

4.1.4 Unidimensional 3PL Model

The final model under investigation was a unidimensional IRT model. As depicted in Figure 4.1(e), this model included a single latent dimension to account for variance between the seven items. Among all latent variable measurement models, a unidimensional structure exemplifies the simplest possible functional form. As discussed below, however, the parametric complexity of each item within the model may allow a unidimensional structure to be more flexible than certain multidimensional models.

4.1.5 Differences in Free Parameters

The first four models listed above (EFA, bifactor, DINA, DINO) involved different multidimensional factor structures. In each structure, all items were fit using a 2-parameter logistic (2PL) curve:

$$P(u_i = 1|\theta) = \frac{1}{1 + \exp[-(c_i + a_i\theta)]}, \quad (4.1)$$

where the probability P of a positive response $u = 1$ to item i, given an ability of θ, is

dependent on two freely estimated item parameters: the intercept c_i and the discrimination a_i. The fifth model under investigation was a simple unidimensional structure, but each item was measured using a 3-parameter logistic (3PL) curve:

$$P(u_i = 1|\theta) = g_i + \frac{(1 - g_i)}{1 + \exp[-(c_i + a_i\theta)]}, \quad (4.2)$$

where the additional g_i parameter represents the lower asymptote (or "pseudo-guessing") parameter. By allowing for variability in the lower asymptote, the 3PL model is necessarily more flexible than the 2PL model, and should therefore be more amenable to fitting noise in the data.

Table 4.1 summarizes the freely estimated parameters in each of the models. Although the EFA, bifactor, DINA, and DINO models differed in the number of discrimination parameters and attribute effects, specification of a 2PL trace line for all seven items resulted in exactly 20 freely estimated parameters in each structure. Controlling for the number of free parameters ensured that any observed differences in fitting propensity were due to the functional form of the models rather than the number of free parameters. Specification of the 3PL for all seven items in the unidimensional model resulted in 21 free parameters. That is, relative to the multidimensional models under consideration, the unidimensional model had an extra free parameter. In keeping with the traditional view of model complexity, as discussed in previous chapters, the enhanced flexibility of the 21-parameter unidimensional 3PL model should cause it to have a higher fitting propensity than the 20-parameter multidimensional models.

Table 4.1. *Parameterizations of the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

Model	Structure
EFA	7 c (intercept) parameters 7 a (discrimination) parameters for Factor 1 + 6 a (discrimination) parameters for Factor 2 **20 free parameters**
Bifactor	7 c (intercept) parameters 7 a (discrimination) parameters for the General Factor 5 a (discrimination) parameters for Specific Factor 1 + 1 a (discrimination) parameters for Specific Factor 2 **20 free parameters**
DINA & DINO	7 $\lambda_{1,0}$ (intercept) parameters 2 $\lambda_{1,1,(1)}$ (main) effect parameters for Attribute 1 1 $\lambda_{1,1,(2)}$ (main) effect parameter for Attribute 2 2 $\lambda_{1,1,(3)}$ (main) effect parameters for Attribute 3 1 $\lambda_{1,2,(1,2)}$ (interaction) effect parameter for Attributes 1 & 2 1 $\lambda_{1,2,(2,3)}$ (interaction) effect parameter for Attributes 2 & 3 3 c (attribute intercept) parameters + 3 a (attribute discrimination) parameters **20 free parameters**
Uni	7 c (item intercept) parameters 7 a (item discrimination) parameters + 7 g (pseudo-guessing) parameters **21 free parameters**

Note. EFA = exploratory factor analysis; DINA = deterministic input noisy and-gate; DINO = deterministic input noisy or-gate; Uni = unidimensional 3PL model.

4.2 Hypotheses

Regarding the performance of these models in the context of any possible data, I offer two hypotheses.

Hypothesis 1: The EFA model will exhibit, on average, the highest fitting propensity.

The EFA model was included as a baseline of sorts, since the exploratory nature of this model should provide it with the highest degree of fitting propensity. That is, unless the underlying "true" data generating mechanism of the chosen random data set just happens to represent a bifactor, DINA, DINO, or unidimensional 3PL model, then the EFA model should always fit best.

Hypothesis 2: The bifactor model will display higher fitting propensity than the DINA and DINO models.

The second hypothesis is that the bifactor item response model, relative to the DINA and DINO models, will fit a greater number of data sets that are randomly sampled from (and uniformly distributed across) the complete data space. As discussed earlier, the bifactor model has become increasingly popular in recent years (e.g., Reise, 2012), due in part to its ability to closely fit the observed data. However, I believe that the functional form of the bifactor model instills in it an undesirable tendency to fit any possible data.

The unidimensional 3PL model was included in this analysis to better understand the notion of complexity in IRT models. Preacher (2006) found that fitting propensities varied when models with the same number of parameters differed in functional form. Among the five factor structures included in the present study, the unidimensional model has the simplest functional form; it is the only model in which each item loads on a single latent variable. This economy of functional form may cause the unidimensional 3PL model to have a drastically reduced fitting propensity. However, all items in the unidimensional model were fit using 3PL trace lines, which increased the model's complexity, as gauged by traditional metrics (i.e., the number of free parameters). Because of this contradiction between functional form and parametric complexity, I do not offer a clear hypothesis regarding the unidimensional 3PL model; I choose instead to simply observe its performance relative to the competing multidimensional models.

4.3 Method

Preacher (2006) noted that expressions of the minimum description length principle, such as the normalized maximum likelihood (Equation 2.8), are intractable due to integration over the complete data space. He concluded, "Until a good analytic approximation can be identified, calculation of an MDL index in the SEM context involves fitting a model to a large number of random data sets" (p. 249). Although it is possible that the intractability of the NML and related expressions may not extend to

IRT, for the purposes of the present study, I followed the same strategy as Preacher (2006): to explore the bifactor model using the MDL principle, I generated random data that were uniformly distributed across the complete data space.

4.3.1 Data Generation

Data generation was accomplished by attaching to all possible response patterns a weight randomly sampled from a unit simplex. Smith and Tromble (2004) demonstrated that sampling from a unit simplex is ideal for situations in which the goal is to obtain random multinomial probability distributions that are uniformly sampled across a range from 0.0 to 1.0. The simulated data included seven dichotomous items, resulting in $2^7 = 128$ unique response patterns. In the first step of the random weighting scheme, $2^7 - 1 = 127$ integers were randomly sampled from an array of numbers between 1 and 1,000,000; the resulting vector **W** of random response pattern frequencies was then sorted in ascending order and values of 0 and 1,000,000 were appended to the ends of **W**. Next, the lag-1 difference (i.e., the difference between the n and $n + 1$ elements of **W**) was computed, and these lagged differences were divided by 1,000,000. The weights were then multiplied by the (theoretical) sample size $N = 10{,}000$ and a unique weight from **W** was appended to each of the $2^7 = 128$ response patterns. This entire process was replicated such that 1,000 unique random data sets were created. The data generation script, written in the R statistical software program (R Core Team,

2014), is presented in Appendix A, along with an example data set.

It is important to note that the data generation process outlined above will not result in data that have some known underlying structure. The data were explicitly designed to represent the *complete* data space, which implies that each of the models should fit well to at least some subset of the data sets. To be precise, the complete data space necessarily includes data that truly are unidimensional in nature, data that truly align to a bifactor structure, and so on. Any model that demonstrates a predisposition to fit well to a relatively large number of these data sets is a model that is remarkably (perhaps overly) flexible. If the bifactor model tends to fit well, for example, that is not because the data were necessarily generated from a bifactor structure; rather, such a finding would highlight a property of the bifactor model itself, as an excessively flexible model that bends to fit any possible data.

4.3.2 Estimation Specifications

Once the randomly weighted data were generated, an R script was written to fit each of the five models to the same 1,000 data sets using the flexMIRT software program (Cai, 2013). The flexMIRT `CaseWeight` command was used to identify the randomly sampled and uniformly distributed vector **W** of response pattern frequencies.

In all models, all items were estimated using the Bock-Aitkin expectation-maximization (EM) algorithm (Bock & Aitkin, 1981) with 49 quadrature points between

-6.0 and 6.0. The EFA, bifactor, and unidimensional models used cross-product approximation to compute standard errors, while the two diagnostic classification models used the Richardson extrapolation method. The diagnostic models also differed from the others by specifying a maximum of 5 M-step iterations, rather than the flexMIRT default of 100. These changes in estimation of the cognitive diagnostic models were advised by the authors of the flexMIRT software (Houts & Cai, 2013). Additionally, to aid in estimation of the unidimensional 3PL model, a Beta(1.0, 4.0) prior was placed on the pseudo-guessing parameter of each item.

In all models, the potential difficulties in estimating the random data parameters were addressed by setting the E-step tolerance at .001 (rather than the more stringent flexMIRT default of .00001) and increasing the maximum number of E-step iterations to 20,000 (from the flexMIRT default of 2,000). These alterations were intended to allow the EM algorithm to achieve an adequate degree of stability in the absence of a converged solution.

For the purposes of this study, estimation convergence was defined as the detection of a local maximum according to a 2nd-order test. Despite the relaxed tolerance and the increase in estimation iterations, there were still a number of replications that did not settle on stable parameter estimates. Table 4.2 displays the convergence rates for each of the dichotomous IRT models under investigation. The unidimensional 3PL model had the highest non-convergence rate—when fit to the unidimensional model,

24.3% of the data sets failed to converge on a stable solution. Non-convergence rates were slightly lower for the EFA (21.0%) and bifactor (18.4%) models. Interestingly, the two cognitive diagnostic models had far greater success with regard to estimation convergence. The DINA model converged on stable estimates in all but 4.7% of the data sets; the DINO model fared even better, obtaining stable solutions in all but 3.9% of the data sets. Ideally, all five models would have achieved 100% convergence in all data sets. We echo the reasoning of Preacher (2006), who argued that estimates computed after 10,000 E-step iterations can be accepted as the final (converged) estimates, despite their possible instability. In specifying a maximum of 20,000 E-step iterations, our aim was to afford further confidence in the non-converged results.

Table 4.2. *Estimation convergence rates of the exploratory factor analytic, bifactor, dynamic input noisy and-gate, dynamic input noisy or-gate, and unidimensional 3PL models.*

Model	% Non-converged	% Converged
EFA	21.0	79.0
Bifactor	18.4	81.6
DINA	4.7	95.2
DINO	3.9	96.1
Uni	24.3	75.7

Note. 1,000 data sets were fit to each model. Convergence stability was based on detection of a local maximum. EFA = exploratory factor analysis; DINA = dynamic input noisy and-gate; DINO = dynamic input noisy or-gate; Uni = unidimensional 3PL model.

4.3.3 Simulation Specifications

To appraise the fitting propensities of various structural models, Preacher (2006) selected the root mean squared residual (RMSR; Jöreskog & Sörbom, 1996) as the appropriate metric of model fit. RMSR was chosen specifically because it does *not* adjust for the number of free parameters or the functional form of the model. RMSR is, in a sense, a "pure" measure of fit that is unswayed by the characteristics of the model. Thus, RMSR allows one to measure the fitting propensities of competing models simply by comparing differences in their fit to the same (random) data.

However, like other common fit measures in SEM, the computation of RMSR requires a correlation matrix based on continuous variables. The classical goodness-of-fit indices that are so common in structural equation and factor analytic models cannot be used in item response modeling because of the categorical nature of the data. Thus, the present analysis focused on five diagnostic measures that were developed specifically for use in categorical data analysis: the Y2/*N* statistic, the D^2 latent distribution fit index, the marginal χ^2 statistic, the LD X^2 local dependence index, and the $S\text{-}X^2$ item-fit statistic.

4.3.4 Y2/*N* Statistic

Perhaps the closest analog to RMSR that currently exists for discrete data is the Y2 statistic (Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen,

2006). This fit statistic is found by summing all the univariate and bivariate marginal chi-squares derived from the contingency tables of item response probabilities:

$$\mathrm{Y2} = N\left[\sum_{i=1}^{I}\frac{(\mathrm{o}_i - \mathrm{e}_i)^2}{\mathrm{e}_i(1-\mathrm{e}_i)} + \sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\frac{\left(\mathrm{o}_{ij} - \mathrm{e}_{ij}\right)^2}{\mathrm{e}_{ij}\left(1-\mathrm{e}_{ij}\right)}\right], \tag{4.3}$$

where N is the sample size, I is the number of items, o_i and e_i are the observed and expected positive response frequencies for item i, and o_{ij} and e_{ij} are the observed and expected positive response frequencies for item pair ij. Y2 denotes the magnitude of the discrepancy between the data and the statistical model; it is a "badness-of-fit" index in that higher values indicate worse fit. To date, no benchmark values have been established for the Y2 statistic. In the present study, Y2 was divided by the sample size N to produce the Y2/N statistic. This slightly modified version of the Y2 statistic is independent of sample size, allowing for easier comparison with future studies that likely will not boast a sample of 10,000 respondents. A practical way to interpret the Y2/N statistic is as a metric of badness-of-fit per respondent.

4.3.5 D^2 Latent Distribution Fit Index

Another characteristic of the overall IRT model is the latent distribution fit. Item response models are routinely implemented under the assumption that each latent variable in the model is normally distributed. Yet, as Thissen and Wainer (2001) observed, "When the population distribution assumed in the IRT model does not well

represent the distribution of theta for the examinees, then the inferred score distribution will depart from the observed score distribution" (p. 130). That is, if dependable IRT scores are desired, it is imperative to assess the discrepancy between the observed and expected distributions of scores.

To address this issue, Li and Cai (2012) suggested a measure of latent distribution fit based on the Cressie-Read (1984) power divergence family of fit statistics. These statistics are designed to assess the closeness of the observed multinomial variables to their expected values. The Cressie-Read family is denoted as:

$$D(\lambda) = \frac{2N}{\lambda(\lambda+1)} \sum_{k=0}^{K} p_k \left\{ \left(\frac{p_k}{\pi_k} \right)^{\lambda} - 1 \right\}, \tag{4.4}$$

where λ is a scalar, N is the sample size, and p_k and π_k are the observed and expected probabilities for summed scores $k = 0, \ldots, K$, where K represents the maximum possible summed score. Cressie and Read (1984) found that two popular measures of absolute fit in IRT are special cases of the equation above. Specifically, when $\lambda = 0$, the result is the loglikelihood ratio G^2 statistic; when $\lambda = 1$, the result is the X^2 statistic.

Li and Cai (2012) evaluated the fit of the latent distribution by first obtaining the observed summed score probabilities. Then the model-implied probabilities were computed using the Lord-Wingersky algorithm (Lord & Wingersky, 1984; Cai, 2014). Chi-square statistics, based on Equation 4.4 above, were then constructed in order to compare the observed and expected probabilities. A simulation study was carried out in

hopes of finding a chi-square distributed statistic that was sensitive to nonnormality in the distribution of theta, though not sensitive to other types of model misspecification. Based on the results of this study, Li and Cai (2012) recommended setting the λ in the equation above at 2/3, resulting in the D^2 index:

$$D^2 = \frac{2N}{\frac{2}{3}\left(\frac{2}{3}+1\right)} \sum_{k=0}^{K} p_k \left\{ \left(\frac{p_k}{\pi_k} \right)^{\frac{2}{3}} - 1 \right\}. \tag{4.5}$$

The authors concluded that the D^2 statistic is a direct measure of latent variable nonnormality that works well in dichotomous and polytomous IRT modeling across various test lengths and sample sizes.

4.3.6 *S*-X^2 Item-Fit Index

The Y2/*N* and D^2 statistics discussed above are test-level indices of model fit; IRT models can also be appraised by focusing on item-level diagnostics, such as the fit of each individual item. Orlando and Thissen (2000, 2003) introduced a method of constructing an item-fit statistic based on the observed and expected proportions of correct and incorrect responses for each summed score in the sample. The observed summed scores are computed directly and the model-implied joint likelihood distributions for each summed score are computed using the Lord-Wingersky algorithm (Lord & Wingersky, 1984; see also Thissen, Pommerich, Billeaud, and Williams, 1995). The observed and expected response pattern frequencies can then be

compared directly via the $S\text{-}X^2$ item-fit index for dichotomous IRT models (Orlando & Thissen, 2000, 2003):

$$S\text{-}X_i^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}\,, \tag{4.6}$$

where O_{ik} and E_{ik} are the observed and expected proportions of respondents with summed score k who responded correctly to item i (E_{ik} is found via Equation 12 in Orlando and Thissen (2000)). This chi-square distributed statistic enables one to assess the degree to which an item response curve is under- or overestimating the proportion of correct responses at different levels of the underlying trait or ability.

4.3.7 Marginal $\chi 2$

In IRT, goodness-of-fit assessment is based on analysis of multiway contingency tables, the cells of which contain the observed and expected correct and incorrect response rates for every item or pair of items. One straightforward method of item-level evaluation in IRT involves using the marginal values of the univariate contingency tables to compute a χ^2 statistic for each item. The null hypothesis of the marginal χ^2 test is that the observed frequencies are equal to the expected frequencies, given the marginal values. That is, if the observed proportions of correct response are identical to the expected proportions, then the marginal χ^2 will equal 0.0.

4.3.8 LD X^2 Local Dependence Index

The final item analysis statistic quantifies one of the primary assumptions of IRT: local independence. IRT models assume that items are only correlated through the underlying latent construct that the item set is designed to measure (Lord & Novick, 1968). If residual correlations exist after accounting for the correlations explained by the latent trait, then the assumption of local independence has been violated. Local dependence (LD) can occur when items have near-identical content (e.g., multiple items that refer to a single reading passage), when the response to one item is conditional on the response to a previous item, when some unmodeled latent dimension exists in the data, and so on. Failure to address LD violations may result in biased item parameter estimation, inaccurate IRT scaled scores, and inflated information functions and reliability estimates, among other problems (Thissen, Steinberg, & Mooney, 1989; Sireci, Thissen, & Wainer, 1991).

Chen and Thissen (1997) developed the LD X^2 index to address local dependence violations in IRT models. To compute this index, phi correlations are calculated for the observed and expected bivariate contingency tables. When the observed correlation is higher than the model-implied correlation for an item pair, the result is positive LD; if the model-implied correlation is higher, then negative LD has been detected within that item pair. The absolute value of the LD X^2 statistics can then be tested against some critical value to determine whether the violation is ignorable (Houts & Cai, 2013).

CHAPTER 5

On the Complexity of IRT Models:

Simulation Results

5.1 Y2/*N* Statistic

For all five models, Table 5.1 displays the overall means and standard deviations of the Y2/*N* statistic for the total, converged, and non-converged analyses, as well as the difference between the converged and non-converged analyses. This table provides a general comparison between all models, as well as a more detailed comparison of the converged and non-converged analyses within each model. Beginning with the between-model comparisons across all 1,000 data sets, Table 5.1 reveals that on average, the EFA and bifactor models produced Y2/*N* values of .05 or lower. That is, on average, the bifactor model was almost as capable as the EFA model with regard to fitting any possible data. The DINA and DINO models tended to have Y2/*N* values of .10, and the unidimensional 3PL model yielded an average Y2/*N* of .13.

Table 5.1 also facilitates within-model comparisons of the converged and non-converged data sets. Although there were differences in each model's convergence rate (as discussed earlier with regard to Table 4.2), the Y2/*N* results were fortunately not

Table 5.1. *Means and standard deviations of the Y2/N statistic in the exploratory factor analytic, bifactor, deterministic noisy input and-gate, deterministic noisy input or-gate, and unidimensional 3PL models.*

	EFA		Bifactor		DINA		DINO		Uni	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.04	.02	.05	.02	.10	.04	.10	.04	.13	.05
Converged	.04	.02	.05	.02	.10	.04	.10	.04	.13	.05
Non-conv.	.04	.02	.05	.02	.10	.05	.11	.05	.13	.05
Difference	.00	.00	.00	.00	.00	-.01	-.01	-.01	.00	.00

Note. *N* = 1000; EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic noisy or-gate model; Uni = unidimensional 3PL model; Non-conv = non-converged analyses.

affected by the lack of convergence. The indistinguishability of the converged and non-converged analyses gives credence to the deeper investigation of Y2/*N* that is shown in later tables, wherein the results are based on all 1,000 data sets regardless of non-convergence.

Although the descriptive statistics hint at differences between the models, the fitting propensities are better expressed through visualizations of Y2/*N*. Figure 5.1 displays the empirical cumulative percentage distribution of the Y2/*N* statistic in the EFA (black), bifactor (green), DINA (blue), DINO (yellow), and unidimensional 3PL (red) models. The curves in the figure simply display the percentage of data sets that achieved a particular value of Y2/*N* when fit to each model. This type of figure allows for the models to be compared in two ways. The first way is by investigating the vertical distance between the curves at some particular value of Y2/*N*. For example,

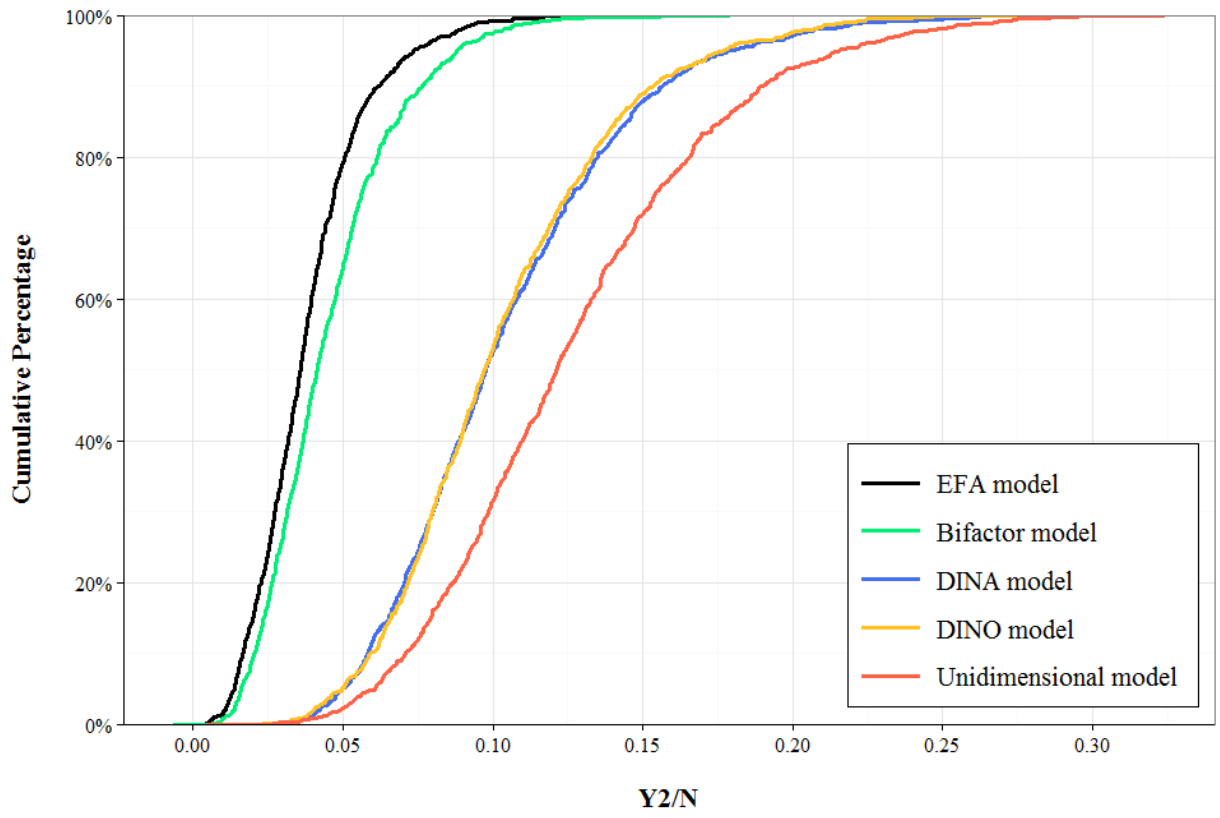


Figure 5.1. Cumulative percentage distributions of the Y2/*N* statistic in the exploratory factor analytic (EFA), bifactor, deterministic noisy input and-gate (DINA), deterministic noisy input or-gate (DINO), and unidimensional 3PL models.

consider the vertical grid line at Y2/*N* = .05. The EFA curve intersects with this line at *y* = 79%, meaning that 790 of all 1,000 data sets had Y2/*N* values of .05 or lower when fit to the EFA model. The green curve reveals that the bifactor model produced a Y2/*N* of .05 or less in 63.5% of all data sets. The diagnostic classification and unidimensional 3PL models were far less likely to yield Y2/*N* values as low as .05. Specifically, Y2/*N* values of .05 or lower were only obtained in 5.0% of data sets fit to the DINA model, 5.2% of the data sets fit to the DINO model, and 2.3% of the data sets fit to the unidimensional

3PL model.

Tables 5.2 and 5.3 display the percentage of all data sets that reached particular values of Y2/*N*. These tables supplement the cumulative percentage distributions displayed in Figure 5.1 by providing precise values that are not clearly visible in the figure. Table 5.2 presents the percentage of data sets that attained Y2/*N* values between .01 and .15 in each of the five models. The results displayed in the first set of rows are especially telling. For instance, the column labeled Y2/*N* ≤ .03 reveals that the bifactor model fit well to 270 of all 1,000 data sets, yet the DINA, DINO, and unidimensional 3PL models achieved that degree of goodness-of-fit in just three data sets. Similarly, the second set of rows shows that almost every single data set achieved a Y2/*N* ≤ .10 when fit with an EFA or bifactor model. The two diagnostic classification models attained this degree of goodness-of-fit in just over half of all data sets, and the unidimensional 3PL model in less than 1/3rd.

Table 5.3 contains the same type of information as the previous table, but for higher values of Y2/*N*. This table highlights the mild fitting propensity of the unidimensional 3PL model. The first column shows that all models but the unidimensional model achieved Y2/*N* ≤ .16 in over 90% of data sets; the unidimensional model, however, which had an additional free parameter, only reached Y2/*N* ≤ .16 in 77.5% of the data sets. In fact, it is not until Y2/*N* ≤ .40 that virtually every data set can be accounted for by the unidimensional 3PL model. To put this into perspective, every

Table 5.2. *Percent of 1,000 data sets that attained Y2/N values between .01 and .15 when fit to the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

	Y2/*N*				
Model	≤ .01	≤ .02	≤ .03	≤ .04	≤ .05
EFA	1.4	14.9	36.1	61.1	79.0
Bifactor	0.9	9.4	27.0	45.9	63.5
DINA	0.0	0.0	0.3	1.2	5.0
DINO	0.0	0.0	0.3	2.0	5.2
Uni	0.0	0.0	0.3	0.8	2.3
	≤ .06	≤ .07	≤ .08	≤ .09	≤ .10
EFA	89.1	93.8	96.4	98.1	99.1
Bifactor	77.1	85.7	91.3	95.5	97.3
DINA	11.8	19.2	30.0	41.1	52.5
DINO	10.2	18.2	30.2	41.4	52.9
Uni	4.9	9.5	16.1	22.3	31.5
	≤ .11	≤ .12	≤ .13	≤ .14	≤ .15
EFA	99.5	99.8	99.8	100.0	100.0
Bifactor	98.8	99.4	99.6	99.8	99.8
DINA	61.4	69.3	76.1	82.8	87.8
DINO	63.6	71.1	77.8	84.5	88.9
Uni	40.2	48.8	57.6	65.6	72.0

Note. All values are percentages of 1,000 data sets. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

Table 5.3. *Percent of 1,000 data sets that attained Y2/N values between .16 and .30 when fit to the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

	Y2/*N*				
Model	≤ .16	≤ .17	≤ .18	≤ .19	≤ .20
EFA	100.0	100.0	100.0	100.0	100.0
Bifactor	99.9	100.0	100.0	100.0	100.0
DINA	90.9	93.6	95.0	96.2	97.0
DINO	91.6	93.5	95.7	96.5	97.6
Uni	77.5	83.3	86.6	90.1	92.6
	≤ .21	≤ .22	≤ .23	≤ .24	≤ .25
EFA	100.0	100.0	100.0	100.0	100.0
Bifactor	100.0	100.0	100.0	100.0	100.0
DINA	98.1	98.7	99.0	99.2	99.3
DINO	98.5	99.1	99.6	99.7	99.9
Uni	93.9	95.4	96.5	97.7	98.1
	≤ .26	≤ .27	≤ .28	≤ .29	≤ .40
EFA	100.0	100.0	100.0	100.0	100.0
Bifactor	100.0	100.0	100.0	100.0	100.0
DINA	99.6	99.8	100.0	100.0	100.0
DINO	100.0	100.0	100.0	100.0	100.0
Uni	98.8	99.3	99.6	99.8	100.0

Note. All values are percentages of 1,000 data sets. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

single data set fit the EFA and bifactor models when Y2/$N \leq .13$ and .17, respectively.

The second way to compare the cumulative percentage distributions is to consider the horizontal discrepancy between the curves in Figure 5.1. Suppose that a researcher is interested in evaluating each model, not by selecting some referent value of Y2/N, but instead against some benchmark percentage. The horizontal grid line at y = 80%, for instance, indicates that 80% of all EFA fittings achieved Y2/$N \leq .05$, 80% of all bifactor fittings resulted in Y2/$N \leq .06$, 80% of all DINA and DINO fittings had Y2/$N \leq$ .13, and 80% of all unidimensional 3PL fittings produced Y2/$N \leq .17$.

Table 5.4 presents the Y2/N values at every 10^{th} percentile for each of the five models. A consistent pattern exists within each row: the EFA model always had the lowest Y2/N value, the bifactor model followed closely behind, the two diagnostic classification models produced higher values (and performed almost identically), and the unidimensional 3PL model offered the highest Y2/N values. A few interesting comparisons can be made. For example, 40% of EFA model fittings yielded Y2/N values of .03 or lower, but not a single DINA or unidimensional model fitting produced Y2/N values of that magnitude. An even more drastic disparity is found by comparing the highest and lowest deciles: 90% of EFA and bifactor model fittings revealed Y2/$N \leq .06$ and .08, respectively, but only 10% of DINA, DINO, and unidimensional model fittings resulted in similar Y2/N statistics.

It is clear from the Y2/N results discussed above that the EFA and bifactor

Table 5.4. *Y2/N values at certain percentages of 1,000 data sets when fit to the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

Percentage of all data sets	Y2/*N*				
	EFA	Bifactor	DINA	DINO	Uni
100%	.13	.17	.27	.26	.40
90%	.06	.08	.16	.15	.19
80%	.05	.06	.13	.13	.17
70%	.04	.05	.12	.12	.15
60%	.04	.05	.11	.11	.13
50%	.04	.04	.10	.10	.12
40%	.03	.04	.09	.09	.11
30%	.03	.03	.08	.08	.10
20%	.02	.03	.07	.07	.09
10%	.02	.02	.06	.06	.07
0%	.00	.01	.03	.02	.03

Note. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

models possess much greater propensities to fit any possible data. These findings, while informative, do not offer any details about the degree of overlap between the models. In the MDL literature, it is not uncommon to see figures showing hypothetical regions of the complete data space that are "occupied" by competing models (see e.g., Pitt, Myung, & Zhang, 2002). It could be, for instance, that even though the DINA and DINO models tend to fit well to approximately the same *percentage* of data sets, the actual data sets that they fit well could be completely different. A series of visualizations were

created to better understand how the five models under investigation interacted within the complete data space.

The "amoeba" plots presented in Figures 5.2 – 5.4 depict the fitting propensities of each model at various levels of Y2/*N*, just as in the tables discussed above, but they also reveal the overlap (and lack thereof) that characterizes these models. In each of these figures, the square area represents the complete data space. The transparent colored regions represent the number of data sets (out of all 1,000 data sets) that fit the corresponding model at a specific value of Y2/*N*. The regions are drawn roughly to scale; the values that accompany each figure indicate the size of each region as well as the precise degree of overlap between regions.

The top panel of Figure 5.2 shows one of the simplest scenarios: Y2/*N* ≤ .01. Here, the EFA (black) model occupied just 1.4% of the complete data space and the bifactor (green) model occupied 0.9%. That is, at this strict Y2/*N* criterion, the EFA model fit well to 14 of the 1,000 random data sets and the bifactor model fit well to 9 data sets. This figure reveals that the bifactor region was not fully subsumed by the EFA region; that is, there were some data sets that fit well to the EFA model but not the bifactor model, and vice versa. As the figure shows, the overlap between the EFA and bifactor models (denoted as region A) occupied 0.4% of the data space, meaning that 4 out of 1,000 data sets were fit extremely well (Y2/*N* ≤ .01) by both models.

Regions B and C in this first amoeba plot highlight the unique data sets that were

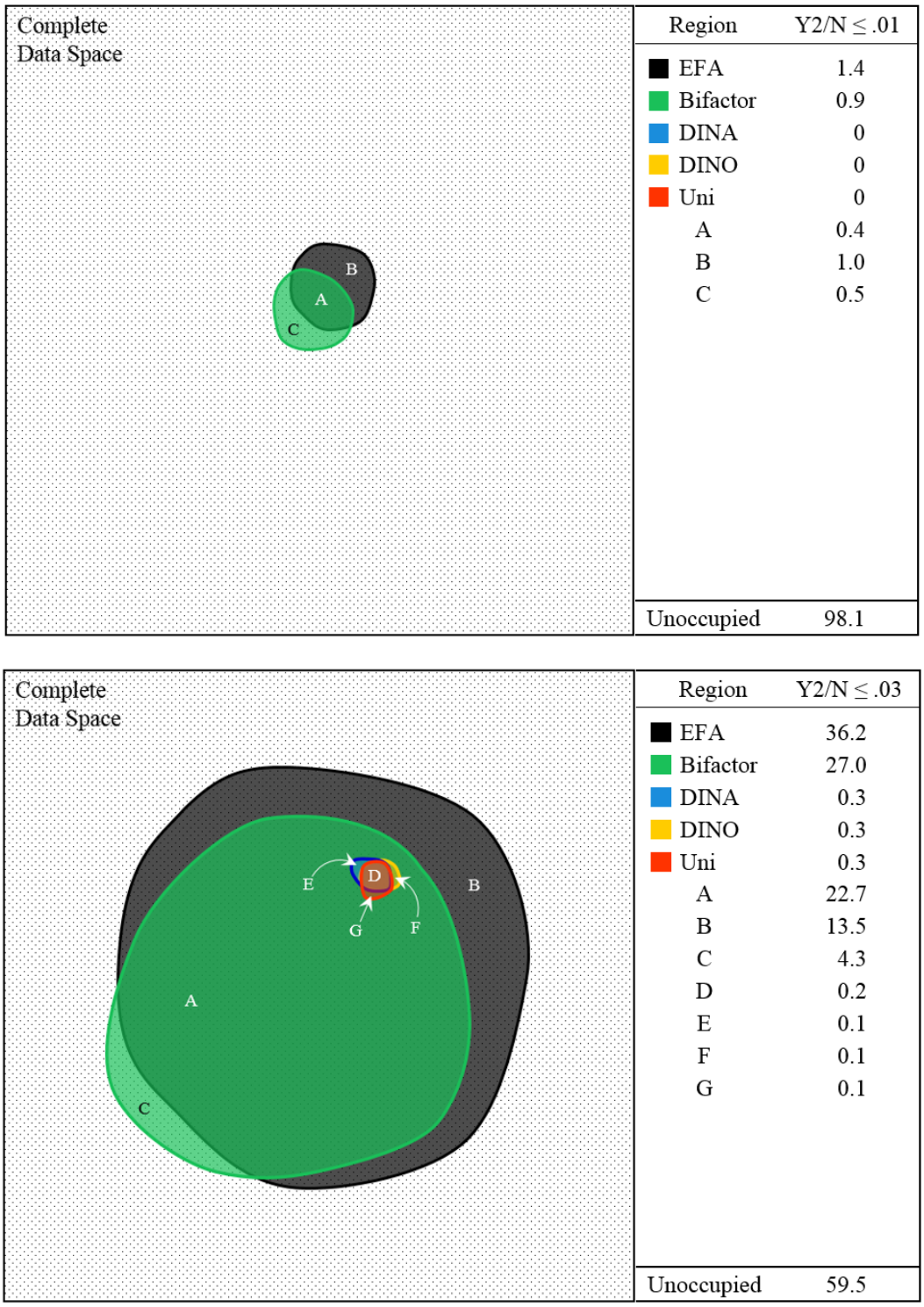
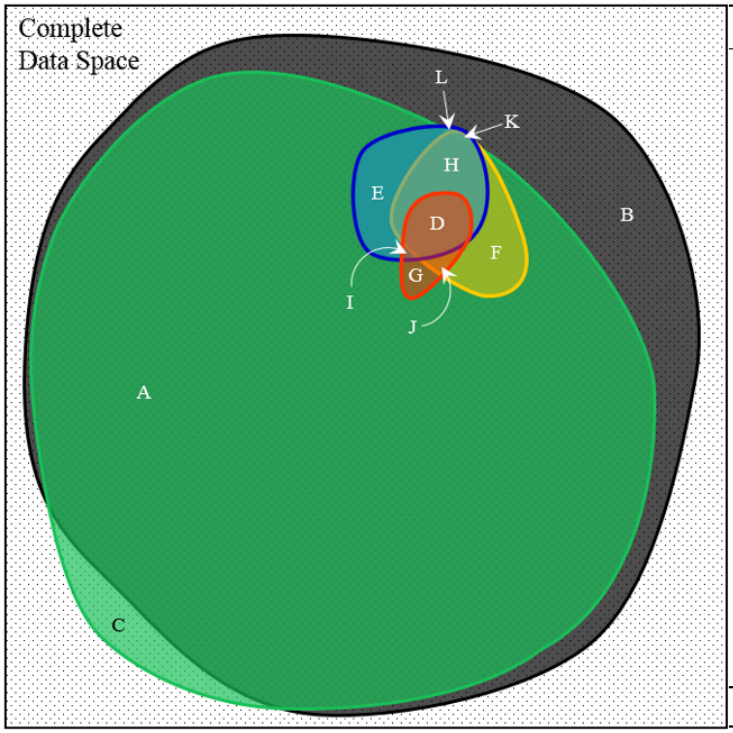


Figure 5.2. Hypothetical regions of the complete data space that were occupied by each model when Y2/*N* ≤ .01 (top) and Y2/*N* ≤ .03 (bottom). All values are percentages of 1,000 data sets. Regions drawn roughly to scale. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

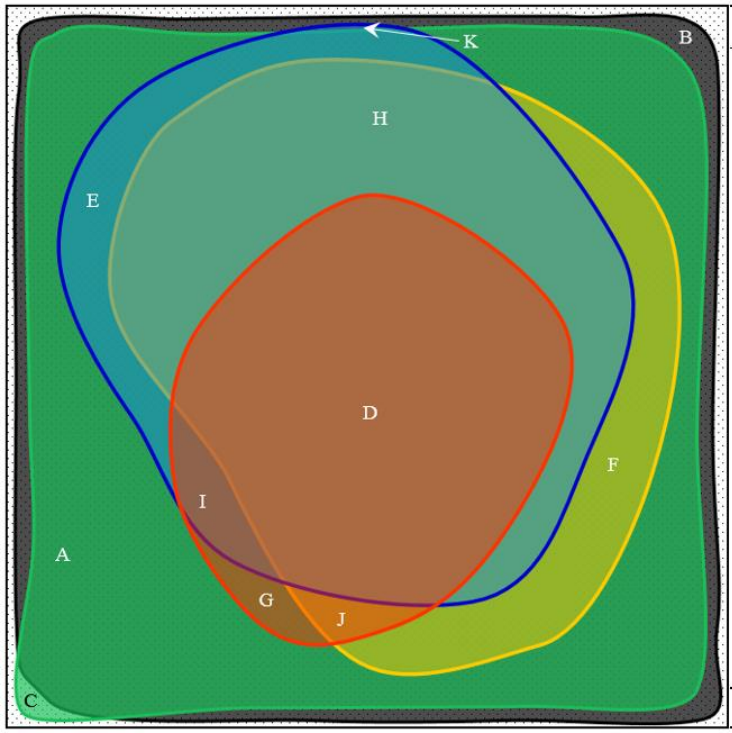
fit well by each model. The EFA model fit 1.0% (region B), or 10 data sets that were not fit by the bifactor model; the bifactor model fit 0.5% (region C), or 5 data sets that were not well by the EFA model. Finally, the values presented in this panel show that at Y2/*N* ≤ .01, the DINA, DINO, and unidimensional 3PL models did not occupy any part of the data space, and that 98.1% of the complete space remained unoccupied by any of the candidate models.

The bottom panel of Figure 5.2 depicts regions of the complete data space that were occupied by each model when Y2/*N* ≤ .03. In this case, the EFA model fit 36.2% of all data sets and the bifactor model fit 27.0%. These two models overlapped such that 22.7% (region A) of all data sets were fit well by both models. Note, however, that 4.3% (region C) of the data sets fit the bifactor model but *not* the EFA model. The DINA, DINO, and unidimensional 3PL models made an appearance when Y2/*N* ≤ .03, though the regions they occupied were quite small and the overlap between them was extensive. Specifically, the diagnostic classification and unidimensional models each fit the same two data sets (region D), and each also fit one unique data set (regions E, F, and G). Finally, at this Y2/*N* benchmark, 59.5% of the total data space was not occupied by any of the models.

The top panel of Figure 5.3 is a visualization of the complete data space when Y2/*N* ≤ .05. Here, the EFA region occupied the majority (79.2%) of the space and the bifactor model was not far behind (63.5%). The overlap between these two models was



Region	Y2/N ≤ .05
EFA	79.2
Bifactor	63.5
DINA	5.1
DINO	5.2
Uni	2.3
A	51.8
B	19.2
C	3.8
D	1.1
E	1.9
F	2.0
G	0.7
H	1.7
I	0.2
J	0.3
K	0.1
L	0.1
Unoccupied	17.1



Region	Y2/N ≤ .10
EFA	99.2
Bifactor	97.3
DINA	52.6
DINO	52.9
Uni	31.5
A	96.9
B	2.1
C	0.4
D	22.8
E	6.4
F	8.1
G	2.5
H	19.4
I	3.8
J	2.4
K	0.2
Unoccupied	0.4

Figure 5.3. Hypothetical regions of the complete data space that were occupied by each model when Y2/*N* ≤ .01 (top) and Y2/*N* ≤ .03 (bottom). All values are percentages of 1,000 data sets. Regions drawn roughly to scale. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

sizeable—over half (51.8%) of all data sets were fit well by both the EFA and bifactor models. Even at this level of Y2/*N*, however, there were still a few data sets (region C: 3.8%) that fit the bifactor model but not the EFA model.

The DINA, DINO, and unidimensional 3PL models were completely subsumed by the bifactor and EFA models; that is, there were not any data sets that fit the diagnostic classification or unidimensional models without also fitting the bifactor or EFA models. However, the blue region shows that 5.1% of all data sets fit the DINA model at Y2/*N* ≤ .05, and regions E + I indicate that 2.1% of data sets fit the DINA model but not the DINO model. The yellow region shows that 5.2% of all data sets fit the DINO model and regions F + J reveal that 2.3% of data sets fit the DINO model but not the DINA model.

The red region indicates that 2.3% of all data sets fit the unidimensional 3PL model at Y2/*N* ≤ .05. While there was some overlap between the unidimensional and diagnostic classification models, there were still 7 data sets (region G) that fit the unidimensional model without fitting either the DINA or DINO models. Another region of interest is region D, which represents the overlap of all five models. This region occupied 1.1% of the complete data space; that is, 11 data sets in the simulation were fit well by all models. Finally, note that when Y2/*N* ≤ .05, only 17.1% of the complete data space was not occupied by any model.

The bottom panel in Figure 5.3 displays the total data space when Y2/*N* ≤ .10.

Here, the EFA and bifactor models fit almost every data set, occupying 99.2% and 97.3%, respectively, of the complete data space. Yet there were still 4 data sets (region C) that fit the bifactor model but not the EFA model. At this level of Y2/*N*, the DINA and DINO regions also showed considerable overlap; each of these models occupied over 52% of the data space, but 42.2% (regions D + H) of all data sets fit both the DINA and DINO models. In the center of this figure, region D indicates that 228 of all 1,000 data sets fit all five models when Y2/*N* ≤ .10. At this relatively high level of Y2/*N*, only 4 data sets were not fit by some model.

Finally, Figure 5.5 provides a summary of the growth in occupation of the data space as Y2/*N* increased from .01 to .05. As in the cumulative percentage distributions shown in Figure 5.1, the amoeba plots shown here indicate that the regions representing the EFA and bifactor models grew very rapidly as Y2/*N* increased. However, the area that fit the bifactor model alone did not seem to change too drastically. When Y2/*N* ≤ .03, there were 43 data sets that only fit the bifactor model; at Y2/*N* ≤ .05, there were 38 data sets that only fit the bifactor model. Further, the diagnostic classification and unidimensional 3PL regions did not grow nearly as quickly as the bifactor and EFA areas, though it is apparent that as Y2/*N* increased, the DINA, DINO, and unidimensional regions grew slightly and began to separate themselves.

Overall, the Y2/*N* results revealed that the fitting propensity of the bifactor model approached that of the EFA model—a model specifically intended to find the

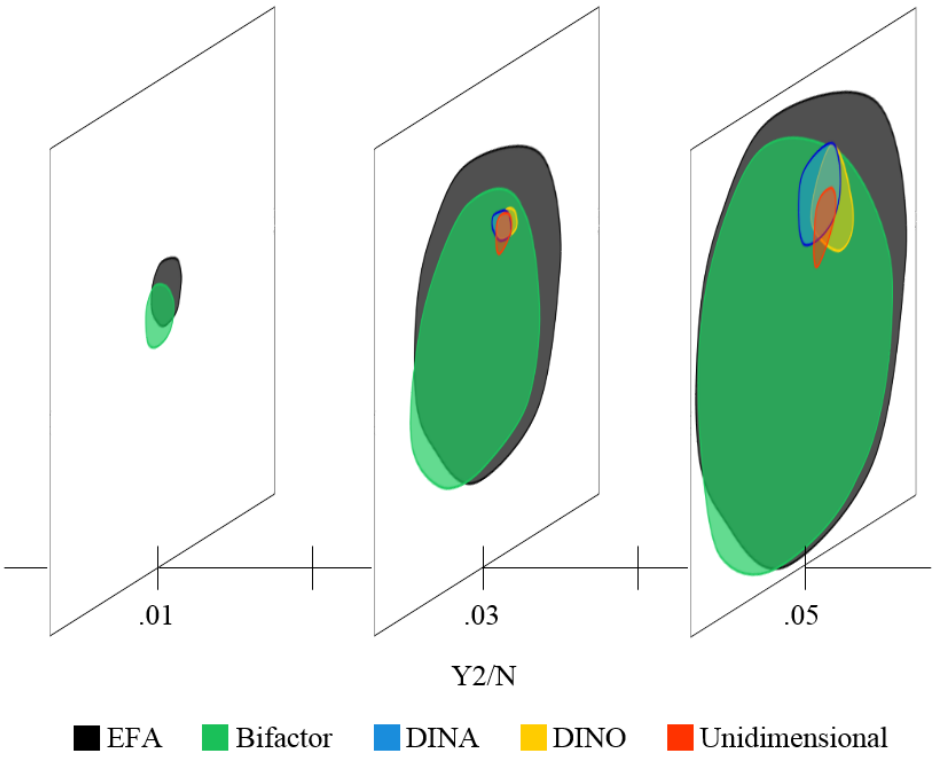


Figure 5.4. Hypothetical regions of the complete data space that are occupied by each model as Y2/*N* increases from .01 to .05. Regions drawn roughly to scale. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional model.

solution that best fits the data. The two diagnostic classification models had far lower fitting propensities and performed very similarly to one another with regard to Y2/*N*. The most counterintuitive finding is related to the unidimensional 3PL model. This model had an additional free parameter which should have supplied it with a superior ability to capture noise in the random data. And yet, the unidimensional model was, by far, the least inclined to fit well. Possible explanations for this will be discussed later.

Aside from shedding some light on a few common IRT models, these results

could guide the interpretation of the Y2/*N* statistic. As mentioned earlier, no cutoff criteria have been established for this statistic. The simulation results, particular those presented in the amoeba plots, indicate that a Y2/*N* cutoff of .01 is probably too low; the DINA, DINO, and unidimensional 3PL models did not fit a single data set at this level of Y2/*N*. At the opposite end of the spectrum, the Y2/*N* cutoff of .10 appears to be too lax—at this benchmark, the EFA model fit all but 8 of 1,000 data sets, the bifactor model fit all but 27 data sets, and overall, only 4 data sets eluded all models. Perhaps a Y2/*N* cutoff of .05 is more appropriate. At this degree of goodness-of-fit, the more flexible models (EFA and bifactor) tended to fit around 2/3rd of all data sets, while the less accommodating models (DINA, DINO, and unidimensional 3PL) tended to fit around 1/20th of all data sets. Thus, a Y2/*N* of .05 or lower is somewhat informative with regard to all of the models under investigation.

5.2 D^2 Latent Distribution Fit Index

Figure 5.5 illustrates the empirical cumulative percentage distributions of the D^2 latent distribution fit index in each model. It is clear from this figure that differences in D^2 were negligible, especially with regard to the EFA, bifactor, DINA, and DINO models, which overlapped so thoroughly that it is problematic to differentiate their curves at all. One could conceivably argue that the unidimensional 3PL model fit the latent distribution slightly worse, but this pattern persisted only through D^2 values of

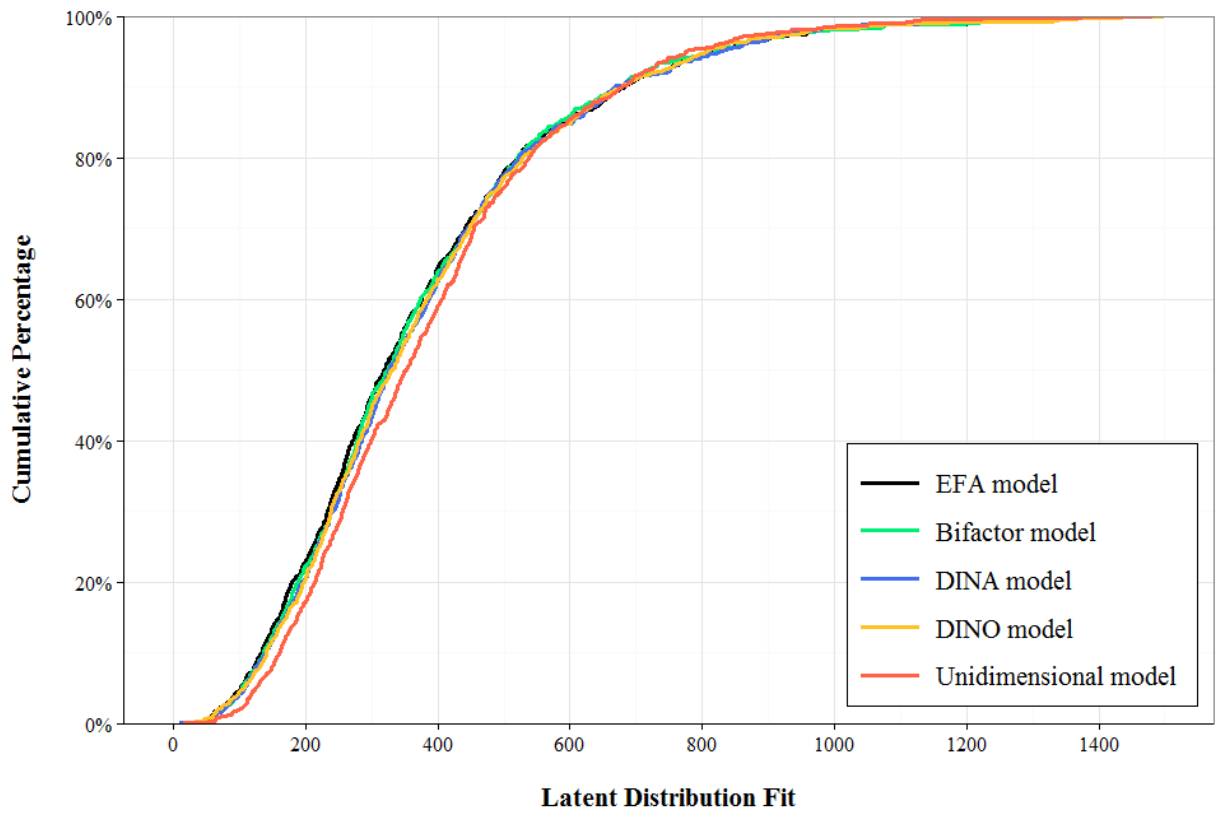


Figure 5.5. Cumulative percentage distributions of the D^2 latent distribution fit index in the exploratory factor analytic (EFA), bifactor, deterministic noisy input and-gate (DINA), deterministic noisy input or-gate (DINO), and unidimensional 3PL models.

approximately 500; beyond this range, the unidimensional model was indistinguishable from the multidimensional models.

Table 5.5 shows just how similar the latent distribution fit values were between the models. Excepting the unidimensional 3PL model, differences in the overall means ranged from 371.67 (EFA model) to 378.82 (DINA model) and differences in standard deviations ranged from 225.38 (DINO model) to 230.51 (EFA model). That is to say, these models were quite similar in terms of latent distribution fit, though the EFA model did have a slight edge on the other multidimensional structures. The

unidimensional model yielded a somewhat higher mean of 389.68 and a lower standard deviation of 214.05, though in absolute terms, it is unclear whether these differences were affected by the presence of an additional free parameter. The models in this analysis are not nested, so a formal statistical comparison cannot be drawn. Overall, however, the D^2 results uncovered the same pattern as the Y2/*N* results: the EFA model performed best, followed by the bifactor, DINA, and DINO models, while the unidimensional 3PL model offered the highest D^2 values.

A comparison of the within-model differences identifies a possible cause of the non-convergence in each model. The latent distribution fit of the unidimensional 3PL model was substantially lower when the model converged (*M* = 382.67, *SD* = 210.94) than when the model did not converge (*M* = 412.88, *SD* = 22.95). The DINA model also

Table 5.5. *Means and standard deviations of the* D^2 *latent distribution fit index in the exploratory factor analytic, bifactor, deterministic noisy input and-gate, deterministic noisy input or-gate, and unidimensional 3PL models.*

	EFA		Bifactor		DINA		DINO		Uni	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	371.67	230.51	372.89	228.00	378.82	228.52	377.60	225.38	389.93	215.24
Converged	371.69	234.74	373.19	225.72	377.77	223.47	377.79	225.08	381.89	212.38
Non-conv.	371.59	214.44	371.59	238.46	400.03	316.38	372.88	235.69	415.00	222.50
Difference	.11	20.30	1.60	-12.74	-22.26	-92.91	4.90	-10.61	-33.11	-10.12

Note. *N* = 1000; EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic noisy or-gate model; Uni = unidimensional 3PL model; Non-conv = non-converged analyses.

produced lower D^2 values when estimation converged (M = 377.77, SD = 223.47) than when estimation did not converge (M = 400.03, SD = 316.38). Perhaps the deterioration in latent distribution fit was the cause (or partial cause) of the non-convergence among these models. This theory remains to be explored.

Overall, the latent distribution fit values did not reveal any glaring discrepancies between the five models, and the amplified D^2 of the unidimensional 3PL model may have been influenced by the existence of an added free parameter. In future MDL-type examinations of IRT models, it appears that the computation and analysis of latent distribution fit may be uninformative.

5.3 *S-X*^{2} Item Fit Index

With regard to the Orlando-Thissen S-X^2 item fit index, the models were even less differentiable than in the exploration of latent distribution fit. For each model, Figure 5.6 illustrates the cumulative percentage distributions of S-X^2 for all seven items. In each item plot, all five curves are almost perfectly superimposed, such that there appears to be a single inverse exponential function in every frame. Tables B.1 through B.5 in Appendix B verify the equivalence between these curves. For each model, these tables present the mean S-X^2 values for each individual item, as well as the mean across all items.

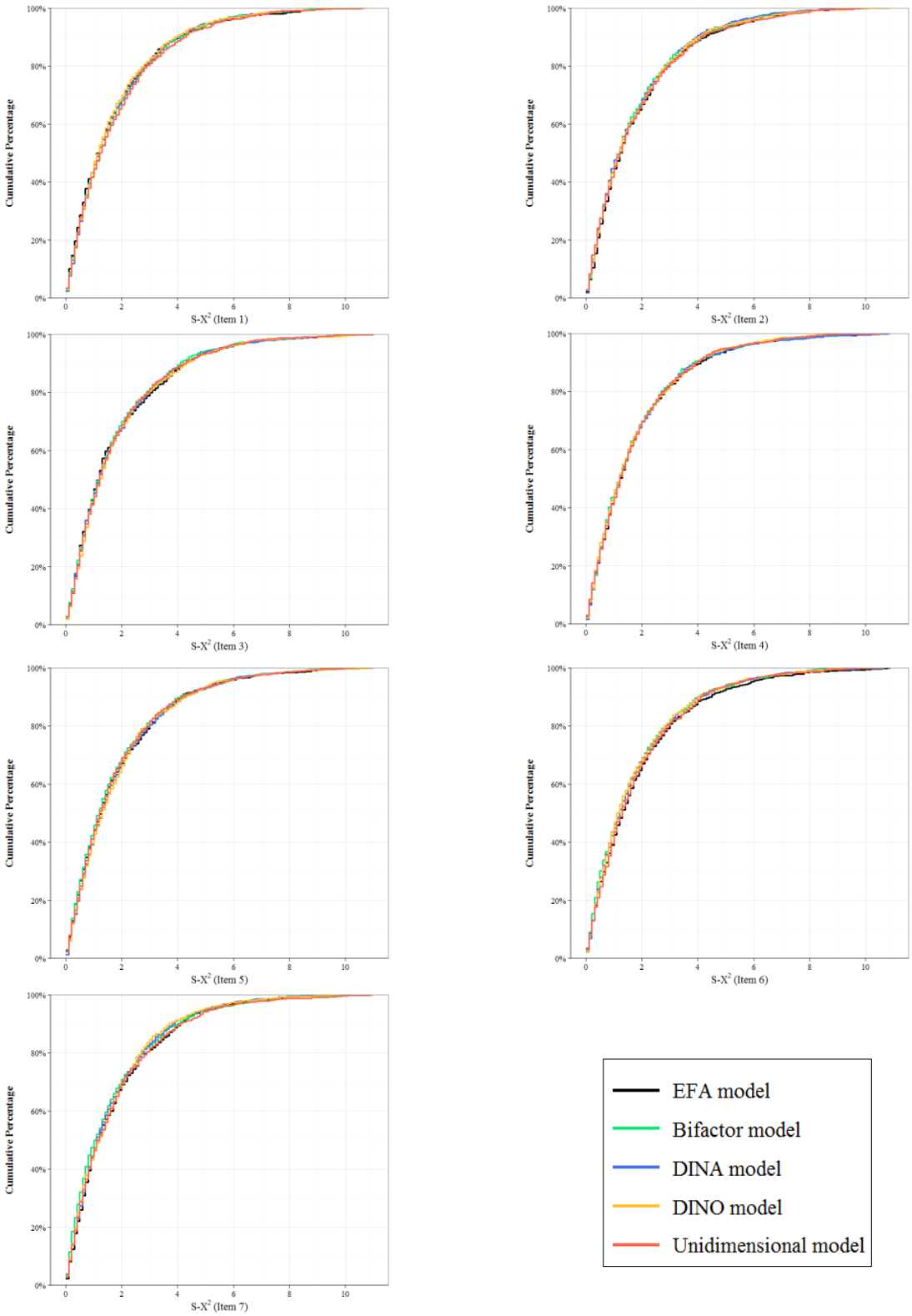


Figure 5.6. Cumulative percentage distributions of the *S-X*² item fit statistics for all items in the exploratory factor analytic (EFA), bifactor, deterministic noisy input and-gate (DINA), deterministic noisy input or-gate (DINO), and unidimensional 3PL models.

Table 5.6 includes the $S\text{-}X^2$ results, aggregated across all items in each model. By contrasting the complete data means, it is clear that the item fit statistics were unable to provide any information that might contribute to a deeper understanding of any between-model differences in fitting propensity. The means ranged from 1.79 in the DINA model to 1.84 in the EFA model—a difference of just .05. The standard deviations were also remarkably similar, with a range of .09. Further, there did not appear to be a coherent pattern related to estimation convergence; in general, the converged and non-converged results were comparable to one another and to the results derived from all data sets. In sum, the $S\text{-}X^2$ item fit index does not appear to be an enlightening metric of differences in fitting propensity.

Table 5.6. *Means and standard deviations of the* S-X^2 *statistic across all items in the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

	EFA		Bifactor		DINA		DINO		Uni	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.84	1.94	1.78	1.88	1.80	1.86	1.79	1.85	1.83	1.89
Converged	1.88	1.96	1.79	1.88	1.79	1.85	1.79	1.85	1.84	1.87
Non-conv.	1.70	1.84	1.73	1.90	1.88	1.94	1.80	1.88	1.78	1.92
Difference	.18	.12	.06	-.03	-.09	-.08	-.01	-.04	.06	-.05

Note. N = 1000; EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic noisy or-gate model; Uni = unidimensional 3PL model.

5.4 Marginal $\chi 2$

The marginal χ^2 results of each model are provided in great detail in Appendix B, Tables B.5 – B.10. To summarize these results, Table 5.7 displays the means and standard deviations of the marginal χ^2 values across all items, as well as the converged and non-converged results and the disparity between them. The between-model differences across all datasets confirm that, on average, the four multidimensional models performed nearly identically, while the unidimensional 3PL model had a higher mean marginal χ^2. Within models, there were slight differences between the converged and non-converged results, but there was a not a systematic pattern that might implicate the marginal χ^2 as the cause of the non-convergence.

On the surface, the marginal χ^2 results seem to tell much the same story (or lack

Table 5.7. *Overall means and standard deviations of the marginal χ^2 values in the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

	EFA		Bifactor		DINA		DINO		Uni	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.04	.30	.04	.30	.05	.30	.04	.23	.07	.38
Converged	.03	.27	.02	.26	.06	.31	.04	.23	.07	.41
Non-conv.	.07	.37	.08	.43	.00	.02	.01	.03	.04	.26
Difference	-.04	-.10	-.06	-.18	.05	.28	.04	.20	.03	.15

Note. N = 1000; EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic noisy or-gate model; Uni = unidimensional 3PL model; Non-conv = non-converged analyses.

thereof) as the D^2 and $S\text{-}X^2$ results from the previous sections. Aside from a few trivial deviations, the cumulative percentage curves of each item (displayed in Figure 5.7) seem uninteresting due to their similarity. However, despite the visual equivalence between the curves, there are a few meaningful findings that can be drawn from these results.

Notice that the *y*-axis in these plots has a lower limit of 80%; this indicates that the vast majority of data sets, regardless of the particular model, had marginal χ^2 values of exactly 0.0. That is, the observed and expected univariate marginal values from the IRT contingency tables were exactly identical in most cases. The first point of interest is that when marginal $\chi^2 = 0$, the unidimensional 3PL model (the red line) always reported the lowest curve. In other words, for every item, there were fewer data sets that produced identical marginals when fit to the unidimensional 3PL model. This is because the multidimensional structures that characterize the other models caused them to better represent the noise existent in the random "observed" data.

Another interesting trend is related to Items 6 and 7 in Figure 5.8. In these items, the bifactor model stood out from the other models by having a higher percentage of marginal χ^2 values that equaled 0.0. This outcome is related to the multifaceted structure of the bifactor model (as shown earlier in Figure 5.1(b)), wherein a latent factor is specifically included to account for the residual dependence between Items 6 and 7. By including this specific factor, the bifactor model outperformed the other

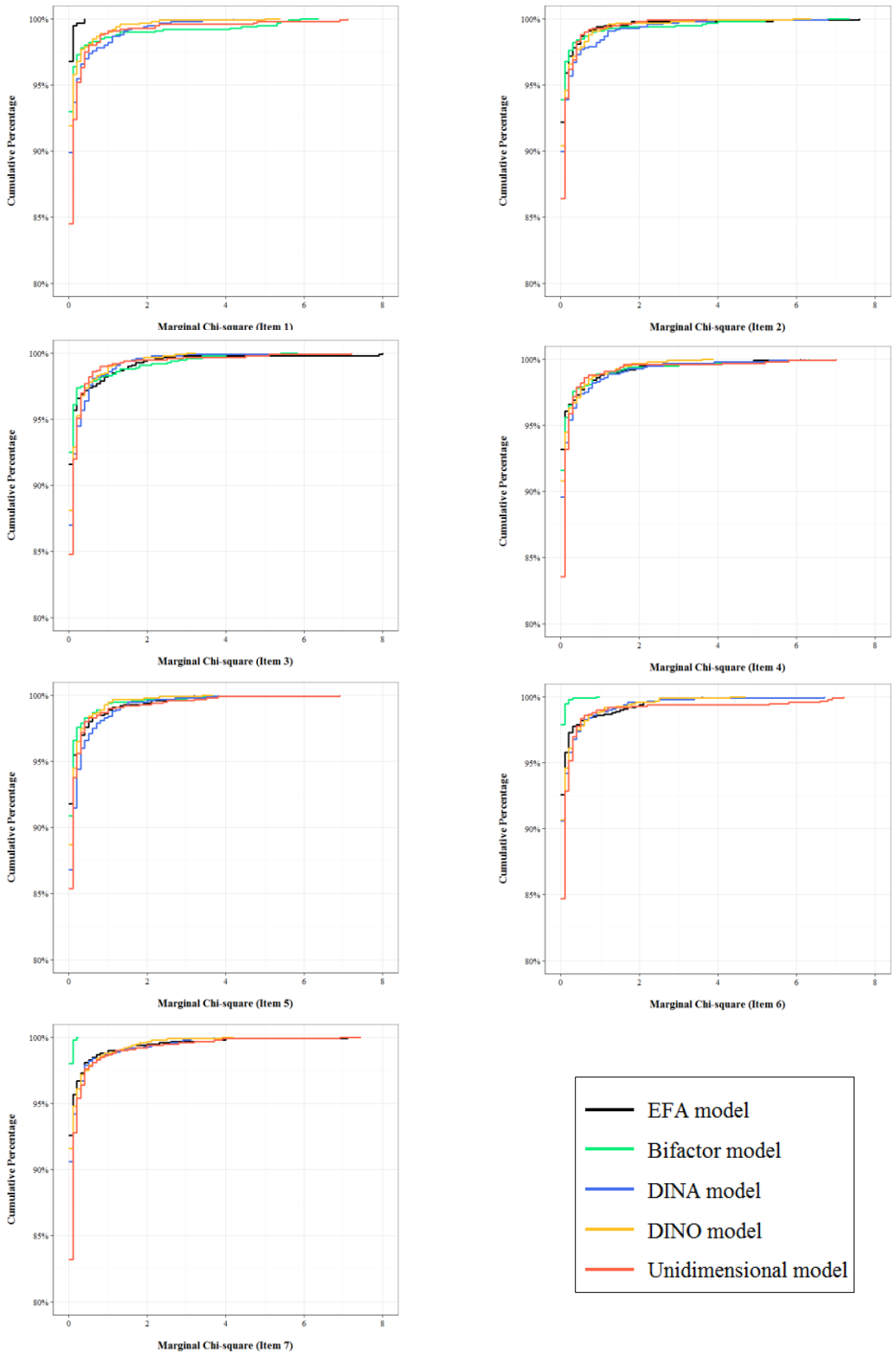


Figure 5.7. Cumulative percentage distributions of the marginal χ^2 values of all items in the exploratory factor analytic (EFA), bifactor, deterministic noisy input and-gate (DINA), deterministic noisy input or-gate (DINO), and unidimensional 3PL models.

models in analyses involving Items 6 and 7. A similar occurrence is shown for Item 1, though the outlying case in this instance is the EFA model rather than the bifactor. It is unclear why the EFA model tended to have lower marginal χ^2 values for Item 1. Perhaps this result is related to the fact that in the EFA model, Item 1 only loaded on the first factor, since its loading on the second factor was fixed at zero (as shown by the missing path in Figure 5.1(a)). This path was arbitrarily chosen to be fixed at zero for the purposes of model identification; if a different path had been chosen, the model would still be properly identified, but the outlying EFA marginal χ^2 values would likely have appeared in a different plot.

To further understand the nuances of the marginal χ^2 statistic for items 1, 6, and 7, the results were tabulated. Table 5.8 displays the percent of all data sets that attained marginal χ^2 values between 0.0 and .5 in each model. The first thing to notice is the high percentage of cases that had values exactly equal to 0.0. For item 1, the EFA model produced a marginal χ^2 of zero in 967 of 1,000 data sets, and over 99% of data sets had marginal χ^2 values less than or equal to .1. Further, the EFA model revealed no marginal χ^2 results higher than .4. The unidimensional 3PL model, on the other hand, produced fewer zero values (84.4%) for Item 1 and its maximum marginal χ^2 statistic was exceptionally high (9.0, to be precise).

Items 6 and 7 reinforce the effect of multidimensionality on the marginal χ^2. As mentioned earlier, the bifactor model was expressly designed to address local

dependence between items 6 and 7. Accordingly, for Item 6, there were 966 data sets that had marginal χ^2 values of precisely zero when fit to the bifactor model, and for Item 7, there were 971. For both items, over 99% of data sets showed marginal χ^2 results

Table 5.8. *Percent of 1,000 data sets that attained particular marginal χ^2 values for items 1, 6, and 7 when fit to the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

		Marginal χ^2					
	Model	= 0	≤ .1	≤ .2	≤ .3	≤ .4	≤ .5
Item 1	EFA	96.7	99.4	99.6	99.6	100.0	
	Bifactor	91.2	94.2	94.7	95.5	95.8	96.0
	DINA	89.8	93.6	95.4	96.5	96.9	97.3
	DINO	91.9	95.8	96.8	97.7	97.9	98.1
	Uni	84.4	92.3	95.1	96.2	97.4	97.9
Item 6	EFA	92.4	95.6	97.1	97.6	97.7	98.0
	Bifactor	96.6	99.5	99.6	99.6	99.7	99.8
	DINA	90.5	94.1	95.7	96.7	97.3	97.8
	DINO	90.7	94.6	96.1	97.0	97.5	97.8
	Uni	84.2	92.3	94.6	96.4	97.2	97.8
Item 7	EFA	92.5	95.6	96.6	97.2	98.0	98.2
	Bifactor	97.1	99.1	99.4	99.7	99.7	99.7
	DINA	90.5	94.1	96.0	96.6	97.8	98.1
	DINO	91.6	94.8	96.1	97.2	97.5	97.9
	Uni	83.2	92.8	95.4	96.4	97.6	97.8

Note. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

less than or equal to .1 when fit with the bifactor model. In sum, the dimensionality of a given IRT model has an impactful effect on the marginal χ^2 statistics of specific items, which may in turn influence the fitting propensity of that model.

5.5 LD X^2 Local Dependence Index

Table 5.8 includes for all models the means and standard deviations of the LD X^2 local dependence index, aggregated across all item pairs (the comprehensive results for every item pair are presented in Appendix B, Tables B.11 – B.15). This table reveals that the EFA and bifactor models were far better equipped to handle local dependence violations. The unidimensional 3PL model fared the worst, as usual. In addressing local dependence, the unidimensional model was handicapped by its meager functional form; the added free parameter did nothing to aid in decreasing the amount of local dependence. This table also confirms that there were inconsequential differences between the converged and non-converged results.

Figures 5.8 and 5.9 present the cumulative percentage distributions of the bivariate LD X^2 index across all item pairs. For a given pair of items, each plot represents the cumulative LD X^2 values across all 1,000 data sets. Note that in an effort to better depict these results, the lower limit of the y-axis in each graph was set at 60%. Upon inspection of these results, a clear trend emerges. As expected, the relatively flexible EFA model (the black curve) was adept at accounting for the local dependence

between all item pairs, and the unidimensional 3PL model (the red curve) was typically the least effective model for addressing local dependence.

There were several notable exceptions to this pattern of results. In 11 of the 21 item pairs, the bifactor model (the green curve) had greater success than the EFA model in handling the local dependence between items. Specifically, the bifactor model was better at capturing the noise caused by local independence violations in item pairs 2 & 1 through 5 & 4, as well as item pair 7 & 6. It is unsurprising that the bifactor model addressed the local dependence involved with these particular item pairs – the specific factors in this bifactor structure, as illustrated in Figure 5.1(b), were explicitly constructed to address the local dependence between Items 1 through 5 (Specific Factor 1) and Items 6 and 7 (Specific Factor 2). What is surprising is that the EFA model, which

Table 5.9. *Overall means and standard deviations of the LD-X^2 across all item pairs in the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noisy or-gate, and unidimensional 3PL models.*

	EFA		Bifactor		DINA		DINO		Uni	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	8.92	21.03	10.76	23.72	25.68	42.67	25.40	42.18	31.59	53.37
Converged	8.53	19.97	10.74	23.78	25.77	42.77	25.38	42.04	31.00	52.01
Non-conv.	10.37	23.98	10.84	22.92	23.86	39.98	25.95	44.11	33.44	56.53
Difference	-1.84	-4.01	-.10	.87	1.92	2.79	-.57	-2.07	-2.45	-4.52

Note. Means and standard deviations computed across all item pairs. EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

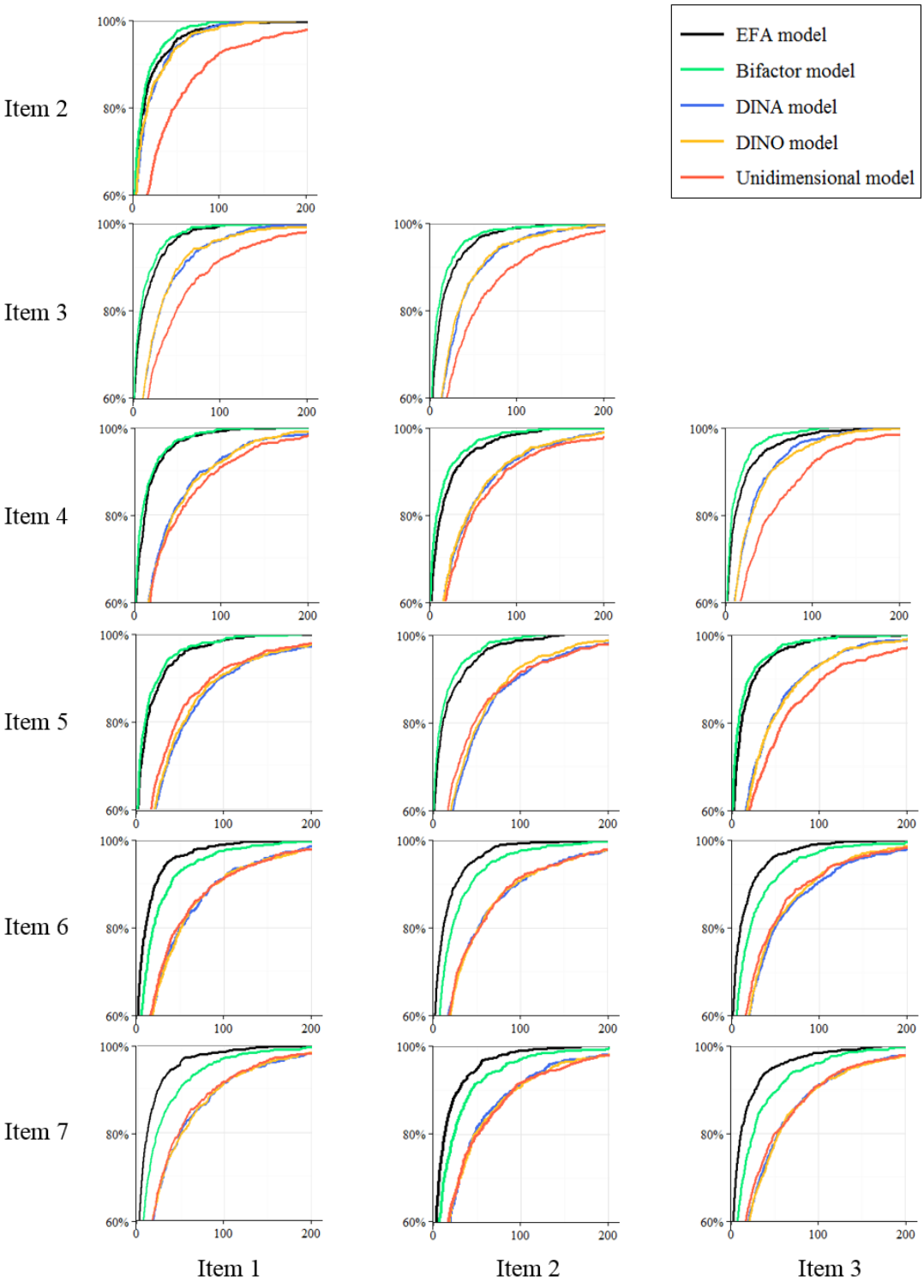


Figure 5.8. Cumulative percentage distributions of the bivariate LD X^2 index across all item pairs in the exploratory factor analytic (EFA), bifactor, deterministic noisy input and-gate (DINA), deterministic noisy input or-gate (DINO), and unidimensional 3PL models (continued in *Figure 5.9*).

allowed all items to load on both factors (save the path from Factor 2 to Item 1, which was fixed at zero to identify the model), was unable to account for the local dependence as successfully as the bifactor model in every situation. Perhaps the bifactor model's high fitting propensity is in part due to its heightened ability to model specific local dependence noise.

Another counterintuitive result from the LD X^2 analyses was the occasional failure of the (multidimensional) diagnostic classification models to manage local dependence violations as effectively as the undimensional 3PL model. Figure 5.8 reveals

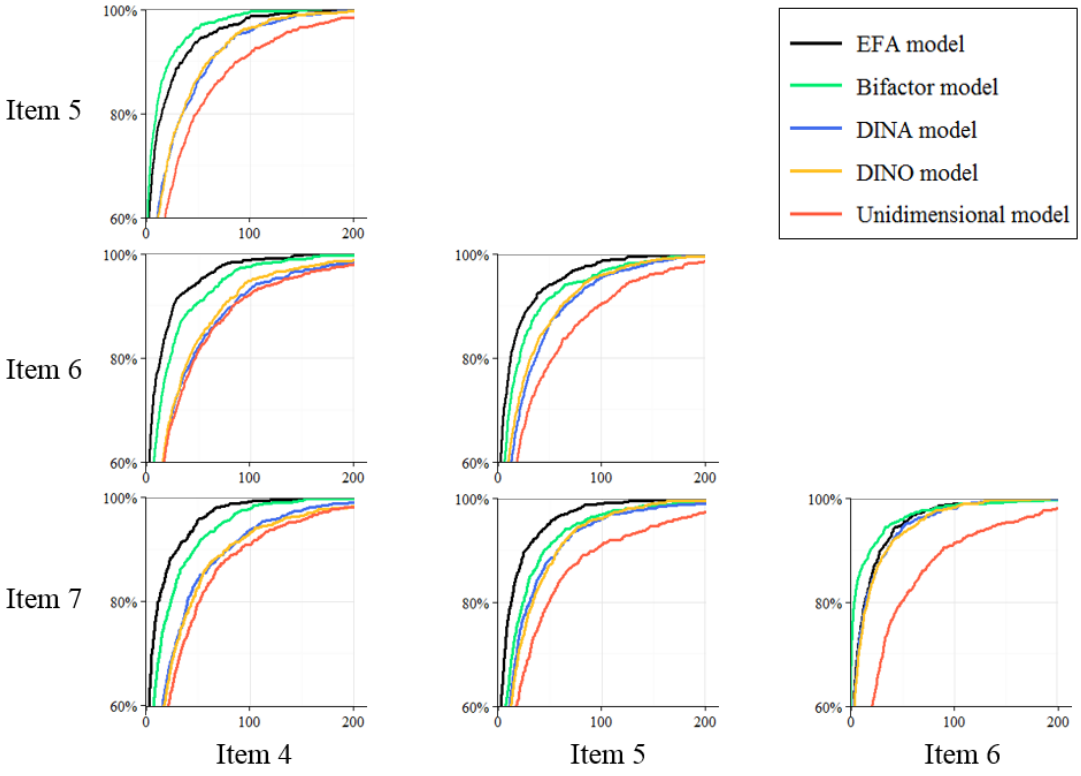


Figure 5.9. Cumulative percentage distributions of the bivariate LD X2 index across all item pairs in the exploratory factor analytic (EFA), bifactor, deterministic noisy input and-gate (DINA), deterministic noisy input or-gate (DINO), and unidimensional 3PL models (continued from *Figure 5.8*).

that among item pairs 5 & 2, 6 & 1, 6 & 2, and 7 & 2, the cumulative percentage distribution of LD X^2 in the unidimensional 3PL model overlapped with that of the DINA and DINO models, thereby indicating that all three models did an equally sufficient job of accounting for the local dependence between these items. In item pairs 5 & 1, 6 & 3, 7 & 1, and 7 & 3, the unidimensional model actually surpassed the diagnostic classification models in its capacity to model the local dependence among these particular item pairs. One possible culprit is the presence of the latent attribute variables in the DINA and DINO models. Because these latent factors are discrete, there was a loss of information that would not have occurred if the items were modeled with a continuous latent variable. Thus, in some cases, the higher-order factor that was employed to model the attribute space in the classification models did not perform as well as the single latent dimension that characterized the unidimensional 3PL model. This finding suggests that the multidimensionality that typifies the DINA and DINO models is not particularly well-suited for modeling local dependence between items.

Local dependence violations between certain items are often trivial enough to ignore. Non-ignorable local dependence can be identified by evaluating the absolute magnitude of each of the Chen-Thissen LD X^2 statistics against some critical value; Houts and Cai (2013) suggest 3.0 as an appropriate criterion. Thus, the column plots in Figure 5.10 depict for four example item pairs the number and percentage of LD X^2

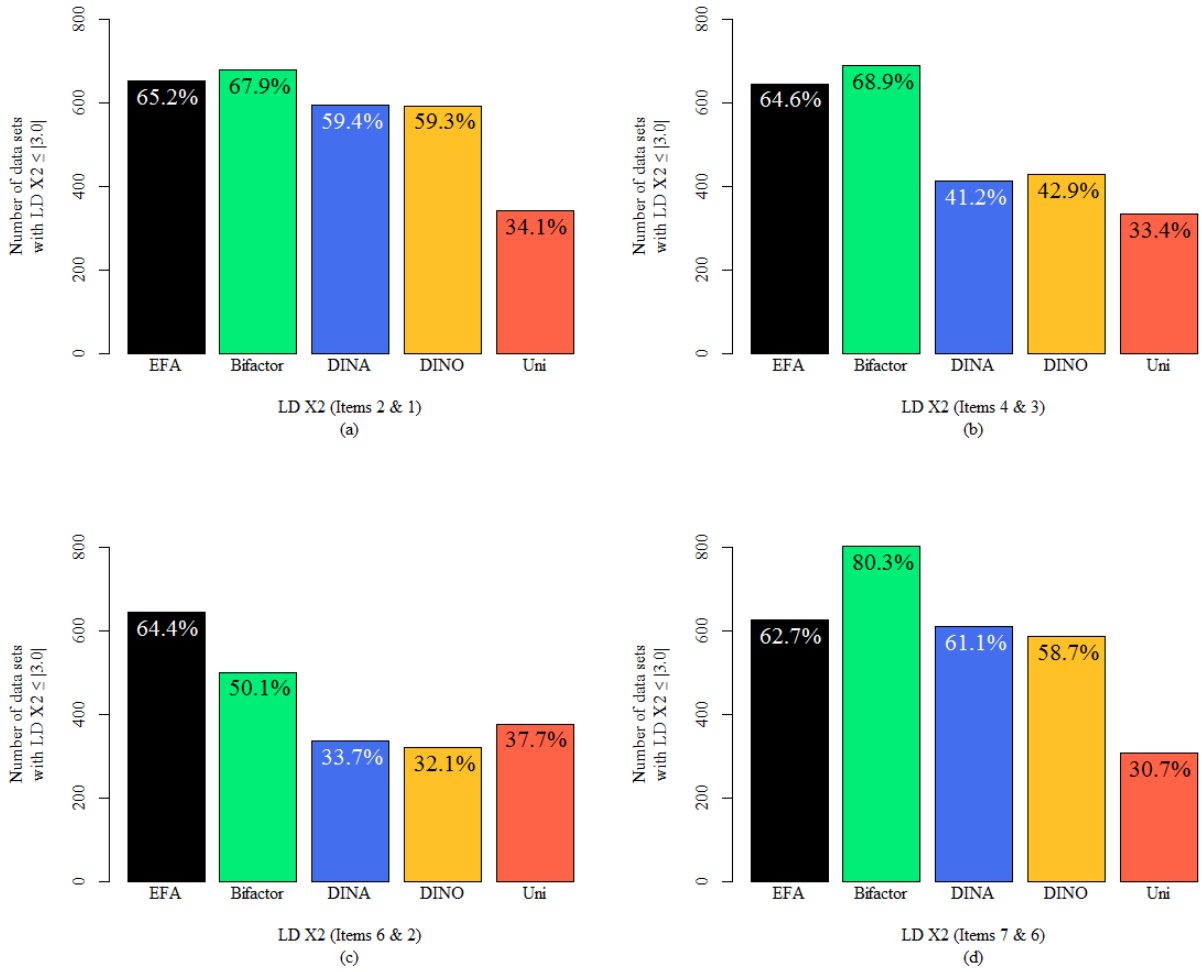


Figure 5.10. Number and percentage of 1,000 data sets that exhibited LD X^2 values ≤ |3.0| in the exploratory factor analytic (EFA), bifactor, deterministic input noisy and-gate (DINA), deterministic input noisy or-gate (DINO), and unidimensional 3PL (Uni) models for item pairs (a) 2 and 1, (b) 4 and 3, (c) 6 and 2, and (d) 7 and 6.

values less than or equal to absolute 3.0. Plot (a) shows the local dependence between Items 2 and 1. Here, all four multidimensional models were effective in reducing the local dependence violations to acceptable LD X^2 levels in approximately 600-680 of the 1,000 data sets, while the unidimensional 3PL model performed expectedly worse. In plot (b), the bifactor and EFA models were just as well-equipped to diminish the local

dependence between Items 4 and 3 as they were in plot (a). The two diagnostic classification models, however, were only able to produce LD X^2 statistics below 3.0 in approximately 42% of the data sets. This result may seem a bit unexpected—in the DINA and DINO models, Items 4 and 3 were both explained by Attribute 2, so one would anticipate a greater reduction in local dependence. However, Item 3 was also associated with Attribute 1; this cross-loading (or "interaction effect") seems to have adversely affected the ability to curb the dependence between these items.

Plot (c) of Figure 5.10 differs from the others in two key ways. First, this plot shows that for item pair 6 and 2, the EFA structure was more effective than the bifactor model at yielding acceptable LD X^2 statistics. Further, the unidimensional 3PL model outperformed both of the multidimensional diagnostic classification models. Overall, the local dependence between Items 6 and 2 was among the most difficult to model; the directional paths of the bifactor and diagnostic structures (as shown in Figure 5.1) were not arranged in a manner conducive to modeling the residual dependence between these particular items. Despite this fact, however, the bifactor model still outperformed the DINA and DINO models.

The final plot in Figure 5.10 illustrates the extent of local dependence between Items 7 and 6. The bifactor model, which included a specific factor that was explicitly intended to explain the residual noise generated by this exact item pair, was unsurprisingly masterful at addressing this dependence. Over 800 of the 1,000 data sets

exhibited LD X^2 values less than or equal to 3.0 when fit with the bifactor model. Of additional interest in plot (d) is the fact that the two diagnostic classification models were almost as successful as the EFA model with regard to reducing the LD X^2 index to a reasonable value. This is perhaps related to the structure of the diagnostic models, wherein Items 6 and 7 both load on one (and only one) attribute factor.

In general, Figure 5.10 underscores the high fitting propensity of the EFA model and the low fitting propensity of the unidimensional 3PL model. Notice that in all four example item pairs, the EFA model was able to reduce the LD X^2 values to tolerable levels in approximately 2/3rds of the data sets, while the unidimensional 3PL model consistently addressed the local dependence in approximately 1/3rd of all data sets. There was some degree of variability in the local dependence management of the bifactor model, though this structure typically addressed the violations quite effectively. The DINA and DINO models were the least consistent, sometimes capturing local dependence noise nearly as well as the EFA and bifactor models, yet occasionally functioning even less effectively than the unidimensional model.

5.6 Overview of Results

Table 5.10 provides a summary of the Y2/N, D^2, S-X^2, marginal χ^2, and LD X^2 results. On average, the Y2/N and LD X^2 indices ranked models in the same sequence: for each of these metrics, the lowest means and standard deviations were returned by the EFA

model, followed closely by the bifactor model, then the DINA and DINO models, and finally the unidimensional 3PL model. The D^2 latent distribution fit index means also expressed this pattern, but the differences between the multidimensional models were not especially impressive. Finally, the means and standard deviations of the $S\text{-}X^2$ and marginal χ^2 values failed to expose any differences between the models (though an in-depth examination of the marginal χ^2 results did turn out to be quite informative).

Table 5.10. *Overall summary of the means and standard deviations of* Y2/N, D^2 *latent distribution fit,* $S\text{-}X^2$*, marginal* χ^2*, and LD* X^2 *statistics in the exploratory factor analytic, bifactor, deterministic input noisy and-gate, deterministic input noise or-gate, and unidimensional 3PL models, across all data sets and all items/item pairs.*

	Y2/*N*		D^2		$S\text{-}X^2$		Marginal χ^2		LD X^2	
Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
EFA	.04	.02	371.67	230.51	1.84	1.94	.04	.30	8.92	21.03
Bifactor	.05	.02	374.08	228.20	1.80	1.90	.05	.36	10.88	23.85
DINA	.10	.04	378.82	228.52	1.80	1.86	.05	.30	25.68	42.67
DINO	.10	.04	377.60	225.38	1.79	1.85	.04	.23	25.40	42.18
Uni	.13	.05	389.68	214.05	1.82	1.88	.06	.39	31.43	53.08

Note. Means and standard deviations computed across all items and all data sets (converged and non-converged). EFA = exploratory factor analytic model; DINA = deterministic input noisy and-gate model; DINO = deterministic input noisy or-gate model; Uni = unidimensional 3PL model.

CHAPTER 6

On the Complexity of IRT Models:

Discussion

It is known that one model may fit the observed data better than another because it has a more flexible functional form or a greater number of estimated parameters (e.g., Collyer, 1985; Cutting, Bruno, Brady, & Moore, 1992). The present study investigated five IRT models that differed in functional form: an exploratory factor analytic model, a bifactor model, a deterministic input noisy and-gate model, a deterministic input noisy or-gate model, and a unidimensional model. All items in the multidimensional models were fit with a 2PL logistic function, resulting in exactly 20 freely estimated parameters in each model. Each item in the unidimensional model was fit with a 3PL logistic function, resulting in 21 freely estimated parameters. Thus, the unidimensional model was simpler in functional form, but more complex in the number of parameters. All five models were fit to 1,000 data sets that were randomly and uniformly sampled from the complete data space. The models were then compared with respect to five statistics intended for categorical data analysis; the cumulative results of these statistics across all data sets functioned as indicators of each model's inherent propensity to fit any possible data.

6.1 Confirmation of Hypotheses

My first hypothesis posited that the EFA model would exhibit, on average, the highest fitting propensity. This prediction was strongly supported by the results. The analyses confirmed that among the candidate structures, the EFA model had the most pliable functional form. Specifically, the Y2/N, D^2, marginal χ^2, and LD X^2 results[2] demonstrated that the EFA model outperformed its competitors in terms of overall model fit, latent distribution fit, recovery of observed response probabilities, and minimization of local dependence violations. This outcome is unsurprising; the exploratory nature of the EFA model means that it is exceedingly adaptable to a wide array of data patterns. This model was included in the study, not to shed new light on the flexibility of an exploratory model, but to serve as a baseline measure of fitting propensity.

The second hypothesis predicted that the bifactor model would display higher fitting propensity than the two diagnostic classification models. This hypothesis was also confirmed by the results: the bifactor model, relative to the DINA and DINO models, had a propensity to fit a greater number of random data sets that were uniformly distributed across the entire data space. In fact, as evidenced by the

[2] The S-X^2 item fit statistic was unable to expose any meaningful differences between the models.

cumulative Y2/N, D^2, marginal χ^2, and LD X^2 metrics, the bifactor model, when fit to random data, was almost as accommodating as the EFA model. Moreover, the amoeba plots (Figures 5.2 and 5.3) uncovered a small number of data sets that actually fit the bifactor model *better* than the EFA model. These findings help to explain the growing popularity of the bifactor model—in model comparison studies that rely solely on goodness-of-fit to the observed data, the highly malleable bifactor model will almost always be chosen as the "best" model. The researcher who employs this model runs the risk of overfitting the data.

6.2 The Importance of Functional Form

The hypotheses discussed above addressed the superior fitting propensities of two particular IRT models. Both the EFA and bifactor models are characterized by relatively complex functional forms. In each case, the items are modeled using numerous cross-loadings on multiple latent dimensions. It is no wonder that these multifaceted models were able to closely represent a substantial proportion of the random data sets. Far less foreseeable were the outcomes returned by the model with the simplest functional form.

The unidimensional 3PL model consistently demonstrated the weakest fitting propensity. The overall model fit results from the Y2/N analysis verified that the unidimensional model struggled to recover the univariate and bivariate marginals of

the "observed" random data. The LD X^2 results revealed that the unidimensional structure was ineffective with regard to modeling local dependence. The D^2 index showed that the unidimensional model was also ill-equipped to account for nonnormality in the distribution of theta scores. Each of the key results indicated weak fitting propensity, *despite the fact that the unidimensional 3PL model included an additional free parameter!*

This finding challenges current notions of IRT model complexity. It suggests that model complexity should not be assessed simply by tallying free parameters; instead, discussions of IRT model complexity should concentrate on the arrangement of the latent variables and structural paths in the model. Measurement researchers should be cautious when using models that are not parsimonious in form (the number of parameters notwithstanding). Models that incorporate multiple latent dimensions, residual factors, cross-loadings, or similar intricacies may have an innate tendency to fit well to any conceivable data, even if such models involve fewer freely estimated parameters.

6.3 Limitations

This study was limited primarily by computational issues. The first limitation relates to our representation of the complete data space. Due to computational burden, I opted to generate only 1,000 data sets that were randomly sampled from and uniformly

distributed across the complete data space. If I had generated more data sets, say, 10,000, then the proxy data space would be even more representative of the actual entirety of the data space. However, this limitation was not debilitating; even with 1,000 random data sets, I was able to demonstrate clear discrepancies between the intrinsic data-fitting capabilities of each model.

The estimation specifications of this study were also limiting. The E-step tolerance of the EM algorithm was relaxed in order to speed up the estimation process. Despite this modification, the random data-fitting procedure was still rather time-consuming. For example, fitting the unidimensional 3PL model to all 1,000 data sets took approximately 30 hours when using a 2.90GHz quad-core processor with 16 GB RAM. If the tolerance between E-step iterations had been left at the default, then estimation would have taken considerably longer.

Furthermore, the estimation process was unable to converge on stable parameter estimates in a sizeable number of data sets (Table 4.2), despite the considerable increase in EM iterations (20,000 cycles). Perhaps with an even greater number of estimation cycles, a different estimator, additional computing time, or other alterations to the estimation process, the convergence rates would improve. However, the models were fitting random nonsensical data with no underlying form; in the cases where data were more noise than signal, one would not expect successful convergence. Thus, while convergence rates may not have been ideal, it is highly unlikely that 100% convergence

across all models and data sets would ever be achieved.

Another potential limitation was the analytic approach itself. I followed the same investigative strategy as Preacher (2006), namely, fitting candidate models to a large number of random data sets. While this tactic produced several compelling findings, alternative formulations of the MDL principle may offer deeper insights into these (and other) IRT models, and by circumventing the tedium of fitting 1,000 data sets, they would likely present these insights with far greater efficiency. A few promising MDL expressions are discussed below.

6.4 Future Research Directions

This line of inquiry opens up a number of promising topics for future research. First, the hypotheses in this study drew attention to the overly flexible nature of the particular EFA and bifactor models that were included in the analysis. In the exact EFA model that was analyzed, the path from Factor 2 to Item 1 was constrained to zero for model identification. The choice to fix this specific path was completely arbitrary. Since this model was to be fit to random data, my thinking was that one EFA structure would be as useful as any other. Yet, could it be that the fitting propensity exhibited by this EFA model was elevated (or diminished) by the chosen arrangement of the variables? How might the outcome compare if, for instance, a path had been fixed from Factor 1 instead of from Factor 2? The same type of question arises when considering the bifactor

results. Would the findings have shifted if other sets of items had been selected to load on the specific factors? In the future, it would be prudent to compare all combinations of factor loadings in these models. Such an all-encompassing analysis would permit one to make claims about the EFA and bifactor models on the whole, rather than simply reporting results that are contingent on particular instantiations of these models.

Another direction of future research relates to the Y2/*N* amoeba plots. The various fitting propensity regions depicted in these figures exposed several interesting nuances. For example, what sort of data patterns characterize the few data sets that fit the bifactor model better than the EFA model? Further, the DINA and DINO models fit approximately the same number of data sets, but these two models did not occupy identical regions of the data space. Is it possible to isolate the type of response pattern that tends to fit better to the DINA model than to the DINO model, or vice versa? The Y2/*N* results, especially in the information-theoretic context of "occupying the complete data space," offer ample fodder for future research.

The simulation study presented herein dealt exclusively with dichotomous item response data. The MDL approach used in this analysis could easily be extended for use in investigations of polytomous response data. For example, two popular models for data with multiple response categories—the graded response model (Samejima, 1969) and the generalized partial credit model (Muraki, 1992)—have the same number of free parameters. However, some researchers have noticed that the graded model seems to fit

data better. As Thissen and Wainer (2001) observed,

> In our experience, fitting hundreds of data sets over two decades, it has almost always been the case that the graded model fits rating data better than does the generalized partial credit model. (p. 151)

An MDL examination of the fitting propensities of these two polytomous IRT models could either support or reject this anecdotal evidence. The findings could also inform the development of large scale education assessments like NAEP, which utilizes both of these polytomous models. If it turns out, for instance, that the graded model has a higher tendency to fit any potential data, then the NAEP developers may wish to use it more sparingly.

The MDL approach could also be used to investigate other latent variable models. The diagnostic classification literature, for instance, comprises a number of latent class-type models (e.g., the LCDM, NIDA, and NIDO models, among many others; see Rupp, Templin, and Henson (2010) for details). Although the DINA and DINO models included in the present study appeared to be quite similar in terms of fitting propensity, a more comprehensive MDL-based overview of the common diagnostic classification models may reveal that some tend to fit random data better than others. Indeed, one could invoke the MDL principle to investigate the data fitting properties of any statistical models, so long as they are of the same class (Rissanen, 2007).

Another future topic of study concerns the role of MDL in substantive

psychological research. A case in point involves the nature of psychopathology. Pickles and Angold (2003) wrote,

> For many years a debate has raged over whether child and adolescent psychopathology should be regarded as consisting of a series of categorical phenomena (with individuals being either cases or noncases of various disorders) or as dimensions with psychopathology being just their negative extremes. (p. 529)

In the child psychopathology literature, some researchers are steadfast in their belief of discrete diagnostic categories (e.g., Sonuga-Barke, 1998) while others argue that psychopathology is a continuum (e.g., Achenbach, 1966). As Lubke and Neale (2006) noted, "... the distinction between categorical and continuous latent variables can be of considerable importance on a theoretical level" (p. 500). One way to distinguish between the latent class perspective and the continuous latent trait perspective would be to assess the proclivity of each approach to fit well to any given data. The less informative of these two perspectives would be that which is inclined to represent any possible data. Based on the performance of the diagnostic classification models in the present study, one would expect categorical latent trait models of child psychopathology to exhibit lower fitting propensities than their continuous counterparts. This hypothesis could be investigated using the principle of MDL, and the results would inform this theoretical debate.

The current study explored the theory of MDL by fitting several IRT models to many random data sets; an appealing extension of this work would focus on the

feasibility, suitability, and effectiveness of MDL in empirical data analysis. Consider a test with a large number of items; a dichotomously-scored test of just 20 items would have 2^20 = 1,048,576 possible response patterns. Conducting an exhaustive information-theoretic evaluation of a model in this data space would require a tremendous amount of computing power and time. To overcome this obstacle, I propose the use of restricted MDL (MDL-R) in the analysis of real data. Rather than considering any *possible* response patterns, MDL-R would instead consider any *plausible* patterns, where "plausible" is operationally defined as "existing in the observed data." Thus, the observed data can be thought of as one particular instantiation out of all possible data sets. For example, administration of a 20-item test may yield a manageable number of response patterns. If 1,000 test takers produced 500 unique 20-item response patterns, then by definition, 1,048,576 – 500 = 1,048,076 possible patterns were not produced. If these observed and non-observed patterns were combined into a single data set, over one million patterns would have frequencies of zero and the remaining 500 would occur at the frequencies seen in the observed data. While a full MDL analysis would incorporate all possible patterns, MDL-R would instead consider only the 500 plausible patterns produced by the sample. From there, the evaluation would follow the same procedure as discussed above, albeit with enormously improved efficiency.

Another important future direction involves the study of various numerical expressions of MDL, especially in the context of IRT. One such expression is known as

the Fisher information approximation (FIA; Rissanen, 1996):

$$\mathrm{FIA} = -\ln f\left(\mathrm{D}\middle|\widehat{\boldsymbol{\theta}}^{*}(\mathrm{D})\right) + \frac{k}{2}\ln\frac{N}{2\pi} + \ln\int_{\boldsymbol{\theta}}\sqrt{|\mathrm{I}(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}\,. \tag{6.1}$$

The first term in this equation accounts for goodness-of-fit, where $f(\cdot)$ is the maximum likelihood function of the observed data D. The second term addresses parametric complexity, where k is the number of free model parameters and N is the sample size. The final term accounts for structural complexity, where $|I(\boldsymbol{\theta})|$ is the determinant of the Fisher information matrix of the parameters in $\boldsymbol{\theta}$. In sum, FIA emphasizes the generalizability of a model by accounting for its goodness-of-fit to the data as well as its parametric and structural complexity. Preacher (2006) noted that the integral in Equation 6.1 is difficult to compute because of the highly-parameterized models that are common in SEM; this problem may not exist for certain IRT models.

A similar expression of MDL that is perhaps more appropriate for IRT is the stochastic information complexity (SIC; Hansen & Yu, 2001; Markon & Kreuger, 2004; Rissanen, 1989):

$$\mathrm{SIC} = -\ln f\left(\mathrm{D}\middle|\theta^{*}(\mathrm{D})\right) + \frac{1}{2}\ln\left|N\cdot\ I\left(\hat{\theta}\right)\right|\,. \tag{6.2}$$

Here, the first term is identical to that of FIA, but the second term lacks the computationally difficult integral. To calculate the SIC of a given model, one must compute the log determinant of the covariance matrix that results when the Fisher information matrix is used to estimate standard errors. This method seems especially

well-suited for future IRT analyses.

6.5 Conclusion

Overall, this report presents a novel outlook on the complexity of IRT models. I demonstrated that the bifactor model has an undesirable tendency to fit any possible data, such that it even outperforms the EFA model when both are fit to certain data sets. Further, I found that that an IRT model with more free parameters but a simpler structure may occupy a much narrower region of the complete data space. These findings establish the MDL principle as a promising methodological tool for understanding the inherent properties of all types of latent variable models. I believe that this work makes a major contribution to quantitative psychology, not only by exposing the vices and virtues of several popular IRT models, but by opening up a plethora of new areas of philosophical, theoretical, and practical research in all types of latent variable modeling.

CHAPTER 7

Use of the Framework

The integrative framework presented in Chapter 3 is not merely a collection of model evaluation techniques. It is a blueprint, designed to guide researchers in making comprehensive, well-founded, and defensible appraisals of their models. The intent is not for an individual to choose a frequentist method *or* a Bayesian technique *or* an information-theoretic approach to model evaluation. Instead, the framework allows one to contemplate a given model from three distinct viewpoints and, by weighing the pros and cons of each perspective, to form an overall impression of the model.

7.1 Empirical Example

7.1.1 The Data

To demonstrate the use of the framework, I will consider empirical data from the Program for International Student Assessment (PISA). The data come from a random sample (N = 1,000) of the 25,000+ students who completed the mathematics portion of the Booklet 8 of the 2000 PISA. For the purpose of illustration, I analyzed only the 11 dichotomous items that make up the 15 mathematics items in Booklet 8. Among these 11 items, there are three testlets, which are referred to in the PISA technical reports as

Apples, Growing Up, and Racing Car (Adams & Wu, 2002). After removal of the polytomous items, these testlets were characterized by two items, two items, and four items, respectively. None of the remaining items (Triangles, Carpenter, and Pipelines) were affiliated with a testlet. Due to the testlet format of Booklet 8, previous researchers have fit these data with various multidimensional structures, including a bifactor model for all 15 items (Cai, Yang, & Hansen, 2011). The MDL findings presented in the previous chapters, however, explicated the tendency of the bifactor model to overfit the data. Thus, to illustrate the utility of the model evaluation framework, suppose instead that a researcher wishes to determine whether a unidimensional 3PL model is appropriate for these data.

7.1.2 Frequentist Evaluation

A traditional frequentist analysis of the PISA data was conducted in flexMIRT (Cai, 2013), using the same estimation specifications discussed earlier (Section 4.3.2). This analysis resulted in the 3PL item parameter estimates and standard errors shown in Table 7.1. All items were significantly discriminating, with slopes ranging from .86 to 2.61 and standard errors ranging from .13 to .52. The middle column in this table reveals that the PISA items covered, in fairly even intervals, a breadth of difficulty from relatively easy (e.g., Racing Car Q1, b = -1.68; Racing Car Q2, b = -1.52) to relatively difficult (e.g., Racing Car Q5, b = .92; Carpenter, b = 1.30). The third column showcases

Table 7.1. *Frequentist estimates of the item parameters and standard errors of the unidimensional 3PL model of the PISA data.*

Item	Discrimination	Difficulty	Lower Asymptote
Apples Q1	2.02	.11	.14
	[.35]	[.11]	[.05]
Apples Q2	2.61	.62	.01
	[.44]	[.08]	[.03]
Growing Up Q1	1.29	-.57	.11
	[.20]	[.28]	[.12]
Growing Up Q3	1.12	-.03	.07
	[.19]	[.24]	[.09]
Racing Car Q1	.86	-1.12	.09
	[.13]	[.33]	[.12]
Racing Car Q2	2.19	-1.52	.11
	[.33]	[.20]	[.13]
Racing Car Q3	1.91	-1.68	.14
	[.27]	[.28]	[.15]
Racing Car Q5	1.68	.92	.03
	[.29]	[.10]	[.04]
Triangles	1.05	-.23	.21
	[.20]	[.47]	[.14]
Carpenter	2.06	1.30	.10
	[.52]	[.09]	[.02]
Pipelines	1.51	.40	.29
	[.44]	[.22]	[.08]

Note. N = 1,000. Standard errors are bracketed. $M_2(33) = 61.11$, $p = .002$, $RMSEA_2 = .03$.

the lower asymptote parameter estimates that typify a 3PL model; most of these pseudo-guessing parameters were near zero. The Pipelines item exhibited a sizeable lower asymptote parameter (g = .29, SE = .08), indicating a moderately high probability of providing a correct response, even at the lower limit of mathematics ability. The lower asymptote estimates of the Apples Q1 (g = .14, SE = .05) and Carpenter (g = .10, SE = .02) items did not seem particularly high, but their comparatively small standard errors confirmed that these items also had significantly non-zero lower asymptotes.

As discussed in Chapter 2, a vital component of frequentist model evaluation is goodness-of-fit to the observed data. To assess the global fit of the unidimensional 3PL model to the PISA data, the M_2 index (Maydeu-Olivares & Joe, 2005) was computed. This limited-information fit statistic, derived from the univariate and bivariate moments of the parameter vector $\boldsymbol{\Theta}$, is a common metric of overall fit in IRT models. This test statistic is evaluated against a chi-square distribution, and the resulting p-value reflects the degree of perfect model-data fit.

To assess less-than-perfect, but still excellent fit, Maydeu-Olivares and Joe (2014) proposed the use of the bivariate root mean square error of approximation ($RMSEA_2$) based on the M_2 statistic:

$$\mathrm{RMSEA}_2 = \sqrt{\frac{\mathrm{M}_2 - df_2}{N \times df_2}}. \tag{7.1}$$

Adequate model fit is indicated by $RMSEA_2$ values less than or equal to .05 / (K – 1),

where K is the number of response categories (Maydeu-Olivares & Joe, 2014). In the dichotomous case, $K = 2$, and satisfactory model fit is therefore represented by $RMSEA_2 \leq .05$. The $RMSEA_2$ value is an estimate that is subject to sample-specific fluctuations. Thus, a 95% confidence interval can be computed around the $RMSEA_2$ estimate by:

$$\left(\sqrt{\frac{\hat{L}_2}{N \times df_2}}, \sqrt{\frac{\hat{U}_2}{N \times df_2}} \right), \tag{7.2}$$

where $\hat{L}_2$ and $\hat{U}_2$ are the noncentrality (λ) parameters of the noncentral chi-square distribution function $F_{\chi^2}(M_2, df, \lambda)$ that result in $F_{\chi^2} = .025$ and $F_{\chi^2} = .975$, respectively (Maydeu-Olivares & Joe, 2014).

In the PISA data, the M_2 test statistic was 61.11 on 33 degrees of freedom, $p = .002$, meaning the unidimensional 3PL model did not fit the data with exact precision. The $RMSEA_2$ estimate value of .03, however, was below the recommended threshold and the 95% confidence interval around the true RMSEA parameter was (.015, .042). Thus, the unidimensional 3PL model, while not a perfect representation of the PISA mathematics data, was deemed to have acceptable fit. Overall, a frequentist evaluation of the unidimensional 3PL model would recognize this structure as an informative representation of the PISA data. The model revealed highly discriminating slopes on most items and demonstrated adequate absolute goodness-of-fit according to the M_2 test statistic.

7.1.3 Bayesian Evaluation

Bayesian analysis of the PISA data was carried out using SAS software (SAS Institute, 2011). To specify the Markov Chain Monte Carlo (MCMC) simulation, the discrimination parameters were assumed to fit a lognormal distribution and the difficulty parameters were assumed to be normally distributed. The lower asymptote parameters were specified just as in the frequentist analysis, with a Beta(1.0, 4.0) prior. Table 7.2 presents the item discrimination, difficulty, and lower asymptote parameter estimates produced by SAS Proc MCMC. Here, the maximum likelihood estimates from the frequentist evaluation were used as starting values; thus, the results in Tables 7.1 and 7.2 closely resemble one another. Both approaches identified the PISA items as highly discriminating across a wide range of difficulty. The only meaningful difference between these approaches was a subtle one: in the Bayesian analysis, the Triangles item was not found to have a significantly non-zero pseudo-guessing parameter.

To further explore the Bayesian mode of evaluation, model checking was performed via computation of the posterior predictive distribution. Model checking allows one to evaluate any feature of the data, and IRT offers a surfeit of features worthy of exploration, including model-, item-, and person-fit statistics, local dependence diagnostics, IRT-scaled score ranking, as well as classical test theory measures such as item-total correlations. As an illustration of the power of this approach, I used posterior PMC to investigate two features of the model: item fit and

Table 7.2. *Bayesian MCMC estimates of the item parameters and standard deviations of the unidimensional model of the PISA data.*

Item	Discrimination	Difficulty	Lower Asymptote
Apples Q1	1.99	.12	.13
	[.39]	[.30]	[.06]
Apples Q2	2.57	.65	.02
	[.36]	[.26]	[.01]
Growing Up Q1	1.29	-.52	.09
	[.17]	[.18]	[.07]
Growing Up Q3	1.13	-.04	.05
	[.15]	[.16]	[.05]
Racing Car Q1	.93	-.84	.11
	[.09]	[.20]	[.09]
Racing Car Q2	2.23	-1.42	.13
	[.35]	[.33]	[.10]
Racing Car Q3	1.88	-1.54	.14
	[.28]	[.29]	[.11]
Racing Car Q5	1.67	.99	.03
	[.26]	[.25]	[.02]
Triangles	1.06	-.40	.14
	[.16]	[.24]	[.09]
Carpenter	2.08	1.39	.07
	[.47]	[.57]	[.02]
Pipelines	1.54	.59	.29
	[.51]	[.62]	[.09]

Note. N = 1,000. Standard deviations of the Monte Carlo standard errors are bracketed.

local dependence diagnosis.

Figure 7.1 displays item fit plots for each of the 11 PISA mathematics items. The red line in each plot is the observed proportion of correct response at each possible total score. The dotted gray lines represent the 5th and 95th percentiles of the predicted proportions correct across all replications (R = 500), and the dashed line represents the 50th percentile. For most items, the observed item characteristic curve was within the 5th and 95th percentile boundaries. Although the observed proportions deviated from the predicted proportions in a few minor cases (i.e., cases wherein the red line extended beyond either of the gray dotted lines), the overall trend was that the observed proportions were representative of the predicted proportions. This result enhances our confidence in the item-fit generalizability of the unidimensional 3PL model; the result given by this model with regard to the observed PISA data is closely aligned with the results one would expect across 500 similar data sets.

Posterior PMC was also used to assess how well the unidimensional 3PL model addressed the dimensionality of the PISA data. Chen and Thissen (1997) examined (among other local dependence measures) the standardized log-odds ratio difference, wherein the log-odds ratio of the observed 2×2 contingency table is given by:

$$\tau_{\text{obs}} = \ln\left(\frac{O_{11} \times O_{22}}{O_{12} \times O_{21}}\right). \tag{7.3}$$

The log-odds ratio τ_{exp} of the expected contingency table is found by substituting E_{ij} for

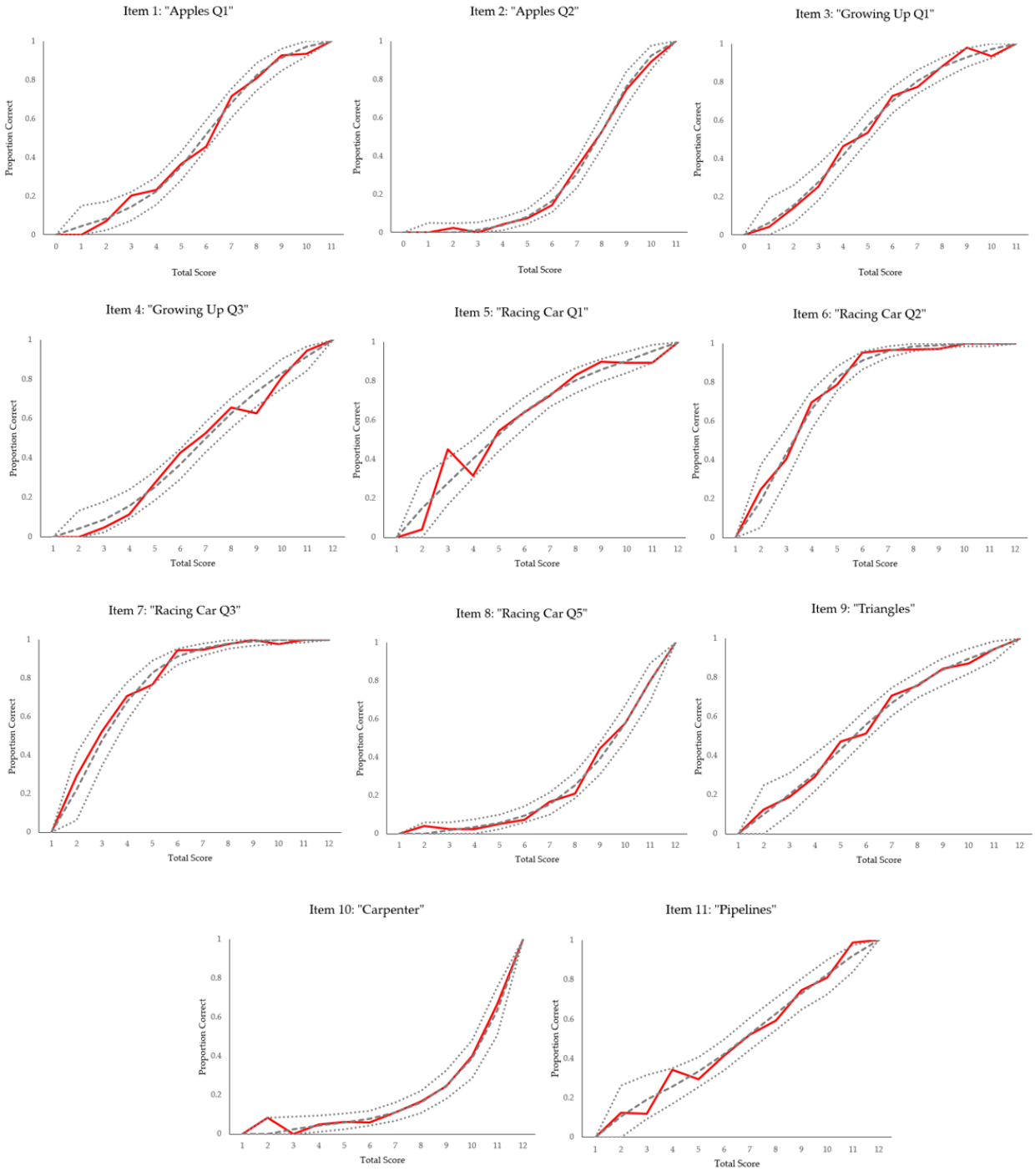


Figure 7.1. Item fit plots of each of the PISA mathematics items. The red lines represent the proportion of correct response at each of the possible total scores. The dotted lines represent the 5th and 95th percentiles of correct response proportions across 500 data sets replicated from the posterior predictive distribution.

the O_{ij} entries in Equation 7.3. These two ratios are then compared via:

$$\frac{\tau_{\text{obs}} - \tau_{\text{exp}}}{\sqrt{\sum_i \sum_j \frac{1}{O_{ij}}}}, \tag{7.4}$$

where the denominator denotes the standard deviation of the log-odds ratio statistic. If the assumption of local dependence has been violated, then the observed log-odds ratio τ_{obs} will be larger than the expected (unidimensional) log-odds ratio τ_{exp}. As Stone and Zhu (2015) note, "The [odds ratio] measure has been found to be effective for checking several aspects of model fit in the [posterior] PMC context" (p. 214).

To check how well the unidimensional 3PL model addressed the local dependence of the PISA data, I examined the posterior predictive *p*-values (*PPP* values) of the log-odds ratio difference in the observed and predicted data (R = 500 replications). Here, the *PPP* value for each of the $n(n\text{-}1)/2 = 55$ pairwise comparisons is simply the proportion of R data sets in which the predicted log-odds ratio is higher than (or equal to) the observed log-odds ratio. If the *PPP* value for a given item pair is less than or equal to .05, then the items are exhibiting significant positive local dependence; if the *PPP* values is greater than or equal to .95, then the items are showing significant negative local dependence. Thus, in truly unidimensional data, the *PPP* values for each item pair would fall between .05 and .95.

The *PPP* values of the PISA data ranged from 0 to .996. Nineteen item pairs had *PPP* values of exactly zero, meaning that across R = 500 replicated data sets, the

predicted (unidimensional) log-odds ratio was *never* higher than the observed log-odds ratio. In total, 29 of the 55 item pairs exhibited significant positive local dependence and 4 pairs demonstrated significant negative local dependence. These results are visualized in Figure 7.2, which shows a pie plot of the *PPP* value of each pairwise comparison among the 11 PISA mathematics items. The black portion of each circle indicates the magnitude of the *PPP* value for that particular item pair. This figure illuminates a few compelling patterns regarding the assumption of unidimensionality. For example, the log-odds ratio difference between Items 2 (Apples Q2) and 1 (Apples Q1) had a *PPP* value of 0, meaning that the assumption of unidimensionality between these items has been violated. Conversely, the log-odds ratio difference between Items 9 (Triangles) and 4 (Growing Up Q3) had a *PPP* value of .996; the assumption of local independence in this item pair was only violated in two (or .4%) of the replicated data sets. There are many other patterns to be dissected in Figure 7.2 (though there does not appear to be a systematic pattern that reflects residual dependence between the posited testlets). The general conclusion based on these results is that the pairwise local dependencies are not as ignorable as one would expect in truly unidimensional data.

As mentioned earlier, model checking enables the researcher to scrutinize any possible feature of the data. The examples above explored the item fit and local independence of the observed data, relative to 500 data sets replicated from the posterior predictive distribution. Posterior (or prior) PMC could be used in a similar

manner to critique the observed data with reference to numerous diagnostic measures, statistics, and indices, thereby providing a more nuanced model appraisal than that granted by frequentist methods.

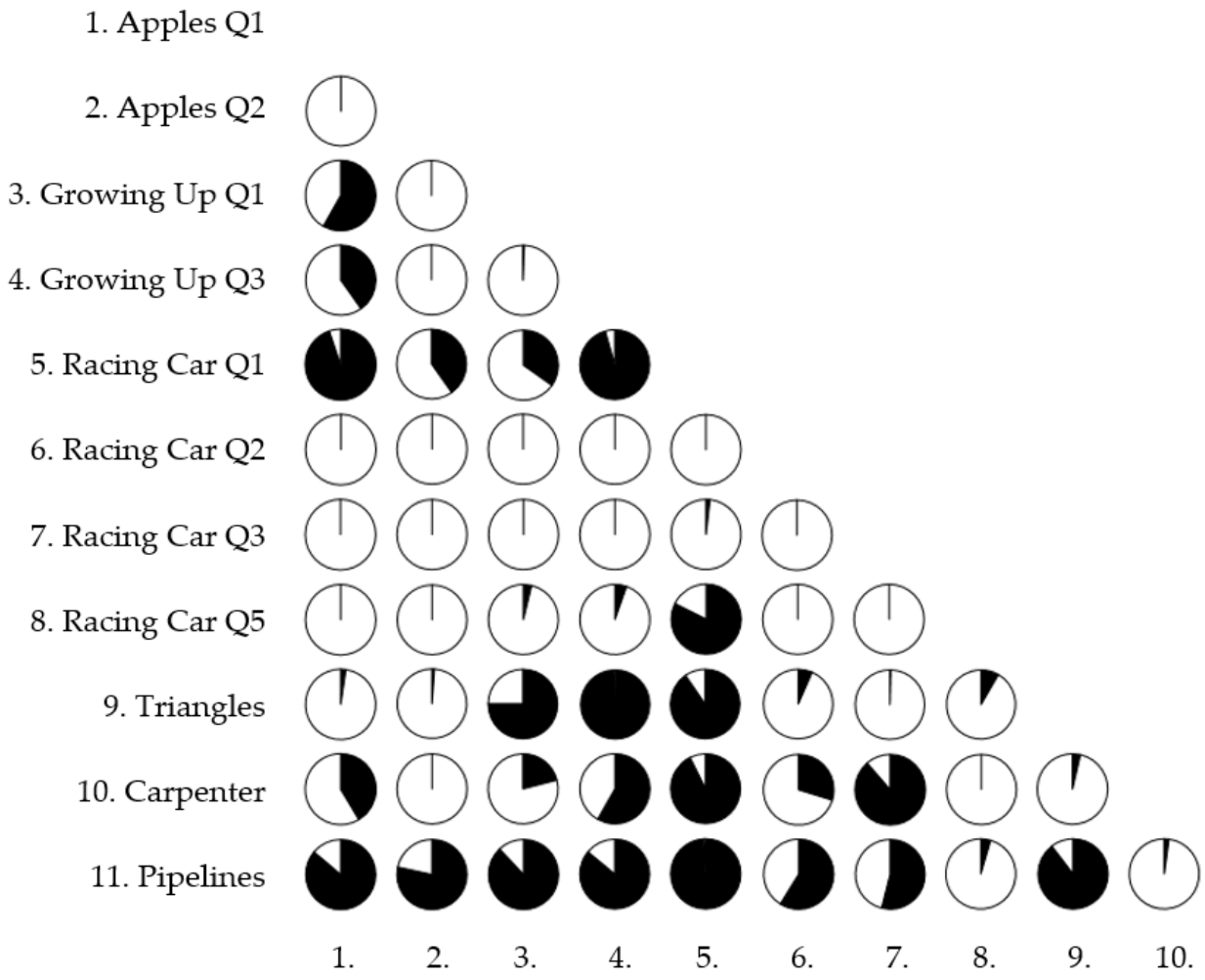


Figure 7.2. Pie plots of the pairwise log-odds ratio differences between the 11 PISA mathematics items.

7.1.4 Information-theoretic Evaluation

The information-theoretic evaluation of the unidimensional 3PL model was essentially carried out already in Chapters 4, 5, and 6. Of course, these chapters were not concerned with fitting empirical item response data. Indeed, it is difficult to recognize the role of real data in an MDL analysis; the principle of minimum description length provides insights about the intrinsic capability of a model to fit well, not to some particular observed data, but relative to all possible data sets. Thus, the primary goal in an empirical MDL analysis is to define the complete data space, relative to the sample data.

The PISA data under investigation comprised 11 dichotomous math items. Implementation of the MDL strategy used in earlier chapters therefore involved generating 1,000 binary data sets, each with 2^11 = 2,048 possible item response patterns. The frequencies of each response pattern were then randomly and uniformly sampled from a unit simplex, just as described in Section 4.3.1. This resulted in a large number of random data sets with the same number of items as the PISA data. These data sets served as a representation of the total data space; the unidimensional 3PL model was then fit to every data set.

In the simulated MDL analyses of earlier chapters, it was necessary to find a fit statistic (e.g., Y2/*N*) that would not be affected by the number of free parameters or other features of the different models under investigation. In the MDL analysis of a

single model, this constraint is not necessary. Instead, one can choose any feature of the model, as in PMC, and explore its behavior within the complete data space.

For the purposes of the PISA illustration, the M_2 global fit measure (Maydeu-Olivares & Joe, 2005) was selected as the feature of interest. Figure 7.3 displays the cumulative percentage distribution of the M_2 statistic as calculated in each of the 1,000 data sets. M_2 values ranged from 56.44 to 243.76, with an average of 140.00. The M_2 value in the observed PISA data was 61.11, as depicted by the green dot in the figure. This indicates that, relative to the realized data, model-data fit as indexed by M_2 was considerably worse in the vast majority of the data space. In fact, only three of the 1,000 random data sets provided better fit to the unidimensional 3PL model. In other words, this model does not possess an innate ability to produce low M_2 values. The 11-item unidimensional 3PL model is extremely unlikely to fit well by chance, so the fact that the observed data resulted in such a low M_2 gives credence to the hypothesis that a unidimensional 3PL model is an acceptable representation of the PISA data.

7.1.5 Conclusion

The integrative framework is a theoretical unification of the prevailing philosophies of statistical inference. It is also a research instrument, designed to be used in the service of comprehensive, well-informed model evaluation. This chapter presented an illustration of the framework's functionality by exploring the unidimensional 3PL IRT model in the

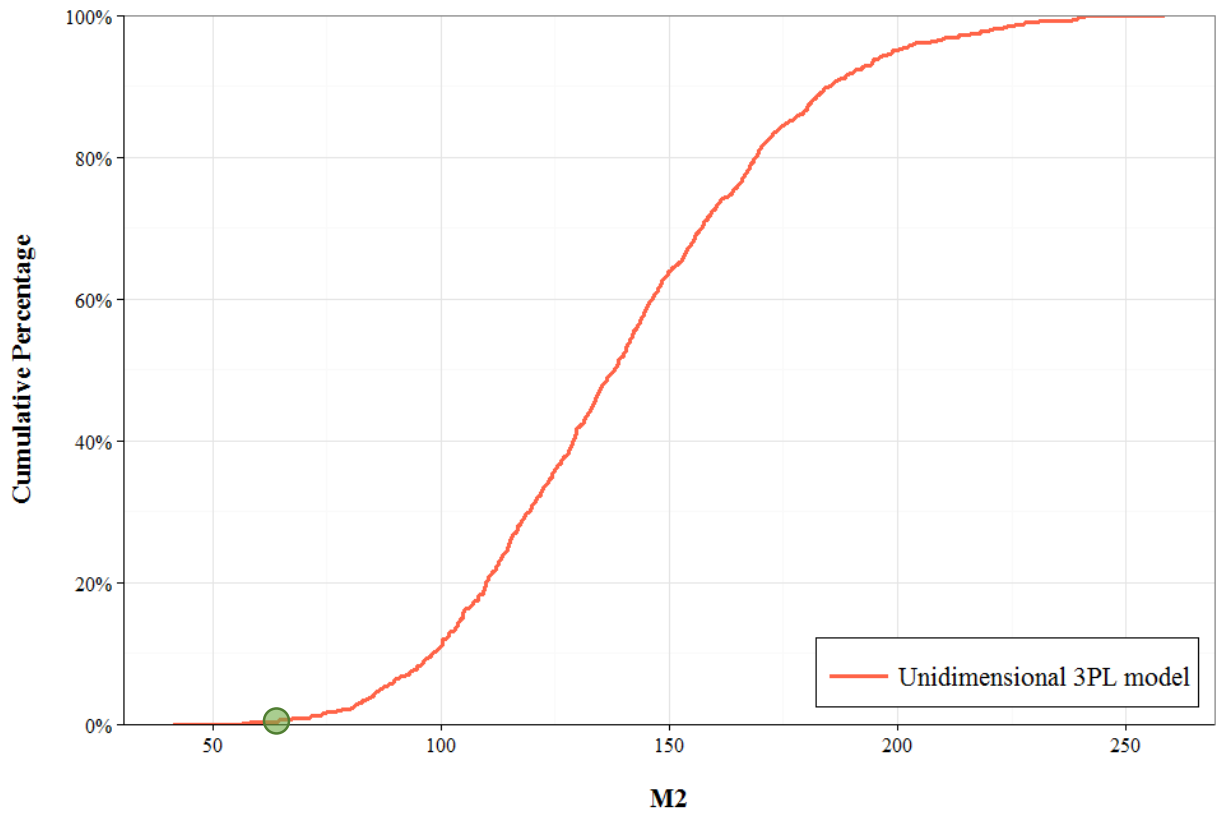


Figure 7.3. Cumulative percentage distribution of M_2 when fitting a unidimensional 3PL model to the complete data space.

context of real data. The findings from the frequentist perspective confirmed that the 11 PISA mathematics items under consideration encompassed a wide range of difficulty and successfully discriminated between respondents. Moreover, the M_2 statistic from the frequentist appraisal demonstrated that the unidimensional 3PL model fit the PISA data quite well, though not perfectly. The Bayesian evaluation delved into features other than global fit, including item-level fit and local dependence violations. Importantly, these metrics were assessed relative to 500 data sets replicated from the

posterior predictive distribution and not from the observed data alone. The Bayesian approach found that, relative to the predicted data, all 11 items in the observed data fit well, but there were numerous violations of the local independence assumption, though not in any recognizable pattern. Finally, the unidimensional 3PL model was viewed through the lens of the MDL principle. By fitting the same model to 1,000 random data sets, it was determined that the unidimensional 3PL model does not have an inherent tendency to fit well. Thus, the low M_2 value given in the frequentist evaluation can be trusted as a meaningful index of fit and not as the byproduct of overfitting. Overall, this multifaceted appraisal supports the hypothesis that the PISA data are sufficiently represented by a unidimensional 3PL model.

CHAPTER 8

Final Remarks

In psychological and educational measurement, models are often judged exclusively by their ability to describe the observed data, with little regard for the notions of generalizability or complexity. What is needed in psychometrics is a more comprehensive examination of a proposed model—its strengths, weaknesses, flaws, behavior, performance. Bayesian methods equip researchers with the power to explore the generalizability of specific features of the model, relative to replicated data. Information theory allows one to evaluate the inherent fitting tendency of a model, relative to any and all possible data. The framework presented in this dissertation united these three approaches to model evaluation, investigated in great detail the use of the information-theoretic approach in IRT, and demonstrated the usefulness of the framework in empirical data analysis.

8.1 Review of the Findings

The most intriguing findings were related to the simulation study, which presented, for the first time, a formal examination of the principle of MDL in the context of IRT modeling. Five common models were fit to 1,000 random data sets and various test-

level and item-level diagnostics were computed in each data set. Four of these models were multidimensional in structure, and they possessed the same number of freely estimated parameters. As indicated by the Y2/*N* index, the bifactor model tended to fit random data almost as efficiently as the exploratory factor analytic model—a model specifically designed to accommodate an extensive range of data patterns. The other two multidimensional models, the DINA and DINO diagnostic classification models, fit well to almost the exact same *number* of random data sets, yet they did not fit well to the *same* data sets. Perhaps the most enlightening finding was related to the fifth model under investigation: a unidimensional 3PL model that had a simpler functional form than any of the multidimensional models, but additional complexity in terms of an extra freely estimated parameter. Despite this increase in flexibility, the unidimensional 3PL model was shown to be far less inclined to fit any possible data.

In Chapter 7, the integrative framework was used as a tool for empirical data analysis. A unidimensional 3PL model was thoroughly evaluated using 11 mathematics items from the 2000 PISA. The frequentist perspective yielded information about overall fit to the observed data, the Bayesian approach explored item fit and local dependence relative to 500 replicated data sets, and the information-theoretic principle of MDL revealed that the unidimensional 3PL was highly unlikely to achieve good fit by chance.

8.2 Implications for Education Research

8.2.1. Implications of the Framework

The integrative framework is especially well-suited for educational research. Often, substantive education researchers are less interested in how closely a given model fit the particular observed data, and more interested in express features of the model. Consider, for example, the high-stakes issue of classifying English language learner (ELL) students. Abedi (2008) summarized the extant problems with ELL classification, which include differences in operational definitions, lack of standards, and various legislative complications. A global fit assessment of an English proficiency examination would not provide any indication of misclassification of ELL students. Instead of relying on this frequentist tactic, a researcher could consult an alternative column of the framework. Perhaps posterior predictive model checking could be used to discern whether the observed misclassification rate is representative of the expected misclassification rate.

Further, the English language proficiency exams that are typically used in ELL classification are often multidimensional, comprising subscales such as reading, writing, listening, and speaking. Abedi (2008) noted that ELL classification is often carried out using scores from each separate dimension, as well as the composite score. Posterior PMC could also be used to assess the role of dimensionality in ELL classification, perhaps by measuring the observed local dependence relative to the expected local dependence, as in the PISA data analysis of Chapter 7. The MDL

principle could also provide some insight regarding the dimensionality of an English language assessment. For example, the fitting propensity of a model that represents all items along one "English language" dimension (i.e., a model with a simple functional form) would likely be much lower that the fitting propensity of a model that comprises four separate but correlated dimensions (i.e., a model with a more complex functional form). With regard to an educational issue as complex as ELL classification, simple model-data fit metrics will not suffice; the integrative framework presents researchers with multiple appealing alternatives that are designed to answer more comprehensive questions about models.

8.2.2 Implications of the MDL Principle

Aside from its role in the framework, the principle of MDL has great potential in educational measurement research. The National Center for Education Statistics (NCES), for instance, appoints test developers to decide on appropriate models for use in various large-scale assessments, such as the National Assessment of Educational Progress (NAEP). As an illustration, suppose that NCES contractors are presented with an item that could be fit with a number of different IRT models, and the appropriate model must be selected *a priori* and in the absence of data. A reasoned approach, based on evidence, would be to gauge the fitting propensity of the candidate models using the logic of MDL. Then, without collecting or analyzing any data, the NAEP test developers

could select an appropriate, effective model that avoids overfitting and generalizes well. This is but one illustration of the myriad applications of MDL in educational measurement.

8.2 Impact

Overall, I believe that the line of research developed in this dissertation makes an impactful contribution in multiple areas of educational and psychological research. First, this work advances the topic of IRT by introducing the minimum description length principle as a viable tool of model evaluation. I believe that this information-theoretic approach has immense potential in psychological and educational measurement. When comparing models, for example, a psychometrician might ignore classical goodness-of-fit tests or replication methods, opting instead to view the competing models through the lens of MDL. This may lead to the discovery that, although Model B has better goodness-of-fit to the observed data, Model A is less likely to fit any possible data and is, in that sense, the better model. At the very least, application of the MDL principle will provide the field of psychometrics with a better understanding of the relationships between the most common IRT models.

Secondly, the empty cells in the framework illuminate previously unexamined methods of evaluating statistical models. It is certainly feasible that one might sample data from the prior or posterior and then re-fit the model at each iteration. Perhaps a

test quantity could be used to assess the discrepancies between the model and the random data. Many researchers are blind to methods that exist outside of their own area of expertise (sometimes willfully so); by zooming out and considering the other cells in this consolidated framework, one can explore different approaches to model appraisal, including multiple techniques that are entirely novel.

Ultimately, the integrative framework provides a better understanding of the relationships between the frequentist, Bayesian, and information-theoretic approaches to model evaluation. In psychological and educational research, a schism has long existed between the frequentist and Bayesian philosophies, and the information-theoretic approach is completely unfamiliar. Uniting these seemingly disparate perspectives in a single framework will lead to a better understanding of model evaluation, not only in quantitative psychology and educational research methods, but for science as a whole.

APPENDIX A

Data Generation

R code

```
# Number of items
nitems <- 7

# Generate 2^7 = 128 binary response patterns
x <- list(0:1)
pattern <- expand.grid(rep(x,nitems))

# A large number
M <- 10^6

# Sample 2^nitems-1 integers
p <- sample.int(M,size=(2^nitems-1))

# Sort in place
p <- sort(p)

# Append the ends
p <- c(0,p,M)

# Lag-1 difference
p <- diff(p)

# The desired simplex
p <- p/M

# The "sample" size
# N = 10,000 to give the response patterns realistic frequencies
p <- p*10000

# The weighted response pattern data
data <- cbind(pattern,p)
```

Example data (data set 1 of 1,000)

```
# The final column represents # of simulated respondents (out of
# N = 10,000) who provided that particular response pattern

0 0 0 0 0 0 0 0 33.94
1 0 0 0 0 0 0 0 94.61
0 1 0 0 0 0 0 0 32.1
1 1 0 0 0 0 0 0 103.6
0 0 1 0 0 0 0 0 9.59
1 0 1 0 0 0 0 0 83.81
0 1 1 0 0 0 0 0 104.89
1 1 1 0 0 0 0 0 61.2
0 0 0 1 0 0 0 0 36.99
1 0 0 1 0 0 0 0 95.15
0 1 0 1 0 0 0 0 3.14
1 1 0 1 0 0 0 0 73.35
0 0 1 1 0 0 0 0 79.85
1 0 1 1 0 0 0 0 29.34
0 1 1 1 0 0 0 0 108.54
1 1 1 1 0 0 0 0 9.89
0 0 0 0 1 0 0 0 124.01
1 0 0 0 1 0 0 0 69.36
0 1 0 0 1 0 0 0 270.7
1 1 0 0 1 0 0 0 30.26
0 0 1 0 1 0 0 0 16.73
1 0 1 0 1 0 0 0 140.49
0 1 1 0 1 0 0 0 1.33
1 1 1 0 1 0 0 0 193.21
0 0 0 1 1 0 0 0 192.64
1 0 0 1 1 0 0 0 30.02
0 1 0 1 1 0 0 0 75.02
1 1 0 1 1 0 0 0 26.01
0 0 1 1 1 0 0 0 44.72
1 0 1 1 1 0 0 0 47.83
0 1 1 1 1 0 0 0 90.61
1 1 1 1 1 0 0 0 6.28
0 0 0 0 0 1 0 0 28.42
1 0 0 0 0 1 0 0 191.73
0 1 0 0 0 1 0 0 162.98
1 1 0 0 0 1 0 0 154.57
0 0 1 0 0 1 0 0 109.43
1 0 1 0 0 1 0 0 9.67
0 1 1 0 0 1 0 0 89.81
1 1 1 0 0 1 0 0 125.62
0 0 0 1 0 1 0 0 65.83
1 0 0 1 0 1 0 0 52.49
0 1 0 1 0 1 0 0 37.16
1 1 0 1 0 1 0 0 60.23
0 0 1 1 0 1 0 0 8
1 0 1 1 0 1 0 0 5.95
0 1 1 1 0 1 0 0 52.93
1 1 1 1 0 1 0 0 2.05
0 0 0 0 1 1 0 0 412.2
1 0 0 0 1 1 0 0 16.69
0 1 0 0 1 1 0 0 146.15
1 1 0 0 1 1 0 0 321.69
0 0 1 0 1 1 0 0 151.21
1 0 1 0 1 1 0 0 102.96
0 1 1 0 1 1 0 0 21.66
1 1 1 0 1 1 0 0 443.16
0 0 0 1 1 1 0 0 67.05
1 0 0 1 1 1 0 0 7.64
0 1 0 1 1 1 0 0 28.1
1 1 0 1 1 1 0 0 57.97
0 0 1 1 1 1 0 0 84.52
1 0 1 1 1 1 0 0 154.98
0 1 1 1 1 1 0 0 17.63
1 1 1 1 1 1 0 0 20.93
0 0 0 0 0 0 1 0 52.23
1 0 0 0 0 0 1 0 17.08
0 1 0 0 0 0 1 0 208.91
1 1 0 0 0 0 1 0 10.74
0 0 1 0 0 0 1 0 22.88
1 0 1 0 0 0 1 0 152.65
0 1 1 0 0 0 1 0 44.01
1 1 1 0 0 0 1 0 38.31
0 0 0 1 0 0 1 0 54.83
1 0 0 1 0 0 1 0 233.96
0 1 0 1 0 0 1 0 51.5
1 1 0 1 0 0 1 0 312.53
0 0 1 1 0 0 1 0 36.25
1 0 1 1 0 0 1 0 5.24
0 1 1 1 0 0 1 0 105.12
1 1 1 1 0 0 1 0 56.87
0 0 0 0 1 0 1 0 17.66
1 0 0 0 1 0 1 0 3.44
0 1 0 0 1 0 1 0 19.95
1 1 0 0 1 0 1 0 14.22
0 0 1 0 1 0 1 0 75.62
1 0 1 0 1 0 1 0 32.88
0 1 1 0 1 0 1 0 226.89
1 1 1 0 1 0 1 0 36.29
0 0 0 1 1 0 1 0 104.22
1 0 0 1 1 0 1 0 37.89
0 1 0 1 1 0 1 0 37.05
1 1 0 1 1 0 1 0 29.2
0 0 1 1 1 0 1 0 87.21
1 0 1 1 1 0 1 0 159.23
0 1 1 1 1 0 1 0 42.97
1 1 1 1 1 0 1 0 93.94
0 0 0 0 0 1 1 0 95.14
1 0 0 0 0 1 1 0 203.35
0 1 0 0 0 1 1 0 30.22
1 1 0 0 0 1 1 0 6.11
0 0 1 0 0 1 1 0 16.23
1 0 1 0 0 1 1 0 131.6
0 1 1 0 0 1 1 0 124.08
1 1 1 0 0 1 1 0 12.83
0 0 0 1 0 1 1 0 17.01
1 0 0 1 0 1 1 0 0.05
0 1 0 1 0 1 1 0 128.65
1 1 0 1 0 1 1 0 42.71
0 0 1 1 0 1 1 0 129.86
1 0 1 1 0 1 1 0 8.01
0 1 1 1 0 1 1 0 77.01
1 1 1 1 0 1 1 0 14.29
0 0 0 0 1 1 1 0 41.65
1 0 0 0 1 1 1 0 23.26
0 1 0 0 1 1 1 0 74.15
1 1 0 0 1 1 1 0 7.24
0 0 1 0 1 1 1 0 52.88
1 0 1 0 1 1 1 0 60.54
0 1 1 0 1 1 1 0 179.15
1 1 1 0 1 1 1 0 41.85
0 0 0 1 1 1 1 0 18.86
1 0 0 1 1 1 1 0 70.72
0 1 0 1 1 1 1 0 27.38
1 1 0 1 1 1 1 0 278.55
0 0 1 1 1 1 1 0 6.27
1 0 1 1 1 1 1 0 95.64
0 1 1 1 1 1 1 0 8.59
1 1 1 1 1 1 1 0 23.73
```

APPENDIX B

Additional Simulation Results

Table B.1. *Means and standard deviations of the* S-X^2 *item fit statistics in the exploratory factor analytic model.*

	S-X²							
	Item 1		Item 2		Item 3		Item 4	
EFA Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.82	2.08	1.84	1.83	1.85	2.00	1.82	1.89
Converged	1.86	2.11	1.84	1.83	1.90	2.03	1.85	1.94
Non-converged	1.66	1.93	1.82	1.83	1.66	1.85	1.71	1.72
Difference	.20	.18	.02	.00	.24	.18	.14	.12
	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.86	1.93	1.92	1.99	1.80	1.84	1.84	1.94
Converged	1.90	1.94	1.97	2.01	1.85	1.86	1.88	1.96
Non-converged	1.71	1.87	1.74	1.90	1.61	1.78	1.70	1.84
Difference	.19	.07	.23	.11	.24	.08	.18	.12

Note. N_{Total} = 1000; $N_{Converged}$ = 790; $N_{Non\text{-}converged}$ = 210; EFA = exploratory factor analytic model.

Table B.2. *Means and standard deviations of the* S-X^2 *item fit statistics in the bifactor model.*

	$S\text{-}X^2$							
	Item 1		Item 2		Item 3		Item 4	
Bifactor Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.83	1.96	1.76	1.79	1.82	1.96	1.75	1.82
Converged	1.86	1.98	1.76	1.84	1.85	1.98	1.77	1.80
Non-converged	1.71	1.89	1.74	1.60	1.68	1.90	1.68	1.88
Difference	.16	.08	.02	.24	.16	.08	.10	-.08

	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.81	1.93	1.78	1.89	1.68	1.82	1.78	1.88
Converged	1.80	1.81	1.77	1.87	1.69	1.86	1.79	1.88
Non-converged	1.84	2.40	1.83	1.98	1.63	1.67	1.73	1.90
Difference	-.04	-.58	-.06	-.11	.06	.18	.06	-.03

Note. N_{Total} = 1000; $N_{Converged}$ = 816; $N_{Non\text{-}converged}$ = 184.

Table B.3. *Means and standard deviations of the* S-X^2 *item fit statistics in the deterministic input noisy and-gate model.*

	S-X^2							
	Item 1		Item 2		Item 3		Item 4	
DINA Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.78	1.89	1.76	1.78	1.84	1.93	1.79	1.88
Converged	1.79	1.86	1.73	1.76	1.85	1.95	1.79	1.88
Non-converged	1.70	2.49	2.26	2.13	1.66	1.59	1.91	1.97
Difference	.09	-.63	-.53	-.37	.19	.36	-.12	-.09
	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.91	1.99	1.78	1.82	1.70	1.72	1.80	1.86
Converged	1.91	1.99	1.78	1.83	1.69	1.72	1.79	1.85
Non-converged	1.89	1.95	1.84	1.64	1.89	1.78	1.88	1.94
Difference	.02	.04	-.06	.19	-.20	-.06	-.09	-.08

Note. N_{Total} = 1000; $N_{Converged}$ = 953; $N_{Non\text{-}converged}$ = 47; DINA = deterministic input noisy and-gate model.

Table B.4. *Means and standard deviations of the* S-X^2 *item fit statistics in the deterministic input noisy or-gate model.*

	S-X^2							
	Item 1		Item 2		Item 3		Item 4	
DINO Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.75	1.86	1.79	1.86	1.87	1.97	1.76	1.84
Converged	1.76	1.87	1.80	1.88	1.87	1.95	1.78	1.86
Non-converged	1.52	1.76	1.74	1.45	1.95	2.54	1.27	1.09
Difference	.24	.11	.06	.43	-.09	-.60	.51	.77
	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.88	1.90	1.76	1.78	1.70	1.75	1.79	1.85
Converged	1.85	1.86	1.75	1.79	1.71	1.74	1.79	1.85
Non-converged	2.72	2.64	1.79	1.72	1.59	2.00	1.80	1.88
Difference	-.87	-.79	-.04	.07	.11	-.26	-.01	-.04

Note. N_{Total} = 1000; $N_{Converged}$ = 961; $N_{Non\text{-}converged}$ = 39; DINO = deterministic input noisy or-gate model.

Table B.5. *Means and standard deviations of the* S-X^2 *item fit statistics in the unidimensional 3PL model.*

	S-X²							
	Item 1		Item 2		Item 3		Item 4	
Uni Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.87	2.00	1.83	1.89	1.84	1.92	1.78	1.82
Converged	1.90	2.07	1.82	1.81	1.81	1.84	1.78	1.80
Non-converged	1.79	1.78	1.87	2.14	1.93	2.17	1.78	1.88
Difference	.11	.29	-.05	-.33	-.11	-.34	-.01	-.08
	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	1.84	1.91	1.84	1.81	1.78	1.86	1.83	1.89
Converged	1.84	1.89	1.84	1.77	1.90	1.94	1.84	1.87
Non-converged	1.86	1.98	1.85	1.93	1.42	1.56	1.78	1.92
Difference	-.02	-.09	-.01	-.16	.48	.38	.06	-.05

Note. N_{Total} = 1000; $N_{Converged}$ = 757; $N_{Non\text{-}converged}$ = 243. Uni = unidimensional 3PL model.

Table B.6. *Means and standard deviations of the marginal χ^2 values in the exploratory factor analytic model.*

	Marginal χ^2							
	Item 1		Item 2		Item 3		Item 4	
EFA Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.00	.03	.04	.32	.07	.50	.06	.44
Converged	.00	.03	.02	.22	.05	.43	.06	.46
Non-converged	.01	.03	.09	.56	.14	.70	.06	.36
Difference	.00	-.01	-.06	-.35	-.09	-.26	-.01	.10

	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.04	.24	.04	.25	.04	.33	.04	.30
Converged	.04	.24	.03	.19	.04	.34	.03	.27
Non-converged	.06	.27	.09	.40	.05	.29	.07	.37
Difference	-.02	-.03	-.06	-.21	-.01	.05	-.04	-.10

Note. N_{Total} = 1000; $N_{Converged}$ = 790; $N_{Non\text{-}converged}$ = 210; EFA = exploratory factor analytic model.

Table B.7. *Means and standard deviations of the marginal χ^2 values in the bifactor model.*

	Marginal χ^2							
	Item 1		Item 2		Item 3		Item 4	
Bifactor Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.06	.48	.04	.36	.05	.35	.06	.49
Converged	.06	.46	.03	.29	.03	.24	.04	.39
Non-converged	.09	.55	.09	.59	.14	.63	.16	.80
Difference	-.04	-.09	-.06	-.29	-.10	-.39	-.12	-.41

	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.04	.38	.00	.03	.00	.02	.04	.03
Converged	.03	.37	.00	.03	.00	.01	.02	.26
Non-converged	.09	.41	.01	.04	.00	.02	.08	.43
Difference	-.06	-.04	.00	.00	.00	-.01	-.06	-.18

Note. N_{Total} = 1000; $N_{Converged}$ = 816; $N_{Non\text{-}converged}$ = 184.

Table B.8. *Means and standard deviations of the marginal χ^2 values in the deterministic input noisy and-gate model.*

	Marginal χ^2							
	Item 1		Item 2		Item 3		Item 4	
DINA Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.05	.28	.05	.33	.06	.28	.06	.36
Converged	.06	.29	.06	.34	.06	.28	.06	.37
Non-converged	.00	.03	.00	.01	.01	.02	.00	.03
Difference	.05	.26	.06	.33	.05	.26	.06	.34

	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.06	.27	.05	.30	.05	.27	.05	.30
Converged	.06	.27	.05	.31	.05	.28	.06	.31
Non-converged	.01	.02	.00	.00	.00	.02	.00	.02
Difference	.05	.25	.05	.31	.05	.26	.05	.28

Note. N_{Total} = 1000; $N_{Converged}$ = 953; $N_{Non\text{-}converged}$ = 47; DINA = deterministic input noisy and-gate model.

Table B.9. *Means and standard deviations of the marginal χ^2 values in the deterministic input noisy or-gate model.*

	Marginal χ^2							
	Item 1		Item 2		Item 3		Item 4	
DINO Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.04	.23	.04	.26	.05	.22	.04	.23
Converged	.04	.23	.04	.27	.05	.23	.04	.23
Non-converged	.00	.00	.00	.00	.01	.03	.00	.00
Difference	.04	.23	.04	.27	.04	.20	.04	.23

	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.04	.18	.04	.24	.04	.23	.04	.23
Converged	.04	.19	.04	.25	.04	.24	.04	.23
Non-converged	.01	.02	.01	.02	.02	.10	.01	.03
Difference	.03	.16	.03	.15	.03	.19	.04	.20

Note. N_{Total} = 1000; $N_{Converged}$ = 961; $N_{Non\text{-}converged}$ = 39; DINO = deterministic input noisy or-gate model.

Table B.10. *Means and standard deviations of the marginal χ^2 values in the unidimensional 3PL model.*

	Marginal χ^2							
	Item 1		Item 2		Item 3		Item 4	
Uni Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.06	.29	.06	.39	.05	.27	.06	.34
Converged	.07	.32	.07	.41	.05	.25	.06	.32
Non-converged	.03	.14	.04	.30	.04	.32	.07	.38
Difference	.03	.18	.03	.11	.01	-.07	.00	-.07
	Item 5		Item 6		Item 7		All items	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	.07	.51	.07	.47	.08	.43	.07	.38
Converged	.09	.58	.09	.53	.08	.44	.07	.41
Non-converged	.02	.09	.04	.16	.05	.40	.04	.26
Difference	.07	.49	.05	.37	.03	.04	.03	.15

Note. N_{Total} = 1000; $N_{Converged}$ = 757; $N_{Non\text{-}converged}$ = 243. Uni = unidimensional 3PL model.

Table B.11. *Means and standard deviations of the LD X^2 statistics for each item pair in the exploratory factor analytic model.*

	LD X^2									
	Items 2 & 1		Items 3 & 1		Items 3 & 2		Items 4 & 1		Items 4 & 2	
EFA Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	8.70	21.12	7.66	16.80	8.69	20.68	8.06	17.96	9.97	22.13
Converged	9.17	21.84	8.03	17.52	8.84	19.77	7.99	17.56	8.80	20.60
Non-conv.	6.94	18.11	6.27	13.72	8.13	23.82	8.35	19.44	14.35	26.70
Difference	2.23	3.72	1.76	3.79	.72	-4.05	-.26	-1.88	-5.55	-6.10

	Items 4 & 3		Items 5 & 1		Items 5 & 2		Items 5 & 3		Items 5 & 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	9.08	23.80	9.28	21.19	9.49	21.07	8.61	20.99	10.81	24.38
Converged	8.28	23.06	9.90	21.72	8.74	19.77	7.80	19.51	9.61	21.61
Non-conv.	12.11	26.21	6.93	18.95	12.30	25.21	11.61	25.61	15.31	32.45
Difference	-3.83	-3.14	2.97	2.77	-3.56	-5.44	-3.81	-6.10	-5.70	-10.84

	Items 6 & 1		Items 6 & 2		Items 6 & 3		Items 6 & 4		Items 6 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	7.55	18.55	8.12	18.59	8.02	19.43	9.51	24.91	9.90	22.68
Converged	7.83	19.07	7.31	17.21	7.74	19.71	7.85	17.63	9.74	22.03
Non-conv.	6.50	16.45	11.18	22.83	9.06	18.33	15.75	41.72	10.49	25.01
Difference	1.33	2.63	-3.87	-5.62	-1.32	1.38	-7.90	-24.09	-.75	-2.98

	Items 7 & 1		Items 7 & 2		Items 7 & 3		Items 7 & 4		Items 7 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	8.53	21.72	8.87	20.13	8.95	21.97	8.98	20.41	9.17	21.77
Converged	8.22	20.06	8.70	18.89	8.34	20.64	8.46	19.26	8.78	20.38
Non-conv.	9.71	27.09	9.50	24.26	11.26	26.29	10.94	24.19	10.64	26.36
Difference	-1.49	-7.04	-.80	-5.37	-2.92	-5.65	-2.48	-4.93	-1.87	-5.98

	Items 7 & 6		All item pairs	
	M	*SD*	*M*	*SD*
All data sets	9.34	21.31	8.92	21.03
Converged	9.06	21.46	8.53	19.97
Non-conv.	10.41	20.75	10.37	23.98
Difference	-1.35	.71	-1.84	-4.01

Note. N_{Total} = 1000; $N_{Converged}$ = 790; $N_{Non\text{-}converged}$ = 210; EFA = exploratory factor analytic model; Non-conv = non-converged analyses.

Table B.12. *Means and standard deviations of the LD X^2 statistics for each item pair in the bifactor model.*

Bifactor Model	LD X^2									
	Items 2 & 1		Items 3 & 1		Items 3 & 2		Items 4 & 1		Items 4 & 2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	6.25	14.40	6.17	15.61	6.64	18.27	6.61	16.43	7.28	19.78
Converged	6.49	14.90	6.15	15.97	6.71	17.66	5.74	14.19	7.45	20.72
Non-conv.	5.22	11.93	6.25	13.95	6.33	20.78	10.44	23.63	6.51	14.91
Difference	1.27	2.97	-.11	2.02	.37	-3.12	-4.70	-9.44	.94	5.82
	Items 4 & 3		Items 5 & 1		Items 5 & 2		Items 5 & 3		Items 5 & 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	5.54	14.04	7.29	18.91	6.67	16.31	6.67	19.12	6.90	17.53
Converged	5.61	14.07	6.86	18.39	6.44	15.62	6.42	18.37	6.59	16.46
Non-conv.	5.26	13.97	9.18	21.03	7.72	19.08	7.78	22.17	8.27	21.65
Difference	.35	.10	-2.32	-2.64	-1.28	-3.45	-1.36	-3.81	-1.68	-5.19
	Items 6 & 1		Items 6 & 2		Items 6 & 3		Items 6 & 4		Items 6 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	14.22	32.08	14.21	27.39	14.64	28.92	15.41	31.44	14.88	30.42
Converged	14.72	34.26	14.27	27.58	15.17	29.40	15.15	31.40	14.52	30.04
Non-conv.	12.00	19.54	13.94	26.60	12.27	26.64	16.58	31.67	16.46	32.08
Difference	2.73	14.72	.33	.98	2.90	2.76	-1.43	-.27	-1.93	-2.04
	Items 7 & 1		Items 7 & 2		Items 7 & 3		Items 7 & 4		Items 7 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	17.00	34.29	15.59	31.23	16.58	31.42	15.15	27.57	15.80	30.48
Converged	16.82	34.72	15.65	31.90	17.68	33.30	15.00	27.16	15.65	30.23
Non-conv.	17.82	32.41	15.32	28.14	11.71	20.51	15.83	29.39	16.43	31.62
Difference	-1.00	2.31	.32	3.75	5.96	12.79	-.83	-2.23	-.77	-1.39
	Items 7 & 6		All item pairs							
	M	*SD*	*M*	*SD*						
All data sets	6.45	22.49	10.76	23.72						
Converged	6.49	23.11	10.74	23.78						
Non-conv.	6.29	19.57	10.84	22.92						
Difference	.20	3.54	-.10	.87						

Note. N_{Total} = 1000; $N_{Converged}$ = 816; $N_{Non\text{-}converged}$ = 184. Non-conv = non-converged analyses.

Table B.13. *Means and standard deviations of the LD X^2 statistics for each item pair in the deterministic input noisy and-gate model.*

	LD X^2									
	Items 2 & 1		Items 3 & 1		Items 3 & 2		Items 4 & 1		Items 4 & 2	
DINA Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	10.50	20.95	19.27	31.38	20.85	34.66	27.83	46.94	28.54	46.77
Converged	10.73	21.21	19.15	30.95	20.82	34.65	28.02	46.98	28.30	46.13
Non-conv.	5.84	14.00	21.58	39.30	21.50	35.21	23.94	46.42	33.34	58.61
Difference	4.89	7.21	-2.43	-8.34	-.68	-.56	4.08	.55	-5.04	-12.48

	Items 4 & 3		Items 5 & 1		Items 5 & 2		Items 5 & 3		Items 5 & 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	17.52	28.89	35.17	54.12	33.90	51.12	26.93	42.47	20.69	35.97
Converged	17.47	28.80	35.39	54.96	34.01	50.85	27.16	42.77	20.87	35.92
Non-conv.	18.37	31.07	30.74	32.77	31.71	56.95	22.28	35.87	16.97	37.27
Difference	-.90	-2.27	4.65	22.19	2.29	-6.10	4.89	6.90	3.91	-1.35

	Items 6 & 1		Items 6 & 2		Items 6 & 3		Items 6 & 4		Items 6 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	31.15	50.27	31.95	53.86	32.45	53.17	28.50	50.07	22.10	34.56
Converged	31.60	50.75	32.13	54.34	32.38	53.02	28.95	50.82	22.03	34.10
Non-conv.	21.90	38.73	28.38	43.50	33.83	56.59	19.46	30.20	23.41	43.18
Difference	9.71	12.02	3.74	10.83	-1.45	-3.57	9.49	20.62	-1.38	-9.08

	Items 7 & 1		Items 7 & 2		Items 7 & 3		Items 7 & 4		Items 7 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	31.48	50.25	31.06	55.95	32.71	51.40	25.96	42.54	20.41	38.10
Converged	31.16	50.31	31.23	56.59	32.68	51.38	26.09	42.35	20.47	38.34
Non-conv.	38.00	49.11	27.45	41.13	33.44	52.35	23.42	46.59	19.13	33.19
Difference	-6.84	1.20	3.78	15.46	-.76	-.96	2.67	-4.23	1.34	5.14

	Items 7 & 6		All item pairs	
	M	*SD*	*M*	*SD*
All data sets	10.38	22.63	25.68	42.67
Converged	10.58	22.84	25.77	42.77
Non-conv.	6.31	17.44	23.86	39.98
Difference	4.27	5.40	1.92	2.79

Note. N_{Total} = 1000; $N_{Converged}$ = 953; $N_{Non\text{-}converged}$ = 47; DINA = deterministic input noisy and-gate model; Non-conv = non-converged analyses.

Table B.14. *Means and standard deviations of the LD X^2 statistics for each item pair in the deterministic input noisy or-gate model.*

	LD X^2									
	Items 2 & 1		Items 3 & 1		Items 3 & 2		Items 4 & 1		Items 4 & 2	
DINO Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	10.70	22.32	19.23	32.86	20.05	33.01	28.19	45.33	27.54	44.03
Converged	10.91	22.62	18.85	32.16	20.02	33.11	28.06	45.61	27.38	43.66
Non-conv.	5.64	12.20	28.70	46.55	20.71	30.81	31.28	37.95	31.47	52.87
Difference	5.27	10.42	-9.86	-14.39	-.69	2.30	-3.22	7.67	-4.09	-9.21

	Items 4 & 3		Items 5 & 1		Items 5 & 2		Items 5 & 3		Items 5 & 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	18.52	32.13	33.93	54.60	31.33	47.84	27.45	42.13	20.41	36.41
Converged	18.42	31.96	34.04	53.65	31.72	48.19	27.66	42.14	20.57	36.70
Non-conv.	21.16	36.43	31.13	75.05	21.83	37.59	22.38	42.01	16.41	28.66
Difference	-2.74	-4.47	2.92	-21.39	9.89	10.60	5.27	.13	4.16	8.04

	Items 6 & 1		Items 6 & 2		Items 6 & 3		Items 6 & 4		Items 6 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	31.99	53.77	32.73	53.47	30.16	47.47	26.17	46.33	20.24	35.52
Converged	31.94	54.00	32.47	53.48	30.54	47.65	25.53	44.89	20.39	35.68
Non-conv.	33.18	48.22	39.03	53.72	20.92	42.03	41.98	72.37	16.63	31.43
Difference	-1.24	5.79	-6.55	-.24	9.62	5.63	-16.45	-27.48	3.76	4.25

	Items 7 & 1		Items 7 & 2		Items 7 & 3		Items 7 & 4		Items 7 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	30.95	48.40	31.29	54.34	33.12	50.14	28.59	51.13	19.91	31.40
Converged	30.54	48.52	31.17	52.84	33.10	49.83	28.75	51.64	19.95	31.45
Non-conv.	40.99	44.78	34.15	84.11	33.73	57.83	24.58	36.37	19.07	30.51
Difference	-10.45	3.74	-2.98	-31.27	-.64	-8.00	4.17	15.27	.88	.94

	Items 7 & 6		All item pairs	
	M	*SD*	*M*	*SD*
All data sets	10.88	23.09	25.40	42.18
Converged	10.92	23.02	25.38	42.04
Non-conv.	9.97	24.91	25.95	44.11
Difference	.95	-1.89	-.57	-2.07

Note. N_{Total} = 1000; $N_{Converged}$ = 961; $N_{Non\text{-}converged}$ = 39; DINO = deterministic input noisy or-gate model; Non-conv = non-converged analyses.

Table B.15. *Means and standard deviations of the LD X^2 statistics for each item pair in the unidimensional 3PL model.*

	LD X^2									
	Items 2 & 1		Items 3 & 1		Items 3 & 2		Items 4 & 1		Items 4 & 2	
Uni Model	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	30.87	55.52	30.83	51.64	31.60	51.46	32.32	53.12	32.95	56.82
Converged	31.63	57.97	30.61	51.94	29.57	48.45	32.01	53.39	34.75	59.89
Non-conv.	28.50	47.13	31.52	50.78	37.90	59.52	33.27	52.36	27.35	45.63
Difference	3.13	10.84	-.91	1.16	-8.32	-11.07	-1.25	1.04	7.40	14.26
	Items 4 & 3		Items 5 & 1		Items 5 & 2		Items 5 & 3		Items 5 & 4	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	30.67	49.28	30.65	50.45	30.89	51.92	37.01	61.57	29.88	49.15
Converged	29.59	45.30	29.69	46.12	30.18	49.37	36.39	60.89	29.28	49.82
Non-conv.	34.03	59.99	33.62	62.05	33.10	59.24	38.97	63.75	31.77	47.03
Difference	-4.44	-14.69	-3.93	-15.94	-2.93	-9.87	-2.58	-2.87	-2.49	2.79
	Items 6 & 1		Items 6 & 2		Items 6 & 3		Items 6 & 4		Items 6 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	31.49	57.14	31.51	54.17	29.09	50.22	30.01	51.78	31.75	51.63
Converged	30.92	52.01	28.61	47.19	27.79	48.62	30.53	52.39	33.07	54.77
Non-conv.	33.26	70.87	40.54	71.05	33.12	54.80	28.39	49.91	27.62	40.13
Difference	-2.34	-18.85	-11.93	-23.86	-5.33	-6.18	2.14	2.48	5.45	14.64
	Items 7 & 1		Items 7 & 2		Items 7 & 3		Items 7 & 4		Items 7 & 5	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All data sets	29.58	47.06	32.07	56.86	31.46	52.65	33.21	54.70	32.57	57.78
Converged	29.89	48.39	30.98	52.31	30.77	52.93	31.94	54.44	32.02	56.17
Non-conv.	28.63	42.76	35.43	69.18	33.60	51.80	37.17	55.44	34.27	62.60
Difference	1.26	5.62	-4.45	-16.88	-2.83	1.13	-5.23	-1.00	-2.25	-6.42
	Items 7 & 6		All item pairs							
	M	*SD*	*M*	*SD*						
All data sets	33.02	55.81	31.59	53.37						
Converged	30.70	49.78	31.00	52.01						
Non-conv.	40.26	71.05	33.44	56.53						
Difference	-9.57	-21.28	-2.45	-4.52						

Note. N_{Total} = 1000; $N_{Converged}$ = 757; $N_{Non\text{-}converged}$ = 243. Uni = unidimensional 3PL model; Non-conv = non-converged analyses.

BIBLIOGRAPHY

Abedi, J. (2008). Classification system for English language learners: Issues and recommendations. *Educational Measurement: Issues and Practice, 27*, 17-31.

Achenbach, T. M. (1966). The classification of children's psychiatric symptoms: A factor-analytic study. *Psychological Monographs, 80,* 1-37.

Ackerman, R. A., Donnellan, M. B., & Robbins, R. W. (2012). An item response theory analysis of the narcissistic personality inventory. *Journal of Personality Assessment*, *94,* 141-155.

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris: Organization for Economic Cooperation and Development.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, 419-461.

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

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Bamber, D., & van Santen, J. P. (1985). How many parameters can a model have and still be testable? *Journal of Mathematical Psychology, 29*(4), 443-473.

Bandler, R., & Grinder, J. (1979). *Frogs Into Princes* (Vol. 15). Moab, UT: Real People Press.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual differences, 42*(5), 815-824.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology, 55*:1-15.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research, 27*(4), 525-546.

Bayes, T. (1764). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370-418.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459.

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, 383-430.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*(4), 581-612.

Cai, L. (2013). flexMIRT® version 2.00: A numerical engine for flexible multilevel multidimensional item analysis and test scoring [computer software]. Raleigh-Durham, NC: Vector Psychometric Group.

Cai, L. (2014). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, 1-25. DOI: 10.1007/s11336-014-9411-3

Cai, L., Maydeu-Olivares, A., Coffman, D., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3):265-289.

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

Collyer, C. E. (1985). Comparing strong and weak models by fitting them to computer-generated data. *Perception & Psychophysics, 38*(5), 476-481.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B*, *46*, 440–464.

Cudeck, R. & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin, 109*(3), 512-519.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General, 121*(3), 364-381.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333-353.

Dempster, A. (1974). The direct use of likelihood for significance testing. In *Proceedings of Conference on Foundational Questions in Statistical Inference, Aarhus, May 7-12, 1973*, 335. Dept. of Theoretical Statistics, Institute of Mathematics, University of Aarhus.

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

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

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics, 13*(4), 755-779.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. New York: Chapman & Hall.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*(4), 733-760.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of Psychiatric Research, 43*, 401-410.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*(3), 532-560.

Grünwald, P. (2007). *The Minimum Description Length Principle.* Cambridge, MA: MIT Press.

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

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301-323.

Hansen, M. H., & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association, 96*(454), 746-774.

Hayduk, L., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three–Testing the theory in structural equation models!. *Personality and Individual Differences*, *42*(5), 841-850.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 3*, 45-60.

Houts, C. R., & Cai, L. (2013). *flexMIRT® user's manual version 2: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using full-information item bifactor analysis for graded response data: An illustration with the state metacognition inventory. *Educational and Psychological Measurement, 68*, 695-709.

Irwin, D. E., Stucky, B. D., Langer, M. M., Thissen, D., DeWitt, E. M., Lai, J. S., Yeatts, K. B., Varni, J. W., & DeWalt, D. A. (2012). PROMIS pediatric anger scale: An item response theory analysis. *Quality of Life Research, 21*, 697-706.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Scientific Software International.

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

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 10*(3), 333-351.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 79-86.

Lee, T., Cai, L., & Kuhfeld, M. (in press). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal.*

Li, Z., & Cai, L. (2012, July). *Summed score likelihood based indices for testing latent variable distribution fit in item response theory.* Paper presented at the annual International Meeting of the Psychometric Society, Lincoln, NE. Retrieved from http://www.cse.ucla.edu/downloads/files/SD2-final-4.pdf

Longley, S. L., Calamari, J. E., Wu, K., & Wade, M. (2010). Anxiety as a context for understanding associations between hypochondriasis, obsessive-compulsive, and panic attack symptoms. *Behavior Therapy, 41*, 461-474.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8*, 453-461.

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

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics, 34*, 593-610

Marsh, H. W., & Balla, J. (1994). Goodness of fit in confirmatory factor analysis: The effects of sample size and model parsimony. *Quality and Quantity, 28*(2), 185-217.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*(3), 391-410.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement, 11*, 71-101.

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*(471), 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305-328.

Moody, J. E. (1992). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In Moody, J. E., Hanson, S. J., & Lippmann, R. P., (Eds.), *Advances in Neural Information Processing Systems*, 847-854.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm.

Applied Psychological Measurement, 16, 159-176.

Myung, I. J., & Pitt, M. A. (2001). Mathematical modeling. In Wixted, J., (Ed.), *Stevens' Handbook of Experimental Psychology, Vol. 4: Methodology.* New York: Wiley.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In Lamberts, K. & Goldstone, R., (Eds.), *Handbook of Cognition*. London, UK: Sage Publications Ltd.

Ory, D. T., & Mokhtarian, P. L. (2010). The impact of non-normality, sample size and estimation technique on goodness-of-fit measures in structural equation modeling: Evidence from ten empirical models of travel behavior. *Quality & Quantity, 44*(3), 427-445.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S\text{-}X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298.

Osman, A., Wong, J. L., Bagge, C. L., Freedenthal, S., Gutierrez, P. M., & Lozano, G. (2012). The Depression Anxiety Stress Scales-21 (DASS-21): Further examination of dimensions, scale reliability, and correlates. *Journal of Clinical Psychology*, *68*(12), 1322-1338.

Patrick, C. J., Hicks, B. M., Nichols, P. E., & Kreuger, R. F. (2007). A bifactor approach to modeling the structure of the Psychopathy Checklist-Revised. *Journal of Personality Disorders, 21*, 118-141.

Pickles, A., & Angold, A. (2003). Natural categories or fundamental dimensions: On carving nature at the joints and the rearticulation of psychopathology. *Development and psychopathology*, *15*(03), 529-551.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*(3). 472-491.

Popper, K. R. (1962). *Conjectures and Refutations, Volume 192.* New York: Basic Books.

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research, 41*(3), 227-259.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna,

Austria: R Foundation for Statistical Computing.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*. 19-31.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2012). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*, 5-26.

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika, 61*(3), 509-528.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465-471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics,* 416-431.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, IT-42*(1), 40-47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47*(5), 1712-1717.

Rissanen, J. (2007). *Information and complexity in statistical modeling*. Springer Science & Business Media.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (in press). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment.*

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

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

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph,* No. 17.

SAS Institute. (2014). *The SAS system for Windows. Release 9.4*. SAS Inst.., Cary, NC.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2). 461-464.

Simms, L. S., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety, 25*, E34-E46.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247.

Smith, N. A., & Tromble, R. W. (2004). Sampling uniformly from the unit simplex. *Johns Hopkins University, Tech. Rep. 29.*

Sonuga-Barke, E. J. (1998). Categorical models of childhood disorder: A conceptual and empirical analysis. *Journal of Child Psychology and Psychiatry, 39*(1), 115-133.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Spiegelhalter, D. J., Best, N., Carlin, B. P., & van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. *Research Report, 98-009*.

Stan Development Team (2014). *Stan Modeling Language Users Guide and Reference Manual, Version 2.5.0.* Retrieved from: http://mc-stan.org

Stone, C. A., & Zhu, X. (2015). *Bayesian Analysis of Item Response Theory Models Using SAS®*. Cary, NC: SAS Institute Inc.

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

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement, 26,* 247-260.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test Scoring*. New York: Routledge.

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological assessment*, *24*(1), 101-113.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology, 56*(2), 271-288.

van der Heijden, P., Hart, H., & Dessens, J. (1997). A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behaviour. In Rost, J. & Langeheine, R. (Eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences,* 196-208. New York: Waxmann Münster.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research, 2*(2), 29-48.

Whewell, W. (1840). Aphorism 39 from the Philosophy of the Inductive Sciences. In Y. Elkana (Ed.), *Selected writings on the history of science--William Whewell*. Chicago: Univ. of Chicago Press.

Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: An extension of the bi-factor item response theory model for use in differential item functioning. *Journal of Psychiatric Research, 43*, 1025-1035.