ABSTRACT

Misclassification Errors Informed by Response Time in Item Factor Analysis

R. Noah Padgett, Ph.D.

Chairperson: Grant B. Morgan, Ph.D.

The measurement process necessarily leads to observations measured with error to a degree. In education, researchers often want to obtain measurements of difficult-to-measure constructs such as content knowledge, motivation, affect, and personality. A scale is created using multiple items to triangulate the measurement of the construct of interest using the common information across items. One source of error that is not often accounted for is measurement error in the item response itself. In this study, I propose an approach for measuring latent traits while accounting for item-level measurement error. The proposed approach differentially weighs responses by how long an individual takes to respond to the item, i.e., response time as an absolute measure of time taken on each item−weighing responses by response time discounts the information provided by individuals responding rapidly to items. The result is that individuals with longer response times more heavily inform the estimation of the model, and more highly weighted responses are theorized to more accurately reflect the construct of interest. Utilizing more reliable information provides a foundational step in finding validity evidence for inferences made using scales.

The purpose of this study was two-fold. First, simulation studies were conducted to show how the proposed measurement can be estimated and demonstrate the effects of estimating traditional item-factor models when data are prone to item-level

measurement error. In these studies, I show that the parameter estimates (e.g., factor loadings, residual variances, etc.) may be severely upwardly or downwardly biased. The coverage rates for interval estimates of the parameters were also highly variable across conditions studied and parameters. The results showed that researchers' ability to make valid inferences about the underlying model is limited by how item-level measurement error is modeled. Secondly, the applied studies used data from the National Assessment of Educational Progress (NAEP) 2017 math assessment and an open-source dataset on extroversion. The results from these applied studies demonstrate the applicability of the proposed model and how inferences about reliability may be highly dependent on how item-level measurement error is modeled. Finally, implications and applications to educational research using the proposed methods are discussed.

Misclassification Errors Informed by Response Time in Item Factor Analysis

by

R. Noah Padgett, B.S., M.A.

A Dissertation

Approved by the Department of Educational Psychology

Todd Kettler, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Doctor of Philosophy

Approved by the Dissertation Committee

Grant B. Morgan, Ph.D., Chairperson

Nicholas F. Benson, Ph.D., NCSP

Sara Tomek, Ph.D.

John Seaman, Ph.D.

Accepted by the Graduate School
May 2022

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to acknowledge my friend, mentor, and advisor, Grant Morgan. Throughout my graduate school career, your support and guidance have been a blessing, and I would not have had so many opportunities without you. I certainly would have far fewer laughs and had far fewer gains. Thank you, Grant, for everything.

Next, I would like to acknowledge the support of my committee, Dr. Nicholas Benson, Dr. Sara Tomek, and Dr. John Seaman, throughout this process. Their support and thoughtful questions have helped shape the work throughout and helped push me to think better.

I would like to sincerely thank Dr. Leigh Greathouse for her support throughout my graduate education. She helped support and mentor me throughout my education.

I would like to acknowledge the support of my colleagues at the American Institutes for Research (AIR). My colleagues Sheniqua Jeffrey, Young-Yee Kim, James Zheng, Markus Broer, Juanita Hicks, Fusun Sahin, Ruhan Circi, and Lisa Yarnell at AIR helped tremendously to further my graduate education. These colleagues help also helped with the refinement of the use of the developed methods to NAEP data. I would like to thank the support of Dr. Emmanuel Sikali from the National Center for Educational Statistics. His support during my work at AIR with NAEP has been invaluable in helping shape my direction as a researcher and to help give me the confidence in my knowledge.

Finally, my sincere gratitude goes to my friends and family, who continuously supported me with their tremendous encouragement and coffee runs.

CHAPTER ONE

Introduction

Measurement error occurs in all scientific fields. The use of increasingly sophisticated statistical methods allows researchers to parse the error from the signal of interest. Measurement in educational and psychological investigations is especially prone to measurement error because of the reliance on self-report measures. Models for measurement error include classical test theory in psychometrics (Crocker & Algina, 1986; Lord & Novick, 1968), Berkson measurement error model used in epidemiology (Berkson, 1950), and errors in variables in applied statistics (Fuller, 1987). Understanding the processes that give rise to measurement has been well studied in a wide range of fields. The conceptualization differences across fields yield different underlying assumptions in how measurement error is modeled (Kroc & Zumbo, 2020). The modeling approach used can significantly impact the inferences one makes from available data, as shown in Rigdon et al. (2019) concerning differences in modeling a latent variable approach to modeling measurement error. In light of the differences in inferences that can occur solely due to modeling measurement error, the development of robust methods for accounting for measurement error is crucial to further scientific inquiry.

Typical areas of inquiry in educational and psychological research are based on modeling the interdependence of latent constructs. Attitudinal surveys commonly measure latent constructs. Surveys often operationalize measures into discrete statements with fixed response options (e.g., "Disagree" versus "Agree"). This approach's simple nature to create, implement, and score has led to many measurement tools developed using this format. However, the simple selected-response format may also lead to difficulties appropriately mapping the observed responses onto the latent con-

struct. Latent constructs have been connected to observed responses using classical test theory (Crocker & Algina, 1986), the Rasch model (Rasch, 1960), factor analysis (Brown, 2015), item response theory (de Ayala, 2009), and structural equation models (Bollen, 1989).

The modeling frameworks described above give researchers an approach for describing the measurement process of the construct of primary interest. Measurement error is commonly described as the imperfect relationship between the latent construct and observed response. However, measurement error is often overlooked in the observed responses. The item-level response may not be conceptualized as containing measurement error. For example, classical test theory focuses on the total (observed) score and true score relationship in terms of error, which does not address item-level measurement errors. Item level measurement error is similarly not addressed in commonly used measurement models for categorical outcomes such as the graded response model (Samejima, 1969), item factor analysis (Wirth & Edwards, 2007), and partial credit model (Masters, 1982). Such models do not directly incorporate measurement error for a particular item. The models above conceptualize measurement error post hoc by evaluating the difference between the observed response and an expected score. The expected score is also not required to be on the metric of the original measurement. Having expected scores on a different metric than the original items has the benefit of allowing for a nuanced investigation of the model adequacy. However, the measurement error in the original metric is not captured as part of the model. Not accounting for measurement error may lead to a false belief about the precision of our inferences because the uncertainty in the item-level measurement is ignored.

Models that account for measurement errors have been well documented (Fuller, 1987). A vast literature has grown on measurement error models for continuous outcomes (Bollen, 1989; Jöreskog, 1969), and there is a growing literature on generalized linear models (Carroll et al., 2006; Skrondal & Rabe-Hesketh, 2004). Of particular

interest in this study are methods for modeling measurement error in categorical variables. Measurement error in categorical variables is also called a misclassification error. A misclassification error, or error in measuring a discrete outcome, occurs when the observed categorical response is not the true value. The terms measurement error and misclassification error are used synonymously throughout this work. The errors represent when the observed categorical response is not the true response under ideal conditions.

Misclassification errors have been the focus of methodological research for many decades (Chen et al., 1984; Hochberg, 1977; Naranjo et al., 2019; Press, S. James, 1968; Sposto et al., 1992; Tenenbein, 1970; Yiu & Poon, 2008). However, models that directly account for misclassification are rare in psychological or educational measurement research. Goldstein et al. (2008) illustrated one of the few educational applications of measurement error models incorporated into a two-level linear model of mathematics achievement. Goldstein et al. (2008) demonstrated that, even with a relatively simple two-level model, inferences about model parameters (fixed and random effects) might change when misclassification errors are modeled. With misclassification errors, in particular, they found that incorporating misclassification "will often have little effect on the size of the coefficient but may be expected to increases its standard error." (Goldstein et al., 2008, p. 257).

The effects of misclassification are well demonstrated across contexts, with methods of accounting for such errors being advocated for decades (Goldstein et al., 2008; McGlothlin et al., 2008; Naranjo et al., 2019; Sposto et al., 1992). Advocates of misclassification methods have used expert opinion (Naranjo et al., 2019), historical data (Sposto et al., 1992), double sampling methods (Hochberg, 1977), or assumed known parameters (Roy et al., 2005) to identify misclassification rates or create informative priors for misclassification. Unfortunately, identifying the misclassification rate is one of the prevailing limitations of methods to account for misclassification.

Overcoming the limitation of identifying misclassification rates is the major aim of this work. I aim to show how informative priors of misclassification can be created at the individual response level in the context of computer-based test (CBT) from a single instance of administration.

The remainder of this work is outlined as follows. In Chapter 2, I provide a literature review of measurement models in educational research, misclassification methods in educational research, and joint models of item responses and response time in educational research. In Chapter 3, I introduce the proposed modeling framework for misclassification in item factor analysis. At the end of Chapter 3, I outline my proposed research questions related to the methods developed and hypotheses about these questions. In Chapter 4, I discuss the methods for studying the proposed models. In Chapter 5, I present the results from the two simulation studies and two applied studies. Lastly, in Chapter 6, I discuss the results and implications for using the methods developed in this project for modeling item-level measurement error.

CHAPTER TWO

Literature Review

This chapter thoroughly reviews the literature relevant to a joint modeling paradigm of item responses and response time in the educational and psychological assessment. The topics range in scope from the fundamentals of measurement models to methods for incorporating misclassification into analyses. I have broken this chapter into three notable sections: *Measurement Models of Latent Constructs*, *Misclassification Methods in Educational Research*, and *Measurement Models that Incorporate Response Time*.

Measurement Models of Latent Constructs

Education and social scientists have studied not directly measurable constructs for over a century. For example, Spearman (1904) worked on the measurement of general intelligence, a construct that is theorized to exist but cannot be directly measured. Factor analysis is the method for relating observed indicators to continuous latent variables, such as intelligence. Factor analysis is based on the common factor model, a general model for how the relationships among observed variables are explained by unobserved, latent variables (i.e., factors).

Factors are triangulated by investigating how responses to items covary. When items covary, we aim to explain why they covary by their relationship with the underlying construct (i.e., the factor of interest). When responses covary, the responses to one item are related to responses to another item. Based on how the items are theorized to group together, we then can form an expectation of the level of covariance among the items. We hypothesize that the factor model is then tested to see if we can explain the covariance among the set of items. Aiming to explain the item covari-

ance matrix is why factor analysis, specifically confirmatory factor analysis (CFA), is sometimes called covariance structure analysis (Bollen, 1989; Brown, 2015; Kline, 2016).

Over the years, vast methodological literature has amassed on CFA. Much of this literature is out of scope for this review. However, some seminal pieces are highlighted. In the development of CFA, Jöreskog (1967) helped progress the estimation of CFA under maximum likelihood. Later, Jöreskog (1969) helped build the groundwork for hypothesis testing in CFA and SEM more generally. Possibly the most significant contribution to the CFA literature was the development of the software program LISREL by Jöreskog and Sörbom, where they provided the technical capabilities to estimate a wide range of latent variable models (Jöreskog & Sörbom, 2015). The entirety of the work of Jöreskog and colleagues is out of scope for this review, but much of the future work on CFA rests on the shoulders of these giants.

Much of the work on CFA and SEM was brought together by Bollen (1989), whose text contains the synopsis of much of this early work on latent variable modeling. The general modeling framework of CFA is encapsulated in the following concise model, known as the common factor model:

$$\mathbf{Y}_p = \tau + \mathbf{\Lambda}\eta_p + \varepsilon_p, \tag{2.1}$$

where $\mathbf{Y}_p$ is the vector of observed item responses of person p. τ is the vector of intercepts that is typically fixed to a zero vector because of the use of standardized scores in the estimation. $\mathbf{\Lambda}$ is the factor loading matrix. η_p is the vector of factor scores for person p, and ε_p is the residual error. Equation 2.1 relates the model parameters to the observed responses. However, CFA is a method for modeling how theorized variables can explain patterns of responses. The theorized relationship among variables is often expressed through a path model. An example of a one-factor CFA path model is shown in Figure 2.1.

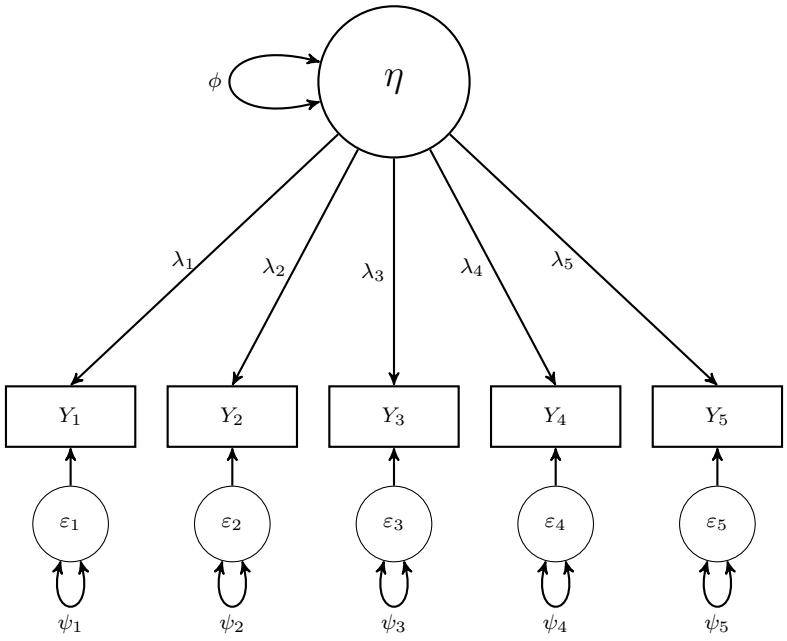


Figure 2.1. Example path diagram of a confirmatory factor analysis model.
Note. The error terms (ε_p) and corresponding residual variances (ψ_p) are typically excluded for simplicity.

Along with relating the observed responses to the latent variables, we need to specify how the variables relate to each other. Our expectation for the relationships among items is formulated as

$$Var(\mathbf{Y}) = \mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Psi}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Theta}, \tag{2.2}$$

where $Var(\mathbf{Y})$ is the model implied covariance matrix which is often shortened to $\mathbf{\Sigma}$. And, where $\mathbf{\Lambda}$ is the estimated factor loading matrix which is sometimes called the pattern matrix, $\mathbf{\Psi}$ is the estimated covariance/correlation matrix among the latent variables, $\mathbf{\Lambda}^{\mathrm{T}}$ is the transpose of the factor loading matrix, and $\mathbf{\Theta}$ is the residual covariance matrix among the observed variables. The residual covariance matrix is typically assumed to be a diagonal matrix, which means that once the factor structure imposed on these data is accounted for, no other relationship among items exists (Brown, 2015). The interested reader is referred to Bollen, 1989; Brown, 2015; Kline, 2016 for more information on CFA.

Categorical CFA - Item Factor Analysis

Traditional confirmatory factory analysis assumes that factor indicators are continuous measures and that these indicators are linearly related to the underlying factors. A continuous measure can (usually) be sufficiently described by its mean and variance, whereas data with discrete categories typically are not appropriately described by its mean and variance. Furthermore, categorical data are unlikely to have a linear relationship with the underlying factor(s). Social scientists rarely obtain continuous data as surveys often use ordered response scales (e.g., Likert-type responses). CFA models that fit with these data types are sometimes called *item factor analysis* and are closely related to some item response theory models (Bollen et al., 2010). The use of traditional CFA, particularly in conjunction with maximum likelihood methods, may not be valid given the restrictions of categorical data.

In educational and psychological research, a common approach to modeling the response to a survey is factor analysis or item response theory (Brown, 2015; de Ayala, 2009; Wirth & Edwards, 2007). In particular, in surveys and educational measurement, truly continuous data are rare. Instead, data are commonly collected using discrete categories to capture information about a respondent, for example, using a Likert-type response format to assess attitudes towards a topic. A commonly used method for analyzing such data is to treat the observed data as representing a discretized underlying continuous response. The process by which this discretizing occurs is described in (Mislevy, 1986; Muthén, 1984).

Let an ordinal response variable (y) take on values $c = 1, 2, \cdots, C$, where C is the total number of response options. The responses to I such items are hypothesized to reflect m latent traits ($\boldsymbol{\eta}$). We would like to relate $\boldsymbol{\eta}$ to y linearly; however, this is not possible due to the discrete nature of y. Instead, we presume that y is the observed manifestation of the categorization process

$$y_i = c, \text{ if } \tau_{c-1} < y_i^* \leq \tau_c, \tag{2.3}$$

where $\tau_0 = -\infty$, $\tau_c = \infty$, and y^* is the continuous latent response variable for item i. The threshold parameters (τ_c) may vary in magnitude and number across items/observed indicators. The linear relationship between y_i^* and η is now possible.

The relationship between the trait of interest, η, and the latent response variable, y_i^*, is modeled by the common factor model. The model for the vector of latent response $\boldsymbol{y}^*$ is

$$\boldsymbol{y}^* = \boldsymbol{\alpha} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2.4}$$

$$\boldsymbol{\Sigma}(\boldsymbol{y}^*) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2.5}$$

where $\boldsymbol{\alpha}$ is the vector of latent response intercepts (typically assumed to be $\mathbf{0}$ within a factor analytic framework), $\boldsymbol{\Lambda}$ is the factor loading matrix, η is the latent trait (typically assumed that $\eta \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Phi})$), $\boldsymbol{\varepsilon}$ is the vector of residual (typically assumed $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Theta})$), and $\boldsymbol{\Sigma}(\boldsymbol{y}^*)$ is the model implied covariance matrix among the latent response variables. In applications, a common assumption is that the item responses are *locally* or *conditionally independent* given the latent trait $\boldsymbol{\eta}$, which results in a diagonal error-covariance matrix, and each item is conditionally independent.

The above model is indeterminate regarding location, scale, and orientation arising from $\boldsymbol{\eta}$ and $\boldsymbol{y}^*$ being unobserved. The indeterminacy of scale and location can be resolved by restricting the parameter space of $\boldsymbol{\eta}$ and $\boldsymbol{y}^*$ to

$$\boldsymbol{\eta} \sim \text{MVN}(\mathbf{0}, \boldsymbol{I}_m)$$

$$\text{diag}\left[\boldsymbol{\Sigma}(\boldsymbol{y}^*)\right] = \mathbf{1} \Rightarrow \theta_i^2 = 1 - \sum_{k=1}^{m} \lambda_{ik}^2.$$

Other restrictions can be made that allow for different interpretations of the model parameters. For example, the factor covariance matrix need not be diagonal or assume unit variances for the factors if the scale and orientation are set by restricting a factor loading to one for each factor. Kamata and Bauer (2008, Table 1, p. 139) described a several approaches to resolving the indeterminacy of the item factor analysis model. The approaches to resolving the indeterminacy are summarized in Table

Table 2.1

Parameterizing latent response formulation of factor analysis with categorical data

	Factor Scale	
LRV Scale	Reference Indicator	Standardized Factor
Marginal	$\lambda_1 = 1$, $\tau_1 = 0$, $V(y^*) = 1$	$E(\eta) = 0$, $V(\eta) = 1$, $V(y^*) = 1$
Conditional	$\lambda_1 = 1$, $\tau_1 = 0$, $V(\varepsilon) = 1$	$E(\eta) = 0$, $V(\eta) = 1$, $V(\varepsilon) = 1$

Note. Marginal = latent response variable has fixed unit variance (total variance, i.e., $V(y^*) = 1$); Conditional = latent response variable has fixed unit residual variance (total variance of latent response variable changes, i.e., $V(\varepsilon) = 1$); λ = factor loading; τ = item threshold; $E(\eta)$ = factor mean; and $V(\eta)$ = factor variance. This table was adapted from Kamata and Bauer (2008) for ease of discussion.

2.1. The labels of marginal versus conditional come from whether the scale of the latent response variable is set directly (marginal) or whether the scale of the latent response variable is achieved through fixing the residual variance (conditional). From testing these approaches, I've found that the setting the scale by fixing the factor variance ($V(\eta) = 1$) and residual variance ($V(\varepsilon) = 1$) tends to result in more efficient sampling when estimating the model using JAGS (Just Another Gibbs Sampler; Plummer et al., 2003) resulting in less time needed to achieve posterior convergence. The orientation indeterminacy can also be resolved by restricting all factor loadings to be positive. Restricting the range of estimates of a parameter is commonly employed in item response theory to estimate item discrimination parameters (Levy & Mislevy, 2016, p. 256).

Once indeterminacy is resolved, the probability of the observed response can be obtained using Equation 2.3. The use of the threshold scheme in Equation 2.3 implies that the observed response for a single item is

$$\begin{aligned} Pr(y_i = c \mid \eta) &= Pr(y_i^* \geq \tau_{c-1} \mid \eta) - Pr(y_i^* \geq \tau_c \mid \eta) \\ &= F\left(\frac{\lambda_i \eta - \tau_{c-1}}{\theta_i}\right) - F\left(\frac{\lambda_i \eta - \tau_c}{\theta_i}\right). \end{aligned} \tag{2.6}$$

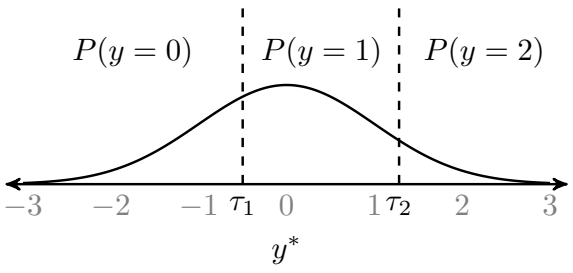


Figure 2.2. Latent response curve and response probabilities for three ordered categories model

The link function (F) is commonly chosen to be the probit ($\Phi(\cdot)$) or logit ($\Psi(\cdot)$) link. An example of this with three response categories is shown in Figure 2.2. The category probabilities can then be computed as the difference between the inequalities $Pr(y = 2) = Pr(y^* \geq \tau_2)$, $Pr(y = 1) = Pr(y^* \geq \tau_1) - Pr(y^* \geq \tau_2)$, and $Pr(y = 0) = 1 - Pr(y^* \geq \tau_1)$. This process is commonly used in item factor analysis (Wirth & Edwards, 2007) and is similar to an ordinal regression model (Agresti, 2010).

The latent response formulation provides a useful framework for developing factor models with nonnormal data. The observed indicators' measurement error is implicitly modeled where item-level measurement error is the difference between the observed categorical response and the continuous latent response. Conceptually, this approach is appealing, especially for scales that use degrees of agreement as to the response anchors. However, this approach does not account for the measurement error in the original metric of the data. Measurement error of discrete outcomes, such as survey item responses, is commonly referred to as a misclassification error. The following section focuses on accounting for misclassification errors in educational data.

Misclassification Methods in Educational Research

The discrete outcomes common in educational contexts are prone to measurement error, and the measurement error of a discrete outcome is commonly known as a misclassification error (Fuller, 1987). Misclassification is an error in measuring a

variable with a small number of discrete possible values. For example, a child may be identified as eligible for the free and reduced lunch program (national school lunch program; NSLP) when they do not qualify. Measurement error in NSLP status was cited as a prevalent issue in educational research by Goldstein et al. (2008). Therefore, the error in identification may be described as a misclassification error.

A misclassification error can be expressed as a probability that an observed discrete measure, y, is not equal to the actual discrete value if measured without measurement error, Y. Suppose that Y is dichotomous, then

$$Pr(x = a|X = b) = p_{ab}. \tag{2.7}$$

Misclassification in the context of educational data can occur when attempting to diagnose a disorder (Falley et al., 2018). Falley and colleagues (2018) investigated how misclassification can be accounted for in the analysis of diagnosing a mathematics disorder (a dichotomous outcome) using a set of covariates that are also possibly measured with error. The methods for accounting for misclassification were found to shift the coefficient estimates to a large degree. They will substantially increase the uncertainty in the estimates (i.e., wider credible intervals) compared to an analysis that ignored potential misclassification. Misclassification changing the parameter coefficients contrasts the finding of Goldstein et al. (2008). Goldstein and colleagues (2008) found that the coefficient will negligibly change magnitude in either direction when misclassification is accounted for, but that uncertainty will increase (larger standard errors). The analyses from the two teams of researchers used different model estimation methods-Bayesian methods (Falley et al., 2018) and maximum likelihood (Goldstein et al., 2008)-which may account for a small degree of the difference in findings of the change in coefficient estimates.

Despite minor differences in conclusions between studies, the impact of misclassification on conclusions has been repeatedly demonstrated (McInturff et al., 2004; Roy et al., 2005; Roy et al., 2013). For example, Burstyn et al. (2014) demonstrated

how a relatively small degree of misclassification of a covariate (5-10%) might change the expected Type II errors rates by at much as 10-15%, on average. Furthermore, authors have demonstrated that misclassification can severely impact parameter estimates. However, disagreements exist regarding the specific effects of misclassification. The dependent factor is the specific analysis conducted and the misclassification rates' assumptions. Researchers have tested the effects of misclassification on results in logistic regression (McGlothlin et al., 2008; Roy et al., 2005; Roy et al., 2013), multilevel linear models (Goldstein et al., 2008), multilevel count models (Nelson et al., 2018), or ordered logistic regression for polytomous outcomes (Eickhoff & Amemiya, 2005; Naranjo et al., 2019). In general, the effects of misclassification appear to be a potential biased estimate of regression coefficients (upwardly or downwardly depending on the analysis) and smaller estimates of uncertainty (standard errors or credible interval widths). Failing to develop approaches to account for such potential sources of bias in results can lead to a false belief about the certainty in one's results.

One understudied context is the effects of misclassification on attitudinal surveys commonly used in educational and psychological research. Measurement error is a common concern, and methods to help model and account for measurement error in the assessed trait are widely used. However, many of the measurement error models in educational research are concerned with measuring the trait. One potential limitation in focusing on trait-level measurement error is that item-level measurement error is not addressed. Addressing item-level measurement error was partially addressed by Yiu and Poon (2008) and Roy and Banerjee (2009). Yiu and Poon (2008) developed an approach to estimating polychoric correlations under misclassification, and Roy and Banerjee (2009) developed a multivariate probit model under misclassification. Polychoric correlations are commonly used in factor analysis as the observed information in a limited information estimation method such as diagonally weighted least squares for estimating CFA models with categorical indicators (Bandalos, 2014; DiStefano &

Morgan, 2014). Probit models are also closely linked to IRT models. However, neither approach is meant to help measure an underlying trait being measured by the set of indicators.

In the next section, measurement models that incorporate response time are introduced. In the following chapter, response time is used to indicate the misclassification rate. The proposed approach for using response time relies on the psychological theory underlying the response process to attitudinal survey items. A class of measurement models has been developed that incorporates response time (Molenaar, Tuerlinckx, & van der Maas, 2015b). Those methods are introduced next.

Measurement Models that Incorporate Response Time

Various methods have been proposed to incorporate response time into an educational and psychological measurement. The methods for modeling response time, especially reaction time, have a long history in psychology. Much of the literature on how researchers have modeled response time directly is the scope for this work, but Luce (1986) provided a comprehensive work on the subject through the '80s. Schnipke and Scrams (2002) and De Boeck and Jeon (2019) provide more recent overviews of this area of research. However, this study is focused on methods that use response time in conjunction with responses to a primary stimulus (e.g., an item on a psychological assessment). Response time for measuring non-cognitive traits such as personality is a less understood domain and is the focus of this work.

Measurement models that use response time in conjunction with responding to test items, in particular, are well documented in educational assessment (Thissen, 1983). Such joint models have been applied to detecting cheating (Boughton et al., 2017) and accounting for insufficient effort responding to surveys (Bowling et al., 2021; Curran, 2016; Dunn et al., 2018; Huang et al., 2012). For these uses, a wide range of statistical models has been developed. A hierarchical model that jointly models

both is recommended to analyze test items using item responses and response times (Becker et al., 2021; De Boeck & Jeon, 2019; Entink et al., 2009; Fox & Marianti, 2016; van der Linden, 2007). The hierarchical approach to jointly modeling responses and response times can be seen as a bivariate generalized linear model using item response and response time as the two outcomes to model. The individual responses are assumed to be locally independent, given the person's abilities. Local, or conditional, independence is a common assumption in IRT models to help build the model that describes the response process for the individual items. The same local independence assumption is expanded to include the response times. The items and response time are assumed independent after conditioning on the trait ability and a random effect interpreted as the person's speed. A multivariate normal distribution has been used to model the latent ability and latent speed factors(van der Linden, 2007). Assuming normality of latent variables is common in IRT, CFA, and other latent variable modeling traditions (Bollen, 1989; de Ayala, 2009; Kline, 2016).

Latent variables describe individual differences among respondents on item and response times. Individual differences are modeled by the joint distribution of the latent variables. van der Linden (2007) modeled the joint distribution of trait (η_{1p}) and speed (η_{2p}) latent variables by a multivariate normal distribution. That is, the joint distribution is

$$\boldsymbol{\eta}_p \sim f(\eta_p; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}), \tag{2.8}$$

$$f(\eta_p; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}) = \frac{\left|\boldsymbol{\Sigma}_{\mathcal{P}}^{-1}\right|^{1/2}}{2\pi} \exp\left[-\frac{1}{2}\left(\boldsymbol{\eta}_p - \boldsymbol{\mu}_P\right)^T \boldsymbol{\Sigma}_{\mathcal{P}}^{-1}\left(\boldsymbol{\eta}_p - \boldsymbol{\mu}_{\mathcal{P}}\right)\right], \tag{2.9}$$

$$\boldsymbol{\mu}_{\mathcal{P}} = (\mu_{\eta_1}, \mu_{\eta_2}), \tag{2.10}$$

$$\boldsymbol{\Sigma}_{\mathcal{P}} = \begin{pmatrix} \sigma_{\eta_1}^2 & \sigma_{\eta_1\eta_2} \\ \sigma_{\eta_1\eta_2} & \sigma_{\eta_2}^2 \end{pmatrix}, \tag{2.11}$$

where $p = 1, \cdots, N$ observations are drawn the hypothetical population $\mathcal{P}$ in the form of draws from the bivariate normal distribution $f(\eta_p; \boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}})$ is the density

function with mean vector $\boldsymbol{\mu}_{\mathcal{P}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathcal{P}}$. The above model captures the individual differences among respondents by capturing a random effect for ability and a random effect for speed of responding to items. The abilities can be thought of as level-2 of the hierarchical model. The level-1 part of the hierarchical model describes how the person parameters connect to item parameters to describe the observed responses to items and response times.

For responses to items, any number of IRT, IFA, or CFA models can be constructed to describe how a person's ability (η_p) is related to observed responses. Within the IRT tradition for dichotomous items, a 3-parameter normal-ogive (3PNO) or logistic (3PL) is a common choice of model as described by van der Linden (2007) when he illustrated the hierarchical approach. For my purposes, the measurement model for an item response is $y_p \sim f(y_p; \eta_p, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ denotes the vector of item parameters for item i.

For item response times, a variety of different models have been proposed. van der Linden (2007) used a lognormal model for the response times (t_{pi}), that is

$$t_{pi} \sim f(t_{pi}; \eta_{2p}, \boldsymbol{\psi}_i), \tag{2.12}$$

$$\boldsymbol{\psi}_i = (\alpha_i, \beta_i), \tag{2.13}$$

$$f(t_{pi}; \alpha_i, \beta_i) = \frac{\alpha_i}{t_{pi}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left[\alpha_i\left(\log(t_{pi}) - (\beta_i - \eta_{1p})\right)\right]^2\right]. \tag{2.14}$$

The response time parameters, $\alpha_i \in \mathcal{R}^+$ and $\beta_i \in \mathcal{R}$, represent the discrimination (factor loading) and time-intensity (location) parameters for the measurement model of response time, respectively. However, a variety of other models exist for describing the response time information (Molenaar, Tuerlinckx, & van der Maas, 2015a; Ranger & Kuhn, 2018; Roskam, 1987; Rouder et al., 2003; Thissen, 1983; van der Maas et al., 2011; Verhelst et al., 1997). In addition, many developed approaches are special cases of the bivariate generalized linear IRT (B-GLIRT) discussed by Molenaar, Tuerlinckx, and van der Maas (2015b).

The B-GLIRT generalization of methods for jointly modeling item responses and response times by Molenaar and colleagues (2015) provided a framework encompassing many previously developed methods. The general model for item responses y_{pi} and response times t_{pi} for person p on item i is described by

$$\begin{aligned} \mathrm{E}(Z_{pi}) &= g_Y\left[E\left(X'_{pi}\right)\right] = \lambda_i \eta_{1p} + \tau_i \\ \text{with } \mathrm{var}(Z_{pi}) &= \sigma^2_{\varepsilon i} \end{aligned} \tag{2.15}$$

$$\begin{aligned} \mathrm{E}(W_{pi}) &= g_T\left[E\left(T'_{pi}\right)\right] = \alpha_i \eta_{2p} + \beta_i + f(\eta_{1p}; \boldsymbol{\rho}) \\ \text{with } \mathrm{var}(W_{pi}) &= \sigma^2_{\omega i}. \end{aligned} \tag{2.16}$$

The linear function of item and person parameters for the item response may differ depending on the parameterization, scale of the item responses and latent variables (binary, ordinal, nominal, continuous), and type of link function used. $g_Y(.)$ is the link function between expected item response and the latent variable (η_{1p}) (e.g., $g_Y(.)$ may be the normal ogive model for dichotomous item responses); λ_i item discrimination parameter; τ_i item location/difficulty parameters; η_{1p} is the latent ability/level of construct for person p; Z_{pi} is the response variable after applying the link function. We can think of this as the underlying latent response variable that is linearly related to the construct(Mislevy, 1986; Muthén, 1984; Wirth & Edwards, 2007); $\sigma^2_{\varepsilon i}$ is the residual variance of Z_{pi} that is not explained by the latent trait. $g_T(.)$ is the link function between the expected response time and the linear combination of response time parameters. The $g_T(.)$ link will commonly be an identity link function. α_i is a time discrimination parameter akin to a factor loading of the item to the underlying latent speed parameter. β_i is a time-intensity parameter that helps measure how long the item takes on average to account for differences in response time per item due to item length or other characteristics not related to the person. η_{2p} is the person speed parameter to capture individual differences in speed across all items W_{pi} is the response time variable after applying a link function which is commonly the log

response time but may vary depending on how response time is measured. $\sigma^2_{\omega i}$ is the residual variable associated with the response time regression function for item i. $f(.)$ is the cross-relation function with parameters $\boldsymbol{\rho} = [\rho_1, \rho_2, ...]$, which will depend on the theory used to inform the relationship between ability/latent construct and the response time. More information about the model parameters can be found in Molenaar, Tuerlinckx, and van der Maas (2015b, p. 58-59).

Additionally, the choice of link functions determines the specific B-GLIRT model. For example, choosing $g_T(.)$ to be the identity function when analyzing the log response time and $g_Y(.)$ to be a probit link for dichotomous responses results in the models described by van der Linden (2007) and Fox et al. (2007). Molenaar, Tuerlinckx, and van der Maas (2015b) described other possible link functions and connections to previous models.

Another innovative approach to incorporating response was proposed by Wise and Demars (2006), who described an "effort-moderated" IRT model for dichotomous items on cognitive tests. The effort-moderated IRT model describes the response process as a two-component mixture model where each response is classified as either a solution behavior (SB) or rapid guess. SB is a fixed value of 1 if the item response time exceeds a specified threshold or 0 otherwise. Wise and Demars (2006) chose the threshold variable for each item after looking at the response time distribution and selecting what appears to be a logical cut-off between SB and a rapid guess. Another approach would be to use a mixture model directly to estimate the threshold as done by (Schnipke & Scrams, 1997). The effort moderated IRT model provides one approach researchers have used to incorporate response as a source of error among respondents.

Utilizing all available information about the response process is necessary for achieving more confidence in research findings. In this study, response time is used in non-cognitive measurement as a source of information on the amount of measurement

error. In the next chapter, the methods for modeling misclassification as a source of error in measurement are discussed. The measurement models discuss how response time can be modeled and incorporated into an item factor analysis framework of modeling while accounting for item-level misclassification.

CHAPTER THREE

Modeling Misclassification in Item Factor Analysis

In this chapter, the methods developed for this study are developed. The methods of item factor analysis, introduced in chapter 2, are expanded to incorporate item-level misclassification. The response time is then described as incorporated into the full misclassi- fication model. Finally, the proposed research questions to evaluate the model are described, and hypotheses are given.

Misclassification in Item Factor Analysis

Observations using categorical responses may be prone to measurement error. Measurement error for categorical responses is also known as a *misclassification error* because of the discrete nature of the data collected. A response would be considered *misclassified* if the observed response does not match what response would occur under ideal conditions. Uncertainty about the relationship between the observed and true categorical response can be incorporated by developing a model to explain how the observed response is likely related to the unobserved "true" response. The "true" response is not known so a natural choice would be to assume that the observed response is the true response. This is assumption is commonly use in educational and psychological measurement. However, suppose this assumption is not tenable, then an approached is needed for mapping an observed response to the unknown true response. In a general measurement model, this relationship between the observed response, true response, and latent variable can be described as shown in Figure 3.1. The latent response variable is then related to the unknown true categorical response instead of the observed response. Adding this layer to the measurement model provides a direct approach for incorporating item-level misclassification into the measurement

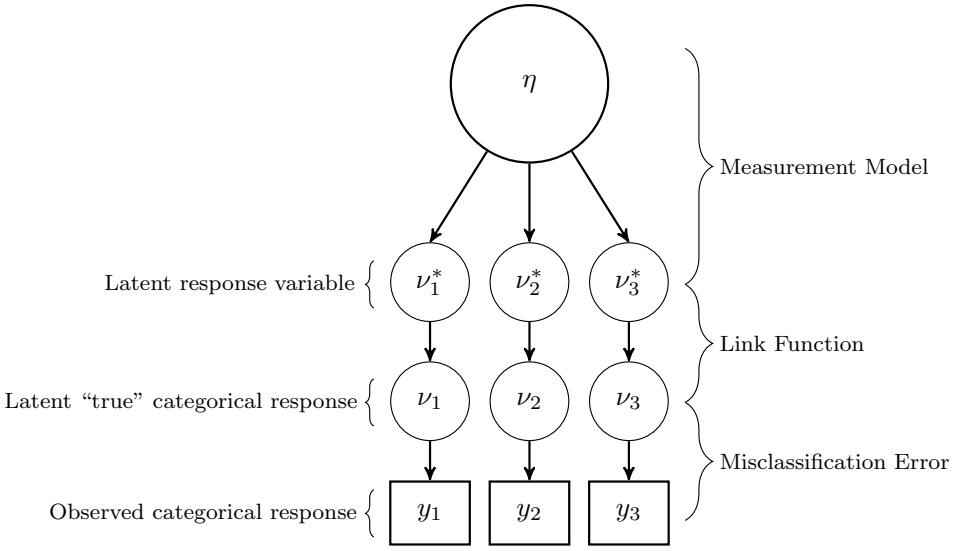


Figure 3.1. Measurement error components

of the latent trait. Mapping the unknown true response to the observed response is accomplished using a conditional probability statement; namely, the probability of the observed response depends on the unknown response (i.e., $Pr(Y \mid \nu)$, where Y is the observe response and ν is the unknown but true response). The conditional probability statement for the observed response can then be combined with the probability model for the true response (i.e., $Pr(\nu)$) to develop the probability model for the observed data.

The probability model for the observed response is defined to be composed of (1) the conditional probability of the observed data given the unknown true response, and (2) the probability model for the true response. The laws of probability can combine both pieces of information to arrive at the overall probability of the observed data. Suppose the observed categorical variable Y takes on values $a = 1, 2, \cdots, C$, and the unobserved categorical variable ν similarly takes on values $b = 1, 2, \cdots, C$. Using (1) and (2) above requires *marginalizing* over all possible values of the unobserved variable ν. Because ν is categorical, marginalizing means summing over the values of

ν to obtain the resulting probability distribution for the observed data Y, which is

$$Pr(Y = a) = \sum_{b=1}^{C} Pr(Y = a \mid \nu = b) Pr(\nu = b). \tag{3.1}$$

The set of all possible $Pr(Y = a \mid \nu = b)$ creates a misclassification matrix. This misclassification matrix is denoted as $\mathbf{\Gamma}$.

Using prior knowledge about how likely misclassified responses are to likely occur, the matrix of probabilities can be built to relate the observed response to the unknown, true item response. The elements of the misclassification matrix will represent a full set of conditional probabilities (e.g., $\gamma_{12} = Pr(Y = 1 \mid \nu = 2)$). For example, an item with three response options would generally be

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}. \tag{3.2}$$

The rows represent the observed response, and the columns represent the unknown true categorical response. The rows of $\mathbf{\Gamma}$ sum to one because the response has to be one of the three options unless a response is missing. The above representation of the misclassification probabilities can form hypotheses about the unknown *true* response ν_{pi}. That is, elements of $\mathbf{\Gamma}$ can be constrained to equality or specific values based on how misclassification of responses is theorized occur. The focus of modeling then shifts from the observed data to the unobserved data, which is theorized to underlie the observed data.

Measurement Model with Misclassification

Next, the measurement model for the unobserved true response is defined. The measurement model incorporates how misclassification relates the unobserved true response to the observed response. The misclassification parameters are defined through a single prior assumed applicable to all respondents and items. Extending the model

with informed misclassification rates is the focus of the next section after the model is defined.

Suppose a researcher obtains $\mathbf{Y}$ responses from $p = 1, \cdots, N$ individuals that responded to $i = 1, \cdots, I$ items. All I items are ordinal with C response categories. The I items are theorized to measure a single trait. An individual is theorized to require more of the trait to endorse a higher response category for each item. The item responses are also prone to misclassification. Prior information is not available to hypothesize how misclassification rates may vary across items or respondents. Therefore, a single misclassification rate is assumed plausible (this assumption will be relaxed later). Let the probability of the observed category c be $\omega_{pic} = p(y_{pi} = c)$. The set of $\boldsymbol{\omega}_{pi} = (\omega_{pi1}, \omega_{pi2}, \cdots, \omega_{piC})$ is the collection of category response probabilities. The observed responses are assumed to follow a categorical distribution: $y_{pi} \sim \text{Categorical}\,(\boldsymbol{\omega}_{pi})$. The Categorical distribution is a multi-category generalization of the Bernoulli distribution. This distribution is defined as $p(y = a \mid \boldsymbol{\omega}) = \frac{\omega_a}{\sum_{c=1}^{C} \omega_c}$.

The observed category response probabilities are defined by using Equation 3.1 with the item factor analysis model to describe the unobserved categorical response probabilities. That is, the observed and unobserved categorical responses are modeled by

$$
\begin{aligned}
y_{pi} \mid \boldsymbol{\omega}_{pi} &\sim \text{Categorical}\,(\boldsymbol{\omega}_{pi}) \text{ for } p = 1, \ldots, N, i = 1, \ldots, I \\
\boldsymbol{\omega}_{pi} &= (\omega_{pi1}, \omega_{pi2}, \ldots, \omega_{pic}) \\
p(y_{pi} = c \mid \boldsymbol{\Gamma}_i, \nu_{pi}) &= \sum_{b=1}^{C} \gamma_{ibc} Pr(\nu_{pi} = b) \\
\gamma_{ibc} &= Pr(y_{pi} = c \mid \nu_{pi} = b) \\
Pr(\nu_{pi} = b \mid \eta_p, \lambda_i, \boldsymbol{\tau}_i, \theta_i) &= \Phi\left((\nu^*_{pi} - \tau_{i,b-1})\theta_i^{-1}\right) - \Phi\left((\nu^*_{pi} - \tau_{i,b-1})\theta_i^{-1}\right), \\
\nu^*_{pi} \mid \eta_p, \lambda_i, \theta_i &\sim \text{N}\left(\lambda_i \eta_p, \theta_i^2\right)
\end{aligned}
$$

where $\mathbf{\Gamma}_i$ represents the misclassification parameters for item i. ν_{pi} represents the unobserved true response. ν^*_{pi} represents the latent response variable on item i by individual p. η_p represents the latent variable value for person p. The item parameters λ_i, $\boldsymbol{\tau}_i$, and θ^2_i represent the factor loadings, item thresholds, and residual variance of latent response variables, respectively.

The priors for the parameters were specified based on the recommendations of Levy and Mislevy (2016, p. 292-295) for Bayesian psychometric modeling. Levy and Mislevy (2016) consider the set of priors described below to be to be *relatively uninformative* compared to priors using content knowledge to inform these distributions.

$$\nu^*_{ip} \sim \mathrm{N}(\lambda_i \eta_p, \theta^2_i) \text{ for } p = 1, ..., N, i = 1, ..., I$$

$$\lambda_i \sim \mathrm{N}^+(0, 4) \text{ for } i = 1, ..., I$$

$$\eta_p \sim \mathrm{N}(0, 1) \text{ for } p = 1, ..., N$$

$$\theta^2_i = 1 \text{ for } i = 1, ..., I$$

$$\tau_{ib} \sim \begin{cases} \mathrm{N}(0, 10) \text{ for } b = 1 \\ \mathrm{N}(0, 10) I(\tau_{i,b-1} < \tau_b) \text{ for } b > 1 \end{cases}$$

$$\boldsymbol{\gamma}_{ib} \sim \mathrm{Dirichlet}(\alpha_{b1}, \alpha_{b2}, \cdots, \alpha_{bC})$$

The above set of priors are considered relatively uninformative because the distributions cover the full range of commonly observed magnitudes for all parameters. For example, unstandardized factor loadings are commonly between zero and four. Additionally, some prior distributions were specified with certainty to values that conform with common practice of how the scale is established for the latent factor and latent response variables. For the latent factor, the location and variability are set to 0 and 1, respectively. The location is a function of the latent variable and factor loadings for the latent response variables, while the variability is fixed to 1. This specification of the latent variables allows the latent factor to be interpreted as standardized and the latent response variables to be conditional on the values of the latent factor (see

Table 2.1 for more details). Other approaches to resolving the indeterminacy of the location of scale or the latent variables can be adopted. The indeterminacy of the orientation is the last component of the latent variable specification that is resolved by restricting the factor loadings to all be positive.

Alternatively, other prior structures have been proposed. For example, van der Linden (2007) used a multivariate normal distribution to specify a prior on the item parameters jointly. The draws from the unrestricted multivariate normal distribution were then transformed to the scale of the variable (e.g., exponential to discrimination and inverse-logit transformation for guessing). This alternative prior structure will be discussed again in the methods and results as part of a prior-to-posterior sensitivity analysis of selected parameters.

The likelihood function under the proposed model is

$$p(\boldsymbol{Y} \mid \boldsymbol{\Gamma}, \boldsymbol{\nu}^*, \boldsymbol{\eta}, \boldsymbol{\Lambda}, \boldsymbol{\tau}) = \prod_{p=1}^{N} \prod_{i=1}^{I} \omega_{pic} = \prod_{p=1}^{N} \prod_{i=1}^{I} p(y_{pi} = c \mid \boldsymbol{\gamma}_c, \nu^*_{pi}, \eta_p, \lambda_i, \tau_{ib}). \tag{3.3}$$

The observations are assumed to be independent, and the item responses are modeled as conditionally independent given the latent variable. The corresponding posterior is

$$\begin{aligned} p(\boldsymbol{\Gamma}, \boldsymbol{\nu}^*, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\Lambda} \mid \boldsymbol{Y}) &\propto \prod_{p=1}^{N} p(\boldsymbol{y}_p \mid \boldsymbol{\Gamma}, \boldsymbol{\nu}^*_p, \boldsymbol{\eta}_p, \boldsymbol{\Lambda}, \boldsymbol{\tau}) p(\boldsymbol{\Gamma}) p(\boldsymbol{\nu}^*_p) p(\boldsymbol{\eta}_p) p(\boldsymbol{\Lambda}) p(\boldsymbol{\tau}) \\ &\propto \prod_{p=1}^{N} \prod_{i=1}^{I} p(y_{pi} = c \mid \boldsymbol{\gamma}_c, \nu^*_{pi}, \boldsymbol{\eta}_p, \boldsymbol{\lambda}_i, \boldsymbol{\tau}_i) \times \\ &\qquad p(\boldsymbol{\gamma}_c) p(\nu^*_{pi}) p(\eta_p) p(\lambda_i) \prod_{b=1}^{C-1} p(\tau_{i,b}) \end{aligned} \tag{3.4}$$

The model described in Equation 3.4 is similarly depicted in Figure 3.2. Again, the model specification diagram clearly defines how the observed category response probabilities function the misclassification parameters ($\boldsymbol{\Gamma}$) and the response probabilities informed by the item factor analysis model (π).

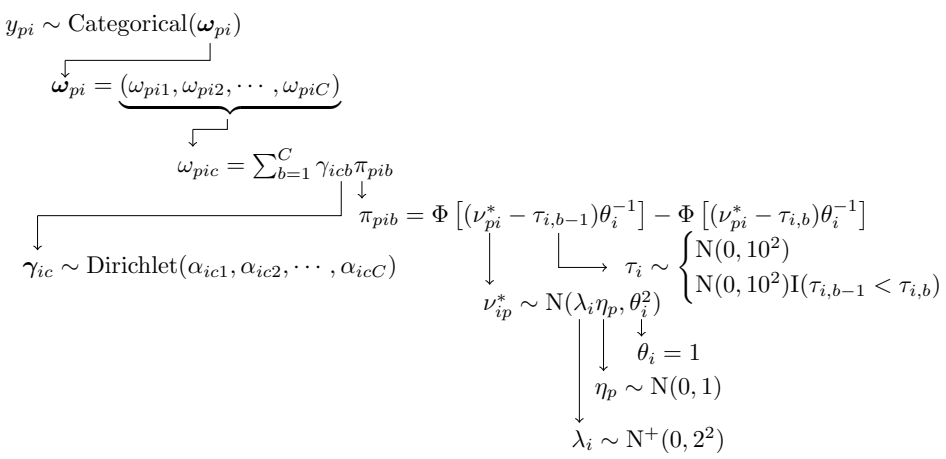


Figure 3.2. Item factor analysis with misclassification

Calibrating Misclassification Error Rates

Misclassification error is directly incorporated into the analysis through the matrix $\boldsymbol{\gamma}$. The elements of $\boldsymbol{\gamma}$ are unidentified in the sense that data from a single test administration are unlikely to update the prior distribution. Defining what the elements could *defensibly* be is then the challenge. I propose that information about the misclassification matrix may be contained in the item response time. Response time data has been shown to increase the precision of estimation of trait scores when incorporated as auxiliary variables in an IRT model (Bolsinova & Tijmstra, 2018). Researchers have developed hypotheses about how response time connects to the underlying latent trait of interest in the modeling item responses in personality assessment. One such hypothesis is the "distance-difficulty hypothesis" (Ferrando & Lorenzo-Seva, 2007a, 2007b; Meng et al., 2014; Molenaar et al., 2021), also known as the inverted-U effect (Kuiper, 1981). The inverted-U effect states that as the distance between a person's ability and the location of an item decreases, the respondent needs more time to respond. This hypothesis implies that individuals may respond rather quickly if they do not perceive that the item is meaningful, introducing construct irrelevant variance into the measurement process. Specifically,

> In personality and attitude testing, if [ability/level of latent trait] is just below [item location] and the subject has to respond quickly, the noise factor in the decision process will play a large role. However, the longer a subject thinks about his or her position, the more likely it becomes that he or she will select the answer option that best fits his or her latent state(van der Maas et al., 2011, p. 342).

A very low response time may then inform the amount of noise introduced into the measurement process. In summary, a longer response time will be treated as indicative of more precision in measuring an individual response. The relationship between response time and measurement error is theorized to be monotonically increasing where response time is measured absolutely on each item for all respondents.

Additionally, the degree of measurement error may also be determined by the distance between a person and item (i.e., the inverted-U effect). So, if we can combine the information about response time and the distance between item-person would give us a more informative measure of the potential degree of measurement error. I propose using the measurement model of response time and the item-person distance as an informative source of measurement error in the observed item response in

$$ELRT_{pi} = \frac{\exp\left(\beta_i - \eta_{2p}\right)}{f(\eta_{1p};\rho)} = \frac{Response\ Time}{Person\ Item\ Distance}. \tag{3.5}$$

$ELRT_{pi}$ is the *effective latent response time* for person p on item i. The use of $\beta_i - \eta_{2p}$ (latent response time) aims to account for the measurement error of response time across items. Next, the latent response time is exponentiated to scale the response time to the positive reals. Then, dividing by the distance function, $f(\eta_p;\rho)$, weights each item response by the individual varying inverted-U effect. Weighting by the inverted-U effect will increase the effective item response time if the distance is small but decrease if the distance is large. Based on the inverted-U hypothesis, a high response time corresponds to a low measurement error, whereas a low response time corresponds to a high measurement error. A link function, denoted $g_E(.)$, is then constructed to map $ELRT_{pi}$ to the item response.

The response time is linked the item response by the misclassification matrix. The misclassification probabilities, $P(y_{pi} = c \mid \nu_{pi} = b)$, are informed by the response time. Informing the elements of the misclassification matrix could then be done by

$$P(y_{pi} = c \mid \nu_{pi} = c) = \frac{1}{1 + \exp\left[-ELRT_{pi}\right]} \tag{3.6}$$

$$P(y_{pi} = c \mid \nu_{pi} = b) = \left(\frac{1}{C-1}\right)\left(1 - \frac{1}{1 + \exp\left[-ELRT_{pi}\right]}\right), \tag{3.7}$$

where C is the number of response categories, equation 3.6 says the link function between the ELRT and probability that the observed response is the "true" response is the inverse logit transformation. Whereas equation 3.7 is the probability of that the observed response is not the "true" response equally divided among the remaining categories. This relatively simple approach would create a symmetric and unique matrix for each participant on each item that can inform the degree of measurement error. The diagonal elements of this matrix would represent the probability that the observed response is the true response. The off-diagonal elements would be the probability of misclassification. As described in equation 3.7, this probability would be equal for all off-diagonal elements of this matrix. However, this can certainly be modified such that elements closer to the diagonal are higher than elements farther out.

Additionally, the probabilities defined in equations 3.6 and 3.7 can be informative priors for the misclassification rates. Misclassificaiton rates are specified using a Dirichlet prior (Naranjo et al., 2019). The relative informativeness of the probabilities defined above can be controlled using a tuning parameter, ξ. Based on informal tests on estimating these models, setting the tuning parameter to 10 or more often results in a relatively stable posterior for the misclassification rates. Other approaches to defining an appropriate magnitude for the tuning parameter can also be used. For example, if a researcher can collect data on a subsample of respondents twice, the tuning parameter can be calibrated. The calibration across repeated sampling could

be conducted by estimating the strength of the relationship between the response time at time 1 to whether the respondents' response changed between timepoints.

Additionally, the information gained by using response time to calibrate misclassification has the benefit of being readily available by researchers using digital assessment. Another benefit of the approach described here is obtaining more accurate parameter estimates and more reliable estimates of uncertainty of the trait measurement model and person parameters. I hypothesize that by controlling for measurement error in the observed response using the response time data, we can obtain more reliable measures of the trait η_p. The use of more reliable trait measures will result in the proposed approach yielding a more comprehensive measurement model to account for uncertainty in item responses when investigating relationships among latent traits/personality constructs. The push to develop approaches that account for the uncertainty in our inferences is a growing concern (Rigdon et al., 2019; Rigdon et al., 2020). And the approach described here aims to describe the response process to account for this uncertainty more accurately.

Full Misclassification Item Factor Analysis with Response Time

The method developed here may be more concisely conveyed, as shown in 3.3. The more concise nature makes it a little clearer how the resulting item response is hypothesized to be influenced by the response time and where misclassification plays a role. The diagram also conveys how the observed y_i are responses that inaccurately reflect the respondents' views/opinions/ability/etc. The response without error ν_i comes together with how misclassification function $g_E(.)$ to form the observed responses. The rest of the model is analogous to previous models that merge responses and response time (e.g., Ferrando et al., 2013; Ferrando & Lorenzo-Seva, 2007a, 2007b; Meng et al., 2014; Molenaar, Tuerlinckx, & van der Maas, 2015b; Ranger, 2013)

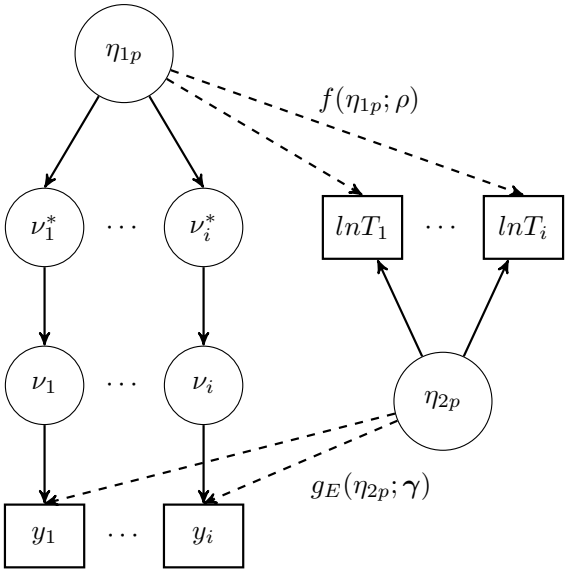


Figure 3.3. Path diagram representation of the proposed joint model of item response, response time, and item misclassification error. *Note.* Residual variances and specific item parameters are omitted from the diagram for ease of discussion.

A path diagram of the full misclassification item factor analysis model with response time is shown in Figure 3.3. Next, the model estimated is described using the model specification diagram previously described for the misclassification in the item factor analysis model. Finally, the full model specification is shown in Figure 3.4.

Prior Justification in Full Model

The priors for the full model are shown in Figure 3.4. The priors for the measurement model portion were the same as the prior models (see *Measurement Model with Misclassification* section for more details). The major changes arise from the inclusion of the model for the log of response time and the hyper-priors for the misclassification rates.

In the response time component, the priors were selected based on the recommendations of Bolsinova and Tijmstra (2018), Merkle and Rosseel (2018), and Molenaar et al. (2021). Specifically, the priors for the response time intercepts and

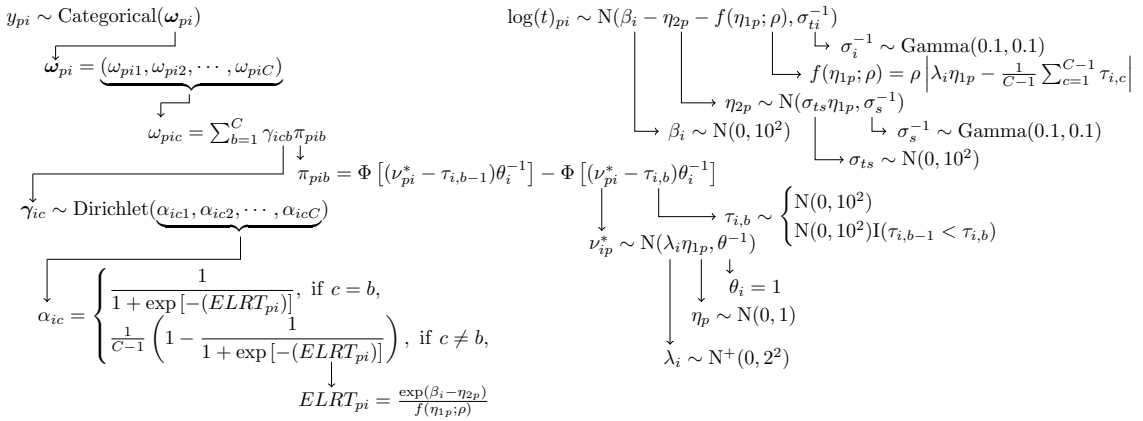


Figure 3.4. Model specification diagram of the proposed joint model of item response, response time, and item misclassification error.

precision have been shown negligibly impact posteriors (Bolsinova & Tijmstra, 2018). Therefore, the choice of the diffuse prior of $\beta_{lrt} \sim \text{N}(0, 10^2)$ for the intercepts and $\sigma_{lrt} \sim \text{Gamma}(0.1, 0.1)$ for the precision are expected to be sufficient without affecting the rest of the model. The prior for the person-item-distance relationship parameter ρ was chosen based on the work of Molenaar et al. (2021). The authors recommended the prior be $\rho \sim \text{N}(0, 10^2)$. The prior for the covariance between the factor σ_{ts} was specified based on the recommendation of Merkle and Rosseel (2018), which specifies $\sigma_{ts} \sim \text{N}(0, 10^2)$.

For the hyper-priors of the misclassification rates, the priors are individually varying based on the effective latent response time (see Equation 3.5). The specification of the link between response time and misclassification is a major contribution of this work. More information regarding the theoretical foundation of this specification is found in section *Calibrating Misclassification Error Rates*.

Research Questions and Hypotheses

The methods described in this chapter provide an approach for modeling item-level misclassification in item factor analysis. The methods need to be tested to discover whether further use can be justified. I will test the hypothesis that modeling

item-level misclassification and response time can provide researchers with a useful inferential tool. The following research questions guide the investigation:

(1) Does modeling item-level misclassification change the inference about scale reliability?

(2) What is the parameter bias and coverage rate of credible intervals in the proposed joint model of item response, response times, and misclassification errors?

(3) Does the use of response time as information of item-level misclassification change inferences about the measurement models for
   (a) NAEP math identity data, or
   (b) extroversion data?

For (1), a simulated dataset will be generated and reduced models estimated (more on this in the methods chapter). The goal is to give a high-level description of how key model summary statistics (i.e., estimates of reliability) can differ among estimated models. For example, the reliability estimates are expected to be noticeably lower when item-level misclassification errors are not modeled.

For (2), bias and coverage are expected to be adequate. No evidence could be found to suggest that the full model cannot be recovered well when the model is correctly and fully specified. However, the number of response categories may influence this expectation because the model requires estimating a misclassification matrix (number of categories by number of categories) for each item and respondent. As the number of categories increases, the sizes quickly expand, resulting in many potentially "unidentified" parameters in the posteriors. The proposed models aim to overcome this potential issue by using response time to inform the misclassification priors. However, this design factor could still be an important aspect of recovering the model parameters adequately.

For (3), applying the proposed model on two separate data sources allows for differences in data peculiarities to test the boundaries of the appropriateness of the proposed methods. Applying the proposed framework to the NAEP math identify data provides a novel application of NAEP's process data of student questionnaire responses. This application shows how item response time can be used with a national dataset to help provide inferences about students. The extroversion data come from the open-source R package *diffIRT* (Molenaar, Tuerlinckx, & van der Maas, 2015a), and all items are dichotomously scored. Applying the proposed modeling framework to this *simple* scenario allows for comparison to existing methods in an open way for others to replicate. Similar to the first simulation study, the results are expected to show that the full model will provide the highest reliability estimates. The reduced models are expected to show that modeling misclassification can provide a useful approach to gaining more information from one's data.

CHAPTER FOUR

Methods

The methods to help address the three research questions are addressed in the following four sections. First, I describe my proposed approach to an applied literature review. Secondly, methods are described for simulating a small dataset with generated with item-level misclassification based on the full model. Using this simulated dataset, I show how inference about the scale reliability may change if misclassification is ignored. Third, I describe the Monte Carlo simulation study to investigate the estimation performance under realistic conditions. Next, I describe the methods and data used in an example using data from the National Assessment of Educational Progress (NAEP) on student mathematics identity/motivation. Lastly, I describe the methods and data used in an example analysis of an extroversion scale with openly available data.

Applied Literature Review

The proposed focused literature review of applied work provides evidence for simulation conditions.

A small subset of education and psychology journals were used in this review where researchers commonly use latent variable models to help answer their research questions. The journals selected for this review are: *Assessment*, *Journal of Psychoeducational Assessment*, *Psychological Assessment*, and *Individual Differences*. The applied literature review was kept focused to allow for an in-depth search of targeted journals. The search was be limited to a five-year range from January 2016 to December 2020. Because of the limited scope of the review, all articles were reviewed that were published within this range that match the key phrase of "("item factor anal-

Table 4.1

Example of information extracted from literature review

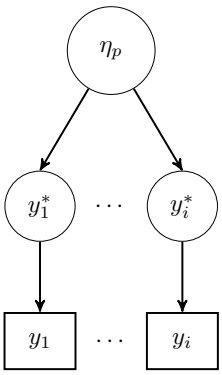
Study	Sample Size	Number Obs. Var.	Indicators per F.	Indicator Dist.	Avg. Std. Loading
Cress, Lambert, & Epstein (2016)	909	42	7-13	4-pt	.74
O'Conner & Fitzgerald (2020)	215	28	13-15	5-pt	.66

Note. Var. = Variables; Obs. = Observed; Avg. = Average; Std. = Standardized. Other information to be extracted include, but is not limited to, journal, number of latent variables, minimum factor loading, maximum factor loading, reliability estimate, how reliability was computed, etc.

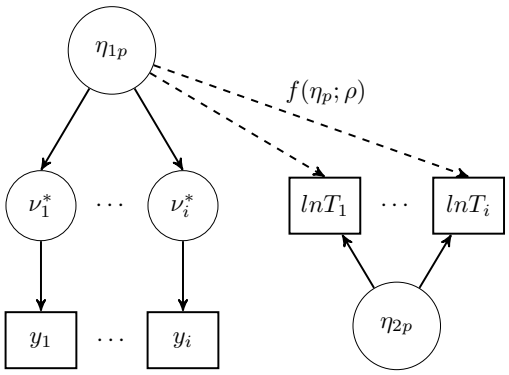
ysis" OR "confirmatory factor analysis" OR "item response theory" OR "structural equation modeling" OR "covariance structure" OR "Rasch" OR "factor structure" OR "psychometric properties)."

The review focused on extracting the data characteristics. Specifically, the extracted characteristics were the journal and author information, sample size, the number of latent variables, the minimum factor loading magnitude, maximum factor loading magnitude, an average of reported factor loadings, reliability estimate(s), and how reliability was computed. In addition, characteristics of the observed indicators include the number of indicators per latent factor, the scale of the indicators (number of response options or continuous), and the distribution of the indicators. Table 4.1 provides an example of the type of information extracted.

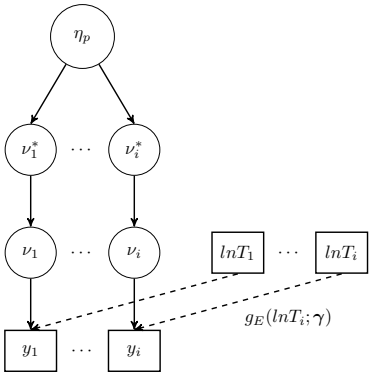
Once the data were extracted, the data helped provide additional validity evidence for the conditions chosen for the simulation studies. First, the extracted information were collated by providing the mean, standard deviation, minimum, median, and maximum values observed. Next, the distribution for the extracted data were plotted to show how the chosen data characteristics for both simulation studies fall within the range of many applied studies. Finally, the chosen simulation and applied



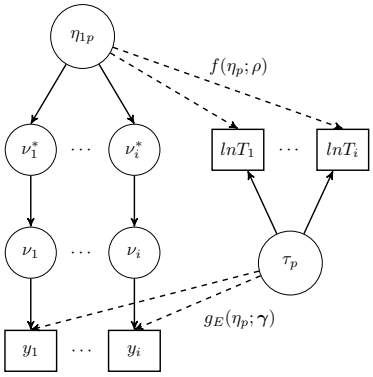
(a) Model 1: Item response only



(b) Model 2: Joint item response and response time



(c) Model 3: Item response only with misclassification



(d) Model 4: Joint model with misclassification

Figure 4.1. Path diagram representation of proposed joint model of item response, response time, and item misclassification error. *Note.* Residual variances and specific item parameters are omitted from the diagram for ease of discussion.

conditions were shown by overlapping the observed distribution and the chosen points in a density plot.

Simulation Study 1: Model Results under Simulated Data

A focused simulation illustrates the use of the proposed joint measurement model for item response and response times with misclassification. The purpose of this study was to demonstrate how item parameter estimates differ when the misclassification is purposefully ignored. A series of four models were estimated to compare results. The models estimated are shown in Figure 4.1.

The results from estimating the four models shown in Figure 4.1 are presented separately. The individual results will demonstrate how the major model parameters differ from true values. Then, the estimates of reliability were compared across models. Reliability was estimated using McDonald's ω (McDonald, 1999, p. 89), which is estimated using

$$\omega = \frac{\left(\sum_{i=1}^{N_i} \lambda_i\right)^2}{\left(\sum_{i=1}^{N_i} \lambda_i\right)^2 + \sum_{i=1}^{N_i} \theta_i^2}, \tag{4.1}$$

where, N_i = number of items, λ_i = factor loading for item i, and θ_i^2 = residual variance of latent response variable i. Because all models were estimated within a Bayesian framework, ω was computed for each posterior draw resulting in a distribution of ω for each model. The resulting distributions were compared graphically and empirically using a simple one-way ANOVA model. The primary outcome of the ANOVA was the effect size estimates η^2 (Maxwell & Delaney, 2004, p. 295-296).

Along with the reliability estimates, the item parameter estimates for the measurement model were compared among models. The parameters to be compared are the factor loadings and item location parameters (e.g., an average of item category thresholds to create a single item location parameter). The estimates were visually compared in a series of scatterplots to illustrate how the estimates differ across the models. Similar to the reliability estimates, the differences among the parameter estimates were evaluated using ANOVA. A two-way ANOVA model was used here where item and model are the two factors. The primary outcome from this analysis will be estimates η^2 effect size for the effect of the model on parameter estimates after partially out the effect of item differences.

Simulation Conditions

Only a single dataset was used in this simulation study. Data characteristics for this simulation study are as follows. The sample size was set to 500. The number of indicators was set to five. The number of response categories was set

to three. The parameter values used to simulate the data are $\lambda = 0.7$, $\sigma_i = 1$, $\mathbb{V}[\eta_1] = 1$, McDonald's-$\omega = 0.83$, $\beta = 1.5$, $\sigma_{lrt} = 0.25$, $\sigma_s = 0.1$, $\rho = 0.1$, and $\tau = \begin{bmatrix} -0.82 & -0.75 & -0.62 & -0.39 & -0.78 \\ 0.78 & 0.88 & 0.83 & 1.03 & 0.88 \end{bmatrix}'$. The item threshold parameters were randomly generated using the code shown in Appendix B. The parameter values are included in each table for each of reference when evaluating the results from each model.

Simulation Study 2: Parameter Recovery

The purpose of this study is to investigate whether the inclusion of misclassification informed by response time in item factor analysis can be used without the need to estimate the complex latent structure of response time. he full model will be estimated under a narrow set of conditions to check if parameter recovery is possible with the full model.

Simulation Conditions

Data characteristics that varied across simulation conditions are the sample size (500, 2500), the number of items (5, 10, 20), and the number of response categories (2, 5) for a total of 18 cells in the design. The exact conditions proposed are subject to change based on the results of the applied literature review. However, the conditions are roughly based on the two applied datasets utilized in this study and conditions used in methodological literature. The parameter values for the measurement and response time models are the same as in simulation study 1.

The number of replications per cell is set to 100. The full model is estimated to take approximately 2 hours (tested on my laptop) but may take longer for some replications, especially for the conditions with a large sample size. To help speed up the simulation, I am looking into how to use the Kodiak server.

The convergence of the posteriors was checked using the criteria of $\hat{R}$ within 0.1 of 1 (Gelman et al., 2013, p. 288). One replication from each condition was extracted for a more in-depth posterior convergence check. This step was conducted before running the full simulation to ensure adequate posterior convergence is likely achieved in most conditions with the selected number of posterior samples, burn-in, and chains.

Sample size. Two sample size conditions were tested. The conditions initially proposed are 500 and 2500. The smaller sample size of 500 approximates a sample size commonly used in primary data collection by individual researchers. For example, the extroversion dataset contains only 143 represents (Molenaar, Tuerlinckx, & van der Maas, 2015a). For example, in a dataset I collected with colleagues for developing the Perceptions of Online Learning Scale (Padgett et al., 2022), the final sample contained approximately 650 respondents. A literature review by DiStefano and Hess (2005) of construct validation papers found that researchers commonly can gain access to approximately 375 (median of 101 studies). Therefore, the lower bound of 500 was selected to represent a typical sample size that is aimed for in construct validation studies.

The larger sample size of 2500 represents the number of cases researchers may use from a large-scale national survey (such as NAEP). The full NAEP data was well over 100,000 respondents, but taking a small random sample (5-10%) is not uncommon when testing the fit more complex latent variable models. The use of a subsample helps reduce the computational burden. Additionally, using the larger sample size allows testing the proposed model under many respondents available in larger studies.

Number of items. The number of items varies across survey scales in practice. The number of items necessarily impacts the estimates of scale reliability as the more

items used leads to a greater amount of information used to triangular the location of each person on the trait of interest. Three conditions were proposed containing 5, 10, and 20 items, respectively. The NAEP math identity scale contains five items, and DiStefano and Hess (2005) found that researchers typically utilized between 3-6 items per unidimensional construct. However, DiStefano and Hess (2005) reported that some researchers have reported using up to 21 indicators for a single construct. The extroversion data used in this dissertation contains ten items. Therefore, the three conditions were selected to capture the realistic conditions of applied researchers.

Number of response categories. The number of response categories determines the size of the misclassification matrix that needs to be estimated. For dichotomous items, the misclassification matrix contains four elements. For items with five response options, the misclassification matrix contains 25 elements. In the full model, the matrices are specified at an individual level, so the number of parameters increases rapidly as the number of categories and respondents increases.

The conditions of two and five response options are representative of the two applied settings. These conditions also generalize to the conditions used by other researchers using attitudinal assessments. The number of conditions for the number of response categories was limited to five. Researchers commonly assume continuity with greater than five response categories (Rhemtulla et al., 2012).

Two additional conditions were added to expand the investigation of the effect of the number of response categories. First, a three-category and seven-category model were added to investigate the effects of this condition more finely. This addition will be described in detail if the results depart significantly from the use of two or five response categories. Otherwise, the results will be included in the appendices and references briefly. The seven-category model took a long time to estimate (almost a week per model), so only 50 replications of these cells were attempted.

Simulation Outcomes

The outcomes from this simulation study were the average relative bias, the coverage rates of 95% credible intervals, and a summary of interval widths. I plan to use the middle 95%-tile of posterior samples for the credible intervals.

The relative bias of parameter estimates will give a high-level summary of the estimation performance. Relative bias was computed across replications within conditions using the posterior median as the point estimate of each parameter. Let θ be the population parameter value, and the replication estimates are denoted by $\hat{\theta}_r$. The condition summary statistics to be reported are the average relative bias, in percent, of the parameter estimates is

$$RB(\theta) = \frac{1}{N_r} \sum_{r=1}^{N_r} \frac{\hat{\theta}_r - \theta}{\theta} \times 100. \tag{4.2}$$

We evaluated the extent of RB as negligible for RB $< |5\%|$, as mild for $| 5\% | \leq$ RB $<| 10\% |$, and unacceptable for RB $>| 10\% |$ (Hoogland & Boomsma, 1998; Muthén & Kaplan, 1985).

The coverage rates are an indicator of the repeated sampling performance of the model under known conditions. For example, a 95% credible interval would be expected to contain the population parameter values 95% of repeated samples under identical conditions. If estimation conditions are appropriate for the proposed model, the coverage rates should be 95% or negligibly different. The coverage rates of the credible intervals were calculated using an indicator function of the posterior summaries across replications. An indicator is created to compute whether the 95% credible interval contains the population parameter values used to simulate the data. The reported coverage rates are the proportion of replications that contain the population parameter values.

The summary of interval widths, especially average interval width, is closely related to the coverage rate as the width of the intervals indicates how precisely the posterior distributions are approximated. If the interval widths vary widely across

replications, then there is evidence that the estimation procedures may be inadequate for the proposed model. Therefore, identifying whether the posteriors are consistent across replications is essential information about the utility of the proposed methods. We may mitigate such issues by altering the parameterization or prior structure if the intervals vary substantially across replications.

Applied Study 1: NAEP Math Identity and Process Data

The National Assessment of Educational Progress (NAEP) is a congressionally mandated assessment of our nation's students across mathematics, reading, science, writing, among other subjects for grades 4, 8, and 12. NAEP is organized by the National Center for Education Statistics (NCES). NAEP transitioned to a digital-based assessment in 2017. The NAEP digital assessment includes a log of all actions taken by students while completing the assessment. The log data are also known as "process data." One of the most commonly used forms of process data is response time. Response time has a long history in educational assessment for cognitive traits (e.g., math ability). However, one of the unique features of the NAEP data is that students also complete a set of background questionnaire items ranging from questions about their family environment to questions about their affective state. Process data are also collected for the non-cognitive component of NAEP.

The NAEP non-cognitive items related to student math identity are used in this study. More broadly, math identity and math motivation are highly related to performance on mathematics assessment (Marsh et al., 1988), including the NAEP math assessment (Zhang et al., 2021). After accounting for all relevant covariates, the differences among students on the non-cognitive trait of motivation/identity accounted for approximately 9% of the variability in NAEP mathematics scores (Zhang et al., 2021). The significant impact of identity on performance determines whether

such claims have sufficient evidence. Probing whether methods can be developed to help shed light on this issue may prove fruitful.

Accounting for the measurement error is the purpose of this application. Additionally, I evaluate how the reliability of the mathematics identity changes depending on how measurement error is incorporated into the model. Measurement error is incorporated using the misclassification approach described in the previous chapter. Misclassification rates are informed by the NAEP process data (response time).

The full model described in the section *Full Misclassification and Measurement Model* is expected to yield the largest increase in estimates of scale reliability. For comparison, the series of reduced models are estimated. The four models shown in Figure 4.1 will be estimated. Model (4) is expected to yield the highest estimated of scale reliability on average. However, an interest also lies in whether the results from Model (3) substantially differ. Suppose we can identify that the response time can be used directly without loss of inferential gains that will tremendously increase the applicability of the proposed methods. The applicability will increase between far fewer parameters that need to be estimated, and results can be obtained quicker.

TThe four models describe the posterior distributions for reliability, estimated using McDonald's ω, under the four models are described. One of the major aims of this application is to demonstrate how inference about scale reliability may depend heavily on how the model incorporates misclassification.

For the NAEP process data example, models were estimated within a Bayesian framework using JAGS (Plummer et al., 2003). Estimation was conducted using the R2jags package (Su & Yajima, 2020) in R. Models were initially estimated using 10000 iterations across four chains. The first 5000 samples from each chain were discarded as burn-in and chains were thinned by a factor of five. The resulting posterior distributions for model parameters were evaluated for convergence using the posterior predictive distribution of item category response proportions, potential scale

reduction factor ($\hat{R}$) being less than 1.1 (Gelman et al., 2013, p. 288), normality/smoothness of posterior density, negligible autocorrelation among samples, mixing of samples between samples, and convergence of Gelman-Rubin-Brook criteria towards 1. A summary of the main model parameters for each model will be available in Appendix D.

Applied Study 2: Extroversion Data

The applied example comes from an open-source personality assessment dataset published as part of the *diffIRT* package (Molenaar, Tuerlinckx, & van der Maas, 2015a) for R. The data contain ten dichotomously scored items measuring aspects of extroversion. The ten items were statements about habits, and respondents were asked whether the habit applied to their personality (i.e., yes/no). The response time, in seconds, was simultaneously recorded.

For items measured dichotomously, the proposed joint model for item response and response time may be seen as a special case of a four-parameter IRT model. The major difference is between a traditional four-parameter IRT model. The proposed model is that the lower and upper asymptotes of the logistic/normal ogive function are informed by the response time weighted by person-item distance

Similar to the NAEP data application, the four models shown in Figure 4.1 will be estimated to these data. Among the estimated model, the posterior distributions of item parameters and ω will be compared. The posterior distributions will initially be compared graphically. Suppose a noticeable difference among models is not achieved. In that case, the posteriors will be compared statistically using a one-way ANOVA model to test whether there is a significant difference among the distributions.

Posterior Sensitivity Analysis

The proposed model is complex and has a necessarily complex corresponding prior structure. An outstanding question from the previous studies is how influential

the prior structure is on these results. A prior-to-posterior sensitivity analysis is conducted using the Extroversion data. The use of the Extroversion dataset allows for an open and rigorous exploration of the sensitivity of these results that others can run, modify, or poke holes in.

The model for the Extroversion dataset is the simplest of the models examined in this project. The model is the simplest because all ten items were dichotomously scored. The measurement and response time model components have 33 parameters in total (10 factor loadings, 10 item thresholds, 10 response time intercepts, 1 person-item distance parameter, 1 speed factor variance, 1 factor covariance). Additionally, there are many more person-specific parameters in the model. The person-specific parameters are the values for each latent variable (2 latent factors, 10 latent response variables, 10 item misclassification matrices $[2 \times 2]$). Therefore, a total of 7,417 parameters are variable in the estimation of the full model for this Extroversion dataset.

Aside from the posterior dimension, the major inferential goals are to understand how reliably we can measure Extroversion. As reliability is a function of the factor loadings, the prior specified for the factoring loadings can significantly influence the induced prior on reliability. Additionally, item-level misclassification is hypothesized. Item-level misclassification rates are informed through individual-varying priors generated as a function of the effective latent response time (ELRT). The informativeness of priors for the misclassification rates approximated by the ELRT can be tuned. The tuning parameter on the prior for the misclassification rates was initially set to one ($\xi = 1$) in all analyses. When $\xi = 1$, the priors for misclassification rates were set to the direct transformation of ELRT, as shown in Figure 3.4; however, the tuning parameter can be altered to change how informative ELRT is as a prior. The sensitivity of the model results to changes in the magnitude of the tuning parameter are testable.

Table 4.2

Posterior sensitivity analyses prior selection

Prior	λ	ξ^{a}
Base[b]	$N^{+}(0, 0.44)$	1
A	$N^{+}(0, 0.01)$	0.1
B	$N^{+}(0, 1)$	10
C	$N^{+}(0, 10)$	$Uniform(0.5, 1.5)$
D	$N(0, 0.44)$	$Gamma(1, 1)$
E	$N(0, 0.01)$	
F	$N(0, 1)$	
G	$N(0, 10)$	

Note. The normal distribution priors are precision-parameterized; therefore, a uncertainty parameter of $0.44 \equiv 1.5$ standard deviation. [a]ξ is a parameter controlling the certainty of the individual varying misclassification parameters. [b]Base prior was used in the Monte Carlo simulation study.

Testing the sensitivity of the posterior of reliability (ω) to decisions of priors for factor loadings and the misclassification tuning parameter is done by re-estimating the full model with different choices for these parameters. The eight prior specifications for the factor loadings and the five specifications for the tuning parameter are shown in Table 4.2. The priors for the factor loading are chosen to have varying degrees of informativeness. The truncated nature of the first four priors aligns with the recommended (Levy & Mislevy, 2016) approach for resolving the indeterminacy of orientation of the latent response variables. The last four priors for the factor loadings remove this restriction. Removing the positive restriction on the factor loadings aligns with priors selected for Bayesian linear factor analysis using continuous indicators (Merkle & Rosseel, 2018). The priors for the tuning parameter were similarly varied to change the relative informativeness of the prior. The last two priors for the tuning parameter were chosen to center the value around one but allow for variabil-

ity across iterations. A graphical summary of the posteriors with the relevant priors layered on the plot is used to help identify posterior sensitivity to the prior structure.

CHAPTER FIVE

Results

Applied Literature Review

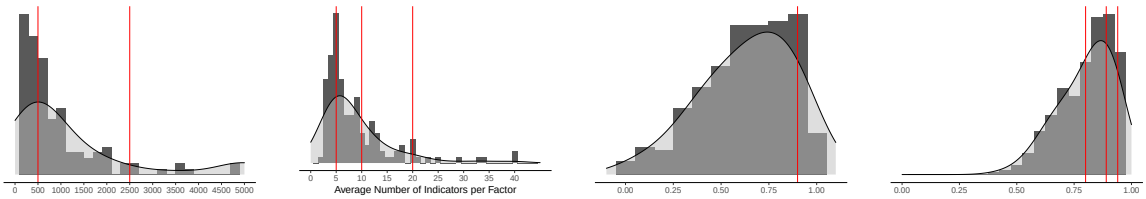
The applied literature review provided a valuable source of information that provides evidence for the measurement models of interest in this research. The complete list of extracted information is compiled and freely available on the accompanying supplemental website. In this section, I describe the information extracted, which is summarized in Table 5.1.

Table 5.1

Information extracted from studies reviewed

Characteristic	Average	SD	Min	Median	Max
Sample Size	4575.1	24451.4	36	603	263683
Number of latent variables	5.7	7.2	1.0	4.0	60.0
Number of observed variables[a]	52.42	83.5	5	24	442
Number of indicators per factor					
Min	7.1	6.6	1	5	40
Avg	9.3	7.2	2.7	6.7	40
Max	13.5	15.8	3	9	135[b]
Factor Loadings (Standardized)					
Min	0.41	0.20	0.00	0.41	0.86
Avg	0.64	0.13	0.23	0.65	0.93
Max	0.85	0.10	0.49	0.87	1.08
Reliability					
Min	.73	.13	.40	.71	.95
Avg	.81	.09	.62	.83	.95
Max	.88	.08	.67	.90	.98
Proportion using McDonald's ω[c]	.16				

Note. [a] Number of variables reportedly collected and analyzed in a measurement model. Many of the reported studies used the collected items to form sum-scores effectively reducing the number of "observed variables" to a small number of construct-specific scores to use in a CFA of those sum-scores. [b] The paper reporting the use of a latent variable with 135 indicators used the Coding subtest of Wechsler Adult Intelligence Scale (WAIS-IV). [c] Proportion using McDonald's ω is conditional on whether reliability was reported.



(a) Sample Size *(b)* Indicators *(c)* Factor Loadings *(d)* Reliability

Figure 5.1. Distribution of measurement model characteristics and chosen conditions for simulations. *Note.* Vertical lines in each plot represent the chosen value(s) as conditions in the simulation.

The data conditions described in Table 5.1 focus on the measurement model of the latent trait. The measurement models for simulation studies 1 and 2 were restricted to a single construct to limit model complexity. However, applied studies commonly use four (median) to six (average of 5.7) latent variables in their studies. As a result, the sample size was highly variable, as shown in Figure 5.1(a). In addition, the number of indicators per factor varied greatly from 1-135 in the studies examined. The distribution of the number of indicators per factor is shown in Figure 5.1(b), demonstrating how the conditions simulated fall within the range many researchers encounter when using latent variable models. Similar conclusions about the factor loadings and reported reliability estimates could be drawn from Figure 5.1(c) and 5.1(d), respectively.

Additionally, the scale of the indicators was also extracted for each study. Twenty-two percent of reviewed studies reported using more than one survey where the observed indicators had a different response scale (i.e., dichotomous and five response options). Studies reporting collecting data using dichotomous indicators (15%), three-(6%), four-(28%), five-(47%), six-(5%), and seven-point (17%) Likert-type response scales. Researchers reported using the total-scores in 10% of studies instead of the item-level responses as part of their unit of analyses. The frequency of use of the response scales observed in applied studies shows that focusing the simulation study

on dichotomous and five-point response scales aligns with over half of the applied studies reviewed.

Simulation Study 1: Simulated Data Analysis

Model 1 Results

Model 1 is similar to traditional factor analyses with categorical indicators. The posterior distributions for the measurement model are summarized in Table 5.2. The estimates of reliability are described in more detail in the *Comparing Posterior Distribution of Reliability* section. The posterior distributions for these parameters converged well in most conditions. The $\hat{R}$ values for all relevant parameters were below 1.10 (see Table 5.2). A more in-depth description of the posterior distributions and convergence criteria (e.g., traceplots, Gelman-Rubin-Brooks convergence criteria plots, and autocorrelation) are reported in Appendix B.

The posterior distributions reveal that the factor loadings were underestimated in this analysis. The underestimation is possibly due to the additional uncertainty with respect to the misclassification not being accounted for in the analysis. The item threshold parameters (τs) were estimated relatively close to the simulated values. Therefore, modeling misclassification likely does not influence the individual item characteristics but may influence how strongly we expect the items to relate to each other, as evident by the underestimated factor loadings.

Model 2 Results

Model 2 is a joint model of item responses and response time as Molenaar, Tuerlinckx, and van der Maas (2015b) recommended. Models that jointly estimate the item responses and response time are expected to yield higher reliability estimates as more information is used to estimate the model. The posterior distributions for the measurement model and response time model are summarized in Table 5.3. The

Table 5.2

Posterior distributions of model 1 summary

Parameter	TV	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$
λ_1	0.70	0.48	0.14	0.21	0.39	0.48	0.57	0.74	1.01
λ_2	0.70	0.38	0.13	0.12	0.29	0.38	0.46	0.63	1.00
λ_3	0.70	0.35	0.13	0.10	0.27	0.35	0.43	0.61	1.02
λ_4	0.70	0.47	0.13	0.20	0.38	0.47	0.55	0.71	1.01
λ_5	0.70	0.35	0.12	0.11	0.27	0.35	0.44	0.59	1.01
$\tau_{1,1}$	-0.82	−0.77	0.10	−0.98	−0.83	−0.77	−0.70	−0.58	1.01
$\tau_{2,1}$	−0.75	−0.73	0.09	−0.92	−0.80	−0.73	−0.67	−0.56	1.00
$\tau_{3,1}$	−0.62	−0.62	0.09	−0.80	−0.68	−0.61	−0.56	−0.45	1.00
$\tau_{4,1}$	−0.39	−0.54	0.09	−0.73	−0.60	−0.54	−0.48	−0.36	1.01
$\tau_{5,1}$	−0.87	−0.81	0.09	−0.98	−0.87	−0.81	−0.75	−0.63	1.00
$\tau_{1,2}$	0.78	0.86	0.11	0.67	0.79	0.85	0.93	1.09	1.01
$\tau_{2,2}$	0.88	0.88	0.10	0.70	0.82	0.88	0.95	1.08	1.00
$\tau_{3,2}$	0.83	0.88	0.09	0.71	0.82	0.88	0.94	1.06	1.00
$\tau_{4,2}$	1.03	1.01	0.11	0.81	0.94	1.01	1.07	1.24	1.00
$\tau_{5,2}$	0.88	0.83	0.09	0.66	0.77	0.83	0.89	1.01	1.00
ω	0.83	0.52	0.06	0.38	0.48	0.52	0.56	0.62	1.01

Note. Reported factor loadings are standardized. TV = True Value.

posterior distributions for these parameters converged pretty well. The $\hat{R}$ values for all relevant parameters were below 1.10 (see Table 5.3). A more in-depth description of the posterior distributions and convergence criteria (e.g., traceplots, Gelman-Rubin-Brooks convergence criteria plots, and autocorrelation) are reported in Appendix B.

Similar to the results of Model 1, the estimated posterior distributions for factor loadings were below the simulated values. The posterior distributions reveal that the factor loadings were underestimated and similar to Model 1. The item threshold parameters (τs) were also similarly estimated relatively close to the simulated values.

For the response time portion of the model, the posterior distributions were relatively close to the simulated values except for ρ. The person-item distance relationship contributes to the response time for each item. In this case, the estimates of ρ were above what we would expect given the simulated value. When ρ is over-estimated, we may incorrectly conclude that the person-item distance has more impact on response

Table 5.3

Posterior distributions of model 2 summary

Parameter	TV	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$
λ_1	0.70	0.40	0.09	0.21	0.34	0.40	0.45	0.58	1.01
λ_2	0.70	0.30	0.09	0.09	0.24	0.30	0.36	0.47	1.04
λ_3	0.70	0.51	0.10	0.30	0.45	0.51	0.58	0.69	1.01
λ_4	0.70	0.40	0.09	0.23	0.34	0.40	0.46	0.58	1.01
λ_5	0.70	0.48	0.09	0.30	0.43	0.49	0.55	0.64	1.01
$\tau_{1,1}$	−0.82	−0.75	0.09	−0.93	−0.81	−0.75	−0.69	−0.58	1.00
$\tau_{2,1}$	−0.75	−0.73	0.08	−0.90	−0.79	−0.73	−0.68	−0.57	1.00
$\tau_{3,1}$	−0.62	−0.67	0.09	−0.86	−0.74	−0.67	−0.61	−0.50	1.01
$\tau_{4,1}$	−0.39	−0.50	0.08	−0.66	−0.55	−0.50	−0.44	−0.33	1.00
$\tau_{5,1}$	−0.87	−0.83	0.09	−1.00	−0.89	−0.83	−0.76	−0.66	1.00
$\tau_{1,2}$	0.78	0.82	0.09	0.65	0.76	0.82	0.88	1.00	1.00
$\tau_{2,2}$	0.88	0.85	0.08	0.69	0.80	0.85	0.91	1.02	1.00
$\tau_{3,2}$	0.83	0.88	0.09	0.70	0.82	0.88	0.95	1.06	1.00
$\tau_{4,2}$	1.03	1.02	0.09	0.84	0.95	1.01	1.08	1.20	1.00
$\tau_{5,2}$	0.88	0.87	0.09	0.69	0.81	0.87	0.93	1.06	1.00
β_{lrt1}	1.50	1.54	0.06	1.43	1.50	1.54	1.58	1.67	1.01
β_{lrt2}	1.50	1.55	0.06	1.44	1.51	1.55	1.59	1.69	1.03
β_{lrt3}	1.50	1.65	0.08	1.49	1.60	1.66	1.71	1.81	1.01
β_{lrt4}	1.50	1.57	0.06	1.46	1.53	1.57	1.61	1.70	1.01
β_{lrt5}	1.50	1.57	0.08	1.42	1.52	1.57	1.62	1.72	1.02
σ_{lrt1}	4.00	3.98	0.30	3.43	3.78	3.97	4.18	4.59	1.00
σ_{lrt2}	4.00	3.96	0.29	3.42	3.75	3.94	4.14	4.57	1.00
σ_{lrt3}	4.00	4.17	0.35	3.55	3.93	4.15	4.39	4.93	1.00
σ_{lrt4}	4.00	4.12	0.31	3.56	3.91	4.11	4.32	4.76	1.00
σ_{lrt5}	4.00	4.71	0.39	4.01	4.43	4.69	4.95	5.54	1.00
σ_s	10.0	10.39	1.47	8.06	9.36	10.21	11.20	13.80	1.02
σ_{ts}	0.07	0.08	0.03	0.02	0.06	0.08	0.10	0.13	1.00
ρ	0.10	0.47	0.14	0.20	0.37	0.46	0.55	0.77	1.04
ω	0.83	0.53	0.06	0.40	0.49	0.53	0.57	0.62	1.01

Note. Reported factor loadings are standardized. TV = True Value.

time than truly exists. In this case, this would mean that a larger distance implies lower response time (positive ρ implies lower response time while negative ρ implies higher response time).

Model 3 Results

Model 3 is the first model to incorporate item-level misclassification into the analysis. The model is similar to traditional factor analyses with categorical indicators but adds the misclassification component. The misclassification rates are directly informed by the observed response time. The posterior distributions for the measurement model are summarized in Table 5.4. The posterior distributions for these parameters converged pretty well. The $\hat{R}$ values for all relevant parameters were below 1.10 (see Table 5.4). A more in-depth description of the posterior distributions and convergence criteria (e.g., traceplots, Gelman-Rubin-Brooks convergence criteria plots, and autocorrelation) are reported in Appendix B.

The resulting posteriors for the factor loadings were closer to the simulated values than the two previous models. The posterior distributions are also, however, more diffuse. The greater uncertainty in the posterior distributions for the measure-

Table 5.4

Posterior distributions of model 3 summary

Parameter	TV	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$
λ_1	0.70	0.62	0.16	0.27	0.52	0.64	0.73	0.89	1.07
λ_2	0.70	0.61	0.16	0.28	0.50	0.62	0.73	0.87	1.01
λ_3	0.70	0.40	0.16	0.08	0.30	0.41	0.51	0.69	1.04
λ_4	0.70	0.63	0.15	0.31	0.53	0.64	0.74	0.87	1.04
λ_5	0.70	0.38	0.17	0.06	0.25	0.38	0.50	0.69	1.01
$\tau_{1,1}$	−0.82	−0.89	0.21	−1.41	−0.99	−0.86	−0.75	−0.58	1.13
$\tau_{2,1}$	−0.75	−0.89	0.18	−1.30	−0.99	−0.87	−0.76	−0.59	1.01
$\tau_{3,1}$	−0.62	−0.63	0.12	−0.87	−0.71	−0.63	−0.54	−0.40	1.01
$\tau_{4,1}$	−0.39	−0.54	0.15	−0.88	−0.63	−0.53	−0.44	−0.27	1.01
$\tau_{5,1}$	−0.87	−0.87	0.14	−1.16	−0.96	−0.87	−0.78	−0.62	1.00
$\tau_{1,2}$	0.78	1.08	0.22	0.76	0.94	1.06	1.18	1.65	1.09
$\tau_{2,2}$	0.88	1.16	0.21	0.83	1.01	1.13	1.27	1.65	1.01
$\tau_{3,2}$	0.83	1.04	0.14	0.78	0.95	1.04	1.13	1.32	1.00
$\tau_{4,2}$	1.03	1.31	0.24	0.94	1.15	1.28	1.43	1.87	1.06
$\tau_{5,2}$	0.88	0.96	0.13	0.70	0.87	0.96	1.05	1.22	1.00
ω	0.83	0.71	0.06	0.57	0.67	0.72	0.75	0.80	1.01

ment model parameters aligns with the increased source of error on the model (i.e., misclassification of responses) and with more parameters in the model.

Model 4 Results

Model 4 is the full model of interest and incorporates item-level misclassification into the analysis informed by the effective latent response time. The posterior distributions for the measurement model are summarized in Table 5.5. The posterior distributions for these parameters converged pretty well. The $\hat{R}$ values for all relevant parameters were below 1.10 (see Table 5.5). A more in-depth description of the posterior distributions and convergence criteria (e.g., traceplots, Gelman-Rubin-Brooks convergence criteria plots, and autocorrelation) are reported in Appendix B.

The factor loadings were closest to the simulated values out of the four estimated four models. The underestimation is not too severe as all five posterior distributions contain the true values within the 95% credible intervals. Similar to Model 3, the posteriors for the item thresholds were more off than the Models 1 and 2. This could indicate that item location parameter estimates shift towards zero when misclassification is incorporated into the model. Although, how the estimates of item location are shifted is yet to be determined.

Similar to the results of Model 2, the estimates for the response time portion of the model were estimated well. All posteriors except for ρ contained the true values. The estimated posterior for ρ was closer to the true value; however, the parameter was still overestimated.

Comparing Posterior Reliability Distributions

In the ANOVA comparing the posterior distribution of ω across the four models, the effect size estimate (η^2) was .76. The large proportion of variability in omega values across models gives evidence that inferences about reliability can be significantly different depending on how these data are modeled. All pairwise comparisons

Table 5.5

Posterior distributions of model 4 summary

Parameter	TV	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$
λ_1	0.70	0.61	0.11	0.37	0.54	0.61	0.69	0.79	1.04
λ_2	0.70	0.55	0.13	0.26	0.48	0.57	0.64	0.75	1.09
λ_3	0.70	0.66	0.11	0.39	0.60	0.67	0.74	0.82	1.05
λ_4	0.70	0.59	0.10	0.38	0.52	0.59	0.66	0.79	1.02
λ_5	0.70	0.58	0.11	0.34	0.51	0.59	0.65	0.77	1.10
$\tau_{1,1}$	−0.82	−0.88	0.14	−1.18	−0.98	−0.88	−0.78	−0.62	1.02
$\tau_{2,1}$	−0.75	−0.88	0.14	−1.16	−0.97	−0.88	−0.78	−0.61	1.01
$\tau_{3,1}$	−0.62	−0.75	0.15	−1.07	−0.85	−0.74	−0.65	−0.48	1.04
$\tau_{4,1}$	−0.39	−0.50	0.12	−0.75	−0.58	−0.49	−0.41	−0.26	1.00
$\tau_{5,1}$	−0.87	−0.92	0.14	−1.21	−1.01	−0.92	−0.83	−0.66	1.00
$\tau_{1,2}$	0.78	0.98	0.15	0.70	0.88	0.98	1.08	1.29	1.01
$\tau_{2,2}$	0.88	1.03	0.14	0.76	0.93	1.03	1.12	1.31	1.00
$\tau_{3,2}$	0.83	1.08	0.15	0.80	0.98	1.07	1.17	1.39	1.01
$\tau_{4,2}$	1.03	1.24	0.17	0.94	1.13	1.23	1.34	1.59	1.05
$\tau_{5,2}$	0.88	1.03	0.15	0.75	0.93	1.02	1.12	1.32	1.02
β_{lrt1}	1.50	1.56	0.07	1.44	1.52	1.56	1.61	1.70	1.02
β_{lrt2}	1.50	1.60	0.08	1.45	1.54	1.59	1.65	1.77	1.08
β_{lrt3}	1.50	1.65	0.10	1.47	1.58	1.65	1.72	1.84	1.06
β_{lrt4}	1.50	1.59	0.07	1.46	1.54	1.59	1.63	1.72	1.05
β_{lrt5}	1.50	1.53	0.08	1.40	1.48	1.53	1.59	1.70	1.08
σ_{lrt1}	4.00	4.01	0.30	3.45	3.80	4.00	4.21	4.64	1.00
σ_{lrt2}	4.00	4.01	0.30	3.44	3.80	4.00	4.20	4.65	1.00
σ_{lrt3}	4.00	4.17	0.37	3.51	3.91	4.13	4.39	4.95	1.01
σ_{lrt4}	4.00	4.13	0.31	3.56	3.92	4.12	4.33	4.76	1.00
σ_{lrt5}	4.00	4.61	0.37	3.93	4.36	4.59	4.85	5.35	1.00
σ_s	10.0	10.68	1.72	8.08	9.49	10.40	11.60	14.67	1.06
σ_{ts}	0.07	0.09	0.03	0.03	0.07	0.09	0.11	0.15	1.00
ρ	0.10	0.29	0.10	0.11	0.22	0.28	0.36	0.49	1.07
ω	0.83	0.75	0.05	0.63	0.71	0.75	0.78	0.83	1.02

Table 5.6

Summary of posterior distribution of reliability

Model	Mean	SD	2.5%	25%	50%	75%	97.5%
Model 1	0.52	0.06	0.38	0.48	0.52	0.56	0.62
Model 2	0.53	0.06	0.40	0.49	0.53	0.57	0.62
Model 3	0.71	0.06	0.57	0.67	0.72	0.75	0.80
Model 4	0.75	0.05	0.63	0.71	0.75	0.78	0.83

were also significantly different, although the differences among estimates for Models 1 and 2 were negligible, and the differences among estimates for Models 3 and 4 were negligible. These results suggest that Model 3 and Model 4 provide practically equivalent interpretations about the reliability estimates of McDonald's ω for a single factor as measured by these items. In general, incorporating misclassification into the item factor analysis can substantially increase estimates of reliability of the factor of interest by bringing the estimates closer to the underlying truth.

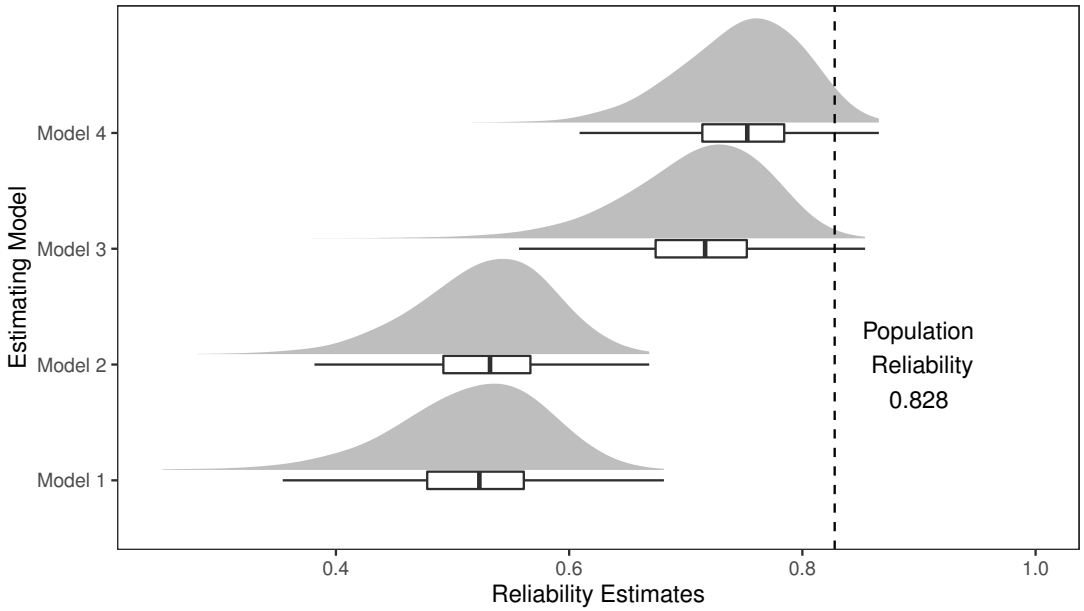


Figure 5.2. Simulation study 1 posterior distribution of reliability across models. *Note.* The posterior distributions presented above are 5000 draws from each posterior, respectively.

Posterior Convergence

The convergence of the posterior distributions was assessed separately for the different parameter groups in the model. For instance, the convergence of the posterior distributions for the factor loadings was aggregated to a single summary statistic. As a result, a posterior was determined to "converge" in this Monte Carlo simulation study by whether the average $\hat{R}$ value for the group of parameters (i.e., average $\hat{R}$ for factor loadings) was below 1.10. Next, the results are described for the dichotomous and polytomous (five-category) results.

The convergence results for models estimated using dichotomous indicators are summarized in Table 5.7. The posteriors converged fairly well for nearly all parameters when the indicators were dichotomous where the average $\hat{R}$ were below 1.10 for most parameters and sample size conditions. Exceptions occurred in conditions where the number of items was five for factor loadings (λ), the total variance of latent response variables (θ), and reliability (ω). Adding more items to the model resulted in higher convergence rates of the posteriors. Convergence rates were above 90% when the number of items was 20 but fell as the number of items decreased depending on the parameter. For instance, the convergence rate for factor loadings fell to 85-90% for models with ten items and 65-68% for models with five items. The convergence of the posterior for reliability fell to 63% in models with five items and a sample size of 500. Still, additional items and a larger sample size increased the convergence rate. The convergence rates tended to be better in conditions with larger sample sizes. A potentially useful aspect of these results is that the variability in $\hat{R}$ across replications became negligible as the sample size and number of items increased, suggesting that convergence becomes more consistent.

Table 5.7

Posterior convergence by $\hat{R}$ of dichotomous items

		Avg $\hat{R}$			SD $\hat{R}$			% Converge		
Parameter	N	5	10	20	5	10	20	5	10	20
λ	500	1.09	1.06	1.04	0.05	0.02	0.01	68	85	95
	2500	1.10	1.05	1.03	0.05	0.01	0.01	65	90	98
τ	500	1.03	1.02	1.01	0.02	0.01	0.00	96	98	100
	2500	1.02	1.01	1.01	0.01	0.00	0.00	98	100	100
θ	500	1.14	1.08	1.05	0.07	0.03	0.01	50	76	92
	2500	1.11	1.05	1.03	0.05	0.01	0.01	62	89	98
β_{lrt}	500	1.07	1.04	1.02	0.04	0.02	0.01	80	93	99
	2500	1.07	1.03	1.02	0.04	0.01	0.01	80	97	100
σ_{lrt}	500	1.01	1.00	1.00	0.01	0.01	0.00	97	100	100
	2500	1.01	1.00	1.00	0.01	0.00	0.00	99	100	100
σ_s	500	1.02	1.01	1.00	0.02	0.01	0.00	99	100	100
	2500	1.02	1.01	1.00	0.02	0.01	0.00	97	100	100
σ_{st}	500	1.04	1.02	1.01	0.04	0.02	0.01	95	100	100
	2500	1.03	1.02	1.01	0.03	0.01	0.01	98	100	100
ρ	500	1.07	1.06	1.04	0.08	0.05	0.02	82	85	97
	2500	1.07	1.03	1.02	0.07	0.03	0.01	80	97	100
ω	500	1.10	1.05	1.03	0.07	0.04	0.03	63	88	96
	2500	1.08	1.03	1.02	0.08	0.02	0.02	78	97	100

Note. Number of items 5, 10, 20 are represented along the columns of this table. Avg $\hat{R}$ = Average $\hat{R}$ across replications; SD $\hat{R}$ = standard deviation of $\hat{R}$ estimates across replications; % Converge = percent of replications with $\hat{R} < 1.10$; λ = factor loading; τ = item threshold; θ = total variance of latent response variable; β_{lrt} = response time model intercept; σ_{lrt} = response time model item residual variance; σ_s = speed latent variable variance; σ_{ts} = covariance between latent variables; ρ = person-item distance relationship; ω = McDonald's ω reliability estimate.

Table 5.8

Posterior convergence by $\hat{R}$ of polytomous (five-category) items

		Avg $\hat{R}$			SD $\hat{R}$			Prop. Converge		
Parameter	N	5	10	20	5	10	20	5	10	20
λ	500	1.07	1.04	1.04	0.03	0.01	0.01	79	93	95
	2500	1.04	1.03	1.03	0.02	0.01	0.01	92	97	97
τ	500	1.02	1.02	1.02	0.01	0.00	0.00	97	100	100
	2500	1.02	1.02	1.02	0.00	0.00	0.00	100	100	100
θ	500	1.08	1.05	1.04	0.04	0.01	0.01	75	92	95
	2500	1.05	1.03	1.03	0.02	0.01	0.01	92	97	97
β_{lrt}	500	1.03	1.02	1.01	0.02	0.02	0.01	93	98	100
	2500	1.03	1.02	1.01	0.02	0.01	0.01	98	100	100
σ_{lrt}	500	1.00	1.00	1.00	0.00	0.00	0.00	100	100	100
	2500	1.00	1.00	1.00	0.00	0.00	0.00	100	100	100
σ_s	500	1.01	1.00	1.00	0.01	0.01	0.00	100	100	100
	2500	1.01	1.00	1.00	0.01	0.00	0.00	100	100	100
σ_{st}	500	1.02	1.01	1.01	0.02	0.01	0.01	99	100	100
	2500	1.02	1.01	1.01	0.01	0.01	0.00	100	100	100
ρ	500	1.04	1.03	1.03	0.04	0.03	0.02	91	95	98
	2500	1.03	1.02	1.02	0.02	0.02	0.02	98	100	100
ω	500	1.05	1.04	1.03	0.04	0.03	0.03	88	97	98
	2500	1.03	1.03	1.02	0.03	0.02	0.02	99	98	99

Note. Number of items 5, 10, 20 are represented along the columns of this table. Avg $\hat{R}$ = Average $\hat{R}$ across replications; SD $\hat{R}$ = standard deviation of $\hat{R}$ estimates across replications; % Converge = percent of replications with $\hat{R} < 1.10$.

The posterior converge results for models estimated with polytomous (five-category) indicators are reported in Table 5.8. The posteriors converged well for all parameters when the indicators had five categories where the average $\hat{R}$ value was below 1.10 for all parameters, number of items, and sample size conditions. The convergence rate of the posterior distributions did fall below 90% for factor loadings (79%), the total variance of latent response (75%), and reliability (88%). The below 90% convergence rates occurred in conditions using five items and a sample size of 500. Similar patterns of convergence were found in models estimated with three categories C.1 and seven categories C.3.

Posterior Median Bias

Relative bias was assessed using the posterior median to estimate the parameter value. This section evaluates how biased the estimates are on average. First, the overall effect of the three design factors (number of categories, number of items, and sample size) was evaluated The effects are summarized for each parameter in Table 5.9 as the η^2 effect size measure. The number of response categories accounted for 1.5% of the variability in relative bias estimates for factor loadings, and 2.5% of the variability in estimates of reliability. Number of items is known to be a major influence on the reliability estimates. Our results captured this finding with sample size accounting for the highest percentage of differences in the relative bias of reliability estimates. Finally, the design factor chosen were found to influence estimates of relative bias for different parameters, and how these factors influenced the results is the remaining focus of this section.

First, models using dichotomous items showed a preponderance of negatively biased estimates for most parameters except the response time model parameters (β_{lrt},

Table 5.9

Effect of design factors on relative bias estimates

	λ	τ	θ	β_{lrt}	σ_{lrt}	σ_s	σ_{ts}	ρ	ω
C	.015	.563	.006	.010	.005	.000	.018	.020	.025
I	.031	.000	.003	.051	.010	.026	.002	.047	.133
N	.001	.006	.035	.065	.006	.001	.001	.116	.049
C:I	.039	.003	.011	.003	.005	.004	.007	.003	.003
C:I	.008	.005	.003	.002	.002	.001	.001	.001	.002
I:N	.014	.000	.002	.001	.003	.003	.002	.002	.008
C:I:N	.012	.001	.004	.005	.001	.004	.004	.012	.010

Note. C = number of categories; I = number of items; N = sample size; λ = factor loading; τ = item threshold; θ = total variance of latent response variable; β_{lrt} = response time model intercept; σ_{lrt} = response time model item residual variance; σ_s = speed latent variable variance; σ_{ts} = covariance between latent variables; ρ = person-item distance relationship; ω = McDonald's ω reliability estimate.

Table 5.10

Posterior bias of dichotomous items

		Average Relative Bias			Average Bias		
Parameter	N	5	10	20	5	10	20
λ	500	−25.57	−13.95	−10.25	−0.23	−0.13	−0.09
	2500	−16.49	−13.87	−13.83	−0.15	−0.12	−0.12
τ	500	−8.28	−2.85	6.40	−0.00	0.01	−0.01
	2500	−2.28	−3.30	−15.94	−0.00	−0.00	−0.00
θ	500	−9.86	−5.68	−4.72	−0.18	−0.10	−0.09
	2500	−11.40	−10.66	−10.84	−0.21	−0.19	−0.20
β_{lrt}	500	0.09	−1.07	−1.31	0.00	−0.02	−0.02
	2500	0.55	0.56	0.30	0.01	0.01	0.00
σ_{lrt}	500	1.29	0.66	0.17	0.05	0.03	0.01
	2500	0.42	0.14	0.08	0.02	0.01	0.00
σ_s	500	11.64	6.25	4.35	1.16	0.63	0.44
	2500	7.68	6.77	5.76	0.77	0.68	0.58
σ_{st}	500	−9.70	−6.54	−4.40	0.01	0.00	0.00
	2500	−7.76	−2.67	−6.57	0.01	0.00	0.00
ρ	500	33.04	−1.20	−7.13	0.03	−0.00	−0.01
	2500	37.40	43.85	36.33	0.04	0.04	0.04
ω	500	−3.94	−1.53	−0.93	−0.03	−0.01	−0.01
	2500	−4.31	−3.34	−1.92	−0.03	−0.03	−0.02

Note. λ = factor loading; τ = item threshold; θ = total variance of latent response variable; β_{lrt} = response time model intercept; σ_{lrt} = response time model item residual variance; σ_s = speed latent variable variance; σ_{ts} = covariance between latent variables; ρ = person-item distance relationship; ω = McDonald's ω reliability estimate.

σ_{lrt}, σ_s). Summaries of the average relative bias and average bias on the scale of the parameter are given in Table 5.10. The measurement model parameters for the factor were more severely negatively biased in conditions with five items and a lower sample size. The negative bias observed for the reliability estimates was negligible on average under all conditions with dichotomous items. The persistent negative aspect of the estimates of relative bias could indicate that the priors chosen for this simulation study were more influential than originally thought. The parameters for the model's response time portion were estimated with negligible bias on average. However, the estimates of the speed factor variance were mildly positively biased indicating the results tended to

Table 5.11

Posterior bias of polytomous (five categories) items

		Average Relative Bias			Average Bias		
Parameter	N	5	10	20	5	10	20
λ	500	−14.66	−13.48	−12.76	−0.13	−0.12	−0.11
	2500	−14.93	−14.61	−14.40	−0.13	−0.13	−0.13
τ	500	−3.81	−6.67	−22.63	0.00	−0.00	0.01
	2500	−8.11	−9.11	−16.54	−0.02	−0.01	−0.01
θ	500	−9.39	−9.83	−9.50	−0.17	−0.18	−0.17
	2500	−11.81	−11.84	−11.76	−0.21	−0.21	−0.21
β_{lrt}	500	0.53	−0.30	−0.82	0.01	−0.00	−0.01
	2500	1.42	0.57	0.39	0.02	0.01	0.01
σ_{lrt}	500	1.00	0.33	0.31	0.04	0.01	0.01
	2500	0.15	0.08	0.13	0.01	0.00	0.01
σ_s	500	10.07	7.25	6.22	1.01	0.73	0.62
	2500	8.09	6.86	5.71	0.81	0.69	0.57
σ_{st}	500	−6.78	−6.13	−4.97	−0.00	−0.00	−0.00
	2500	−1.90	−3.69	−1.87	−0.00	−0.00	−0.00
ρ	500	35.60	21.37	10.46	0.04	0.02	0.01
	2500	65.98	46.40	40.98	0.07	0.05	0.04
ω	500	−4.44	−3.17	−1.69	−0.04	−0.03	−0.02
	2500	−6.57	−3.85	−2.08	−0.05	−0.03	−0.02

estimate more variability in speed than exists. The unacceptable bias observed for the person-item-distance parameter (ρ) is more concerning. The sample size was found to have the largest effect of the estimates of ρ (η^2 =11.6%), which can be seen in the drastically different estimate of relative bias for the conditions with 10 and 20 items for sample sizes 500 versus 2500. The somewhat confusing part of this result is that the lower sample size conditions tended to be less biased on average for this parameter. Similar to the results of the measurement model, this parameter could be highly influenced by the prior in an unforeseen way. When using dichotomous indicators, the overall conclusion is measurement model parameters tend to be negatively biased, response time parameters tend to be unbiased, and reliability estimates tend to be negligibly negatively biased. Next, for models estimated using indicators with five

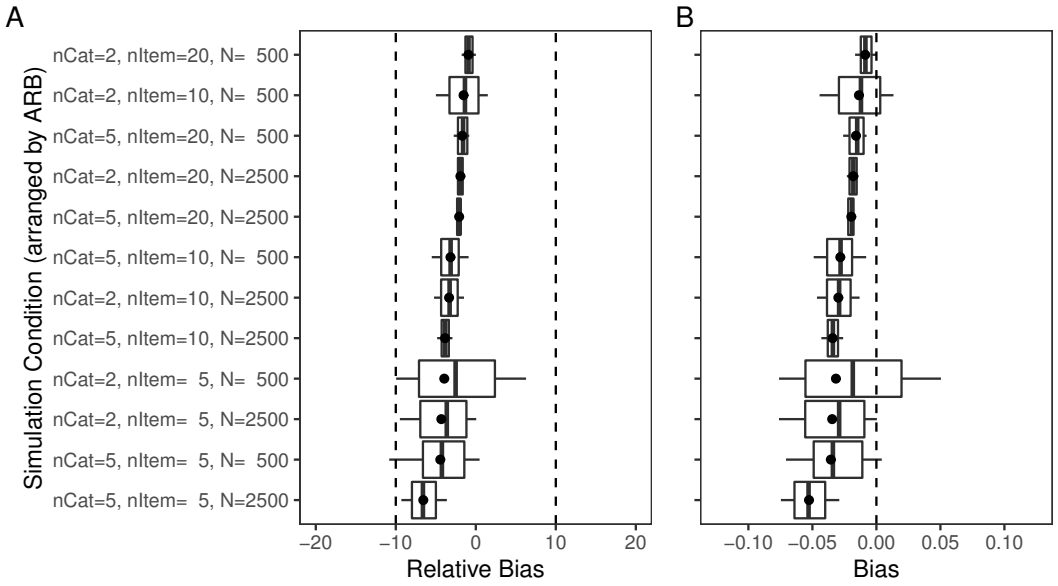


Figure 5.3. Bias of posterior median estimate of factor reliability. (A) relative bias estimates where dashed lines represent ±10% relative bias, (B) average bias of posterior median. ARB = average relative bias.

categories (see Table 5.11), the results were similar to the dichotomous item results but with notable differences. Similar to the previous results, a persistent negative bias was found for most parameters in the latent factor portion. The parameter estimates for the response time portion of the model were also similarly mostly unbiased on average. The variance of the speed factor was mildly biased in conditions with 5-10 items but otherwise acceptable.

The distributions of estimates of relative bias for reliability estimates are shown in more detail in Figure 5.3. On average, the least biased estimates of reliability occurred in the condition with dichotomous items, 20 indicators, and a sample size of 500. The differences between the population values of reliability and the posterior median are plotted in Figure 5.3B. The results show that the reliability estimates are typically within 0.05 of the population value on the scale of reliability. So even though, on average, the estimates are mildly biased, the estimates are still pretty close from a substantive point of view.

Posterior Credible Interval Coverage

Bayesian inference is central to the models discussed in this project, and credible intervals are a key component of interpreting results. The performance of 95% credible intervals to capture the population values of the model parameter is evaluated in this section. Across conditions and parameters, the credible intervals are expected to contain the population value for 95% of replications. Additionally, credible intervals that are narrower on average (i.e., smaller widths) are preferred as this indicates the posterior is more precisely estimating the population values. In this section, the coverage rate and interval widths are evaluated.

First, the coverage rates for the credible intervals across conditions are reported in Table 5.12. Coverage rates of credible intervals for factor loadings (λ) were below the nominal value of 95% in all conditions ranging from 42-92%. Coverage rates of factor loadings tended to be higher for dichotomous indicators and lower sample size conditions. The coverage rates for the item thresholds (τ) were excellent for dichotomous indicators (93-96%). However, the five-category conditions fell slightly

Table 5.12

Credible interval coverage rate

	Dichotomous						Polytomous					
N	500			2500			500			2500		
N Items	5	10	20	5	10	20	5	10	20	5	10	20
λ	88	91	92	83	85	78	88	86	84	68	56	42
τ	93	95	95	96	95	95	92	93	92	79	79	77
θ	88	90	92	83	84	78	89	86	84	68	56	42
β_{lrt}	98	95	93	94	93	94	96	95	91	88	94	94
σ_{lrt}	96	95	95	96	94	95	94	93	95	93	94	95
σ_s	87	95	89	75	64	68	88	89	86	71	62	62
σ_{st}	97	93	96	95	98	94	97	90	94	93	96	96
ρ	100	95	98	86	56	37	92	95	93	62	51	25
ω	94	97	92	79	24	1	89	46	33	5	0	0

Table 5.13

Summary of credible interval widths across conditions and replications

		Avg. Width						Width SD					
		Dichotomous			Polytomous			Dichotomous			Polytomous		
	N	5	10	20	5	10	20	5	10	20	5	10	20
λ	500	1.50	1.28	1.05	0.89	0.65	0.54	0.56	0.38	0.22	0.28	0.09	0.05
	2500	0.73	0.54	0.42	0.36	0.28	0.23	0.14	0.07	0.04	0.04	0.02	0.01
τ	500	0.65	0.63	0.60	0.53	0.49	0.48	0.11	0.09	0.05	0.11	0.05	0.04
	2500	0.28	0.26	0.25	0.22	0.21	0.21	0.03	0.01	0.01	0.02	0.01	0.01
θ	500	3.01	2.53	1.97	1.58	1.06	0.88	2.52	1.80	1.09	1.15	0.37	0.25
	2500	1.20	0.87	0.67	0.57	0.43	0.36	0.51	0.25	0.15	0.13	0.06	0.04
β_{lrt}	500	0.25	0.19	0.15	0.22	0.17	0.14	0.07	0.04	0.02	0.05	0.02	0.01
	2500	0.14	0.10	0.08	0.12	0.08	0.07	0.03	0.01	0.01	0.01	0.01	0.01
σ_{lrt}	500	1.32	1.12	1.05	1.20	1.09	1.04	0.38	0.11	0.08	0.13	0.09	0.07
	2500	0.54	0.49	0.47	0.52	0.48	0.46	0.05	0.02	0.02	0.02	0.02	0.02
σ_s	500	6.41	3.90	3.18	5.09	3.66	3.10	2.05	0.66	0.38	1.08	0.41	0.27
	2500	2.54	1.79	1.46	2.22	1.63	1.38	0.45	0.15	0.08	0.28	0.10	0.06
σ_{st}	500	0.16	0.11	0.09	0.12	0.09	0.07	0.02	0.01	0.01	0.01	0.00	0.00
	2500	0.07	0.05	0.04	0.05	0.04	0.03	0.01	0.00	0.00	0.00	0.00	0.00
ρ	500	0.32	0.16	0.11	0.25	0.17	0.12	0.12	0.04	0.02	0.06	0.03	0.01
	2500	0.17	0.10	0.07	0.15	0.09	0.06	0.03	0.01	0.00	0.02	0.01	0.00
ω	500	0.31	0.10	0.04	0.15	0.06	0.03	0.10	0.03	0.01	0.03	0.01	0.00
	2500	0.13	0.05	0.02	0.07	0.03	0.01	0.03	0.01	0.00	0.01	0.00	0.00

Note. Average width is computed using the average difference of upper 97.5%-tile and lower 2.5%-tile of the probability interval. Width standard deviation (SD) is computed similarly.

for the lower sample size conditions (92-93%) and fell drastically in the larger sample size conditions (77-79%).

A similar pattern was of good coverage for dichotomous indicators. Still, poor coverage in polytomous indicators at a large sample size was observed for the variance of the latent response variable (θ) and speed factor variance (σ_s). Coverage rates were not perfect but acceptable for the response time intercepts (88-98%), response time residual variance (93-96%), and factor covariance (90-97%) across all conditions. The coverage was acceptable for the person-item distance relationship parameter ρ when the sample size was 500 (92-100%), but coverage fell drastically when the sample size increased (25-86%). Lastly, the coverage of the credible intervals for the reliability estimates were acceptable for at the sample size of 500 for dichotomous indicators

(92-94%) but not for polytomous indicators (33-89%). At the higher sample size conditions, the posterior distributions for reliability did not cover the population value (0-79%).

The last component of the credible intervals evaluated in this project was the interval widths. The average width and the variability in width size across replications are reported in Table 5.13. For average width, an indication of the certainty of the posterior distribution, the widths varied substantially across conditions. For the factor loadings, the average width ranged from 0.23-1.50, where the narrowest intervals occurred for the polytomous items and larger sample size. This pattern occurred for all parameters. However, for the item threshold parameters (τ), the average widths remained fairly consistent across sample size conditions. The variability in width length across replications was not substantial for item thresholds (SD ranged from 0.00-0.11), response time intercepts (0.01-0.05), response time residual variance (0.02-0.13) factor covariance (0.00-0.01), person-item distance parameter (0.00-0.06), and reliability (0.00-0.03). More substantial variability in interval widths was observed for factor loadings (0.01-0.28), latent response variance (0.04-1.15), and speed factor variance (0.06-1.08). More consistent interval widths are preferred, giving more confidence in the repeatability of results across repeated samples.

Applied Study 1: NAEP Math Identity and Process Data

NAEP Response Time for Math Identity Items

The response time distribution for each of the NAEP math motivation items is shown in Figure 5.4. The response time distributions for the full NAEP sample are shown in Figure 5.4A. The response time distributions for the analytic sample are shown in Figure 5.4B. The response time distributions show the potentially bimodal distribution for items 2-5, and item 1 does not show substantial evidence of more than one mode. The computation of response times used the time from entering the

item to responding to the first statement as the response time for item 1. Responding to item 1, therefore, includes reading the item stem and the first statement. Reading both components lead to the necessarily higher response time for item 1 for most respondents.

NAEP Math Identity Scale Reliability Estimates

The posterior distributions for estimates of scale reliability (ω) across the four estimated models are shown in Figure 5.5. The results show how differentially weighting item responses using response time can drastically shift the posterior distribution for ω to higher values. The differences between using the observed response time versus the ELRT as informative of the misclassification rates were negligible. However, the model using the observed response time as informative the misclassification rates was

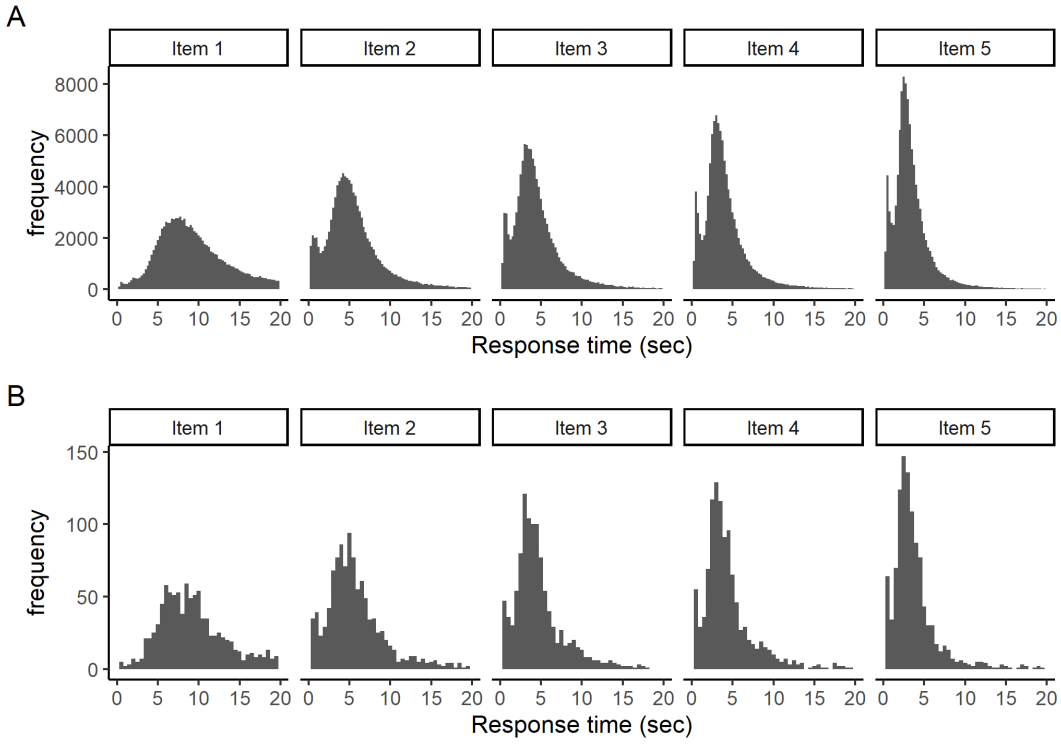


Figure 5.4. NAEP math identity item response times. (A) response time distributions for all eligible NAEP data, (B) response time distribution for analytic sample (1% random sample of full NAEP data).

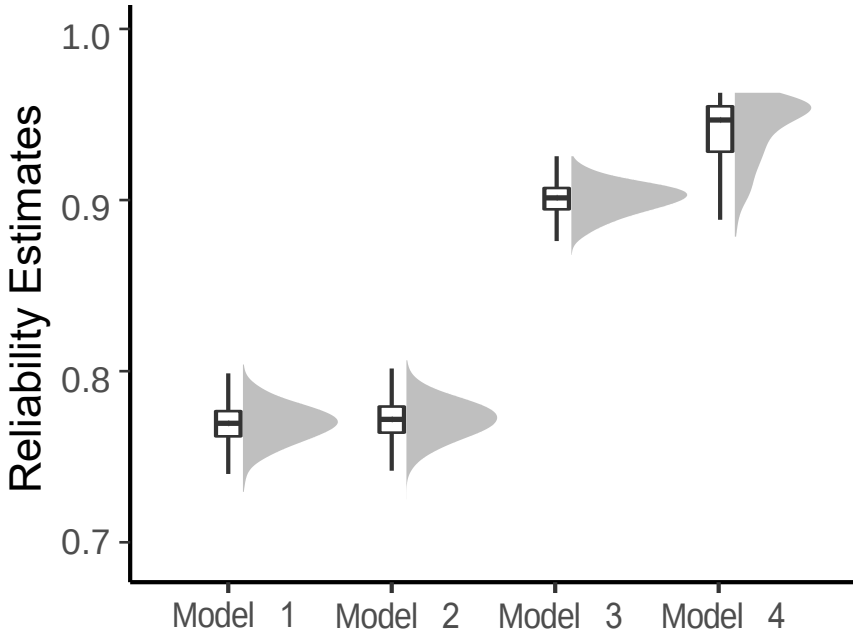


Figure 5.5. NAEP math identity scale reliability estimates.

estimated more quickly as this model appears to sample the posterior distributions more efficiently.

Applied Study 2: Extroversion Data

In this section, the extroversion data and full model results will be more rigorously tested to demonstrate the prior-to-posterior sensitivity of the full model. First, the results across the four models are described. Then, the prior-to-posterior sensitivity analyses are discussed.

Model Results

The full results of the four estimated models are given in Appendix E. Models 1-3 did not differ substantially from the results of the full model. Of major interest at this point is how reliability estimates differ across models. The posterior distributions of ω across the four models are shown in Figure 5.6. The models all had posterior means and medians above 0.90, which indicates that the differences in responses to

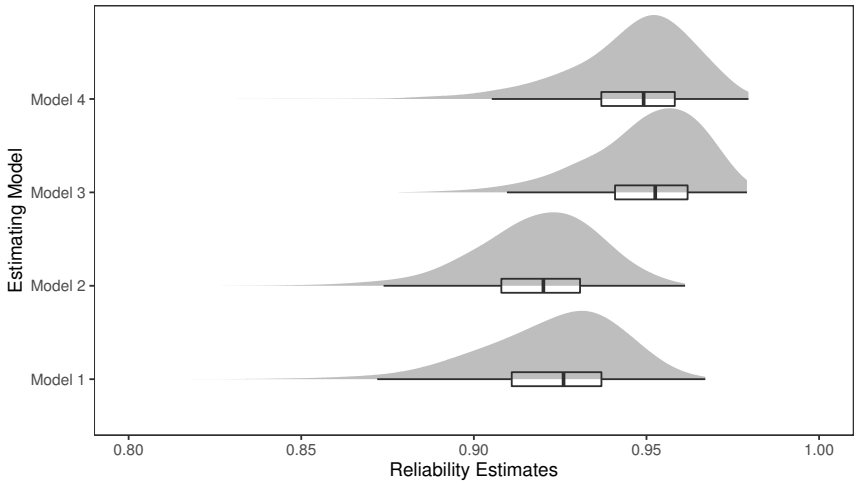


Figure 5.6. Extroversion scale reliability estimates.

the items may be attributable to a single latent factor because McDonald McDonald (1999, p. 89) described how ω can be interpreted as the the "Omega is the square of the correlation between the [total score] and a [factor] (Property 1)." Based on these results, the conclusion is that these items may be used to estimate a single latent factor with a high degree of reliability. However, how defensible is this conclusion is the purpose of the sensitivity analysis that is to follow.

Prior-Posterior Sensitivity Analyses

The estimates of ω for the extroversion data were heavily dependent on the prior structure specified. In the first part of the sensitivity analysis, the prior distribution for the factor loadings was varied across the six prior specifications outlined in Table 4.2. The posteriors and prior distributions are shown in Figure 5.7. Changing the priors from the theoretically needed positive truncated normal distributions to the more relaxed priors used in Bayesian factor analysis with continuous indicators did not appear to influence most posteriors substantially An exception occurred when the prior precision was 10 and the posteriors were heavily pulled towards zero. Too much precision (10) or too diffuse (0.01) resulted in the posteriors pulled towards zero

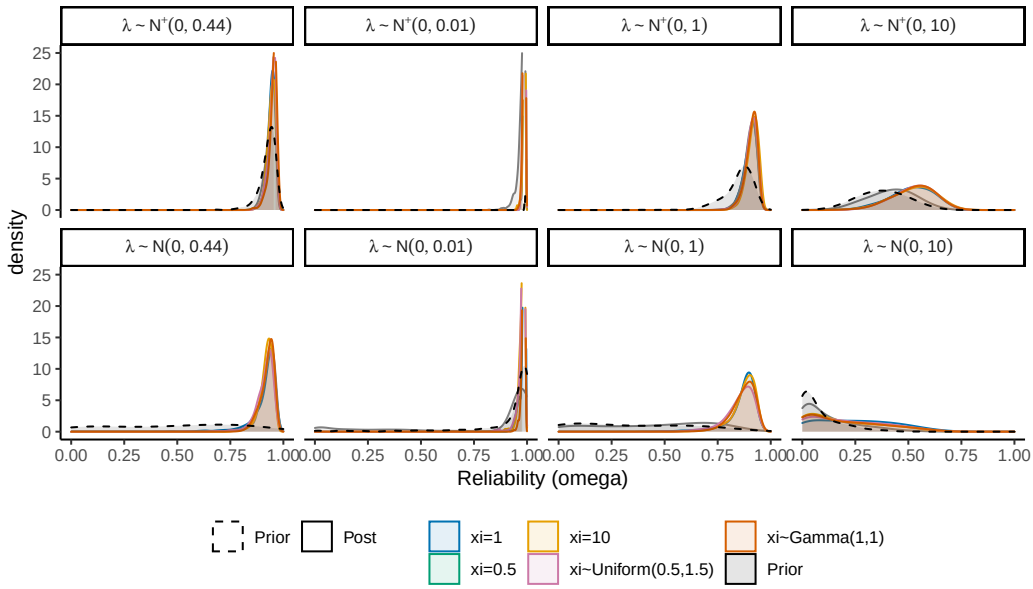


Figure 5.7. Reliability posterior sensitivity across varying factor loading priors. *Note.* The normal priors are precision parameterized.

or one, respectively. Additionally, the value chosen for the tuning pattern ξ did not seem to influence the posterior distribution of ω in most cases. For example, when the priors for the factor loadings were $N(0, 1)$, the tuning parameter was influential only when $\xi = 0.1$. Most cases' lack of influence of the tuning parameter suggests that the chosen value of $\xi = 1$ for the rest of this project is likely defensible and did not influence the results meaningfully. A more detailed numerical summary of the posterior distributions shown in Figure 5.7 is given in the appendix in Table E.5.

CHAPTER SIX

Discussion

The discussion for this dissertation is broken up into three major sections. The first general comments on the dissertation are discussed. The general discussion includes comments on the utility of the proposed methods. Next, two sections are devoted to the simulation results and real data analyses. These sections include a component focused on recommendations for analysts. Lastly, directions for future research are discussed and I end with a few concluding remarks.

General Discussion

A major aim of psychometric modeling is to explain the response process that gave rise to the observed data. The modeling goal is to develop methods representing the data generating process, leading to factor analysis, item response theory, and Rasch models. These methods can be grouped as measurement models for an underlying construct of interest. To measure this underlying construct, a set of items, or indicators, are used to reflect the construct. The observed response is caused by the level of trait of a respondent in a structured way. Depending on the measurement model, the structured relationship between observed responses and latent traits may be constant across all levels of the trait (linear factor analysis) or depend on the level of the trait (nonlinear factor analysis, IRT, or Rasch models). Any variability in the observed response after the trait has been accounted for is measurement error. The measurement error component is then used to help assess how well the data conform to the hypothesized measurement model.

The hypothesized measurement model forms the basis of the more general latent variable modeling framework for modeling complex systems. The complex systems can

be viewed as representing the data generating process. The data generating process is a far too often overlooked modeling opportunity in educational and psychological research. In this project, I aimed to develop an approach to modeling the data generating process underlying responses to self-report items that incorporates external information about the respondents. Information about responses (i.e., response time) was incorporated to explain the measurement error component of the model. Decomposing the error sources is the spirit of generalizability theory (G-Theory; Brennan, 2005, 2010; Webb & Shavelson, 2005). The connection to G-Theory is utilized in a modeling perspective instead of a design perspective. The models for decomposing measurement error can be flexibly defined within a general latent variable modeling framework.

The latent variable modeling perspective can provide a useful framework for formally testing hypotheses about difficult-to-measure constructs. Such investigations may arise from data collected from individuals through surveys can be analyzed in various ways depending on the type of data and the analyst's goal. However, the data collected are prone to confounding factors out of the researchers' control. And one unavoidable part of any measurement process is the issue of measurement error. Evaluating the impact of measurement error on inferences is about evaluating the validity of inferences.

The methods developed in this project were found to benefit the evaluation of reliability in simulated and real data analyses. Single estimates of reliability for a measurement tool are limited in information but provide a useful summary point for understanding the strength of the relationship between the observed scores and the unobserved trait (p. 89, property 1 McDonald, 1999) or the potential consistency of scores across similar conditions (p. 89, property 2 McDonald, 1999). The limited information provided by ω may be overcome if an IRT approach to reliability through the *information function* is taken (see Wise & Demars, 2006). Evaluating the infor-

mation provided across the range of trait scores may yield different inferences about measuring different parts of the latent dimension. A limitation of ω or information approach is the lack of ability to probe for potential issues in the results based on theory-driven hypotheses about measurement error.

In this work, measurement error modeled as misclassified responses was shown to provide a flexible approach for evaluating the sensitivity of reliability evidence to different schemes of incorporating response time as information in the model evaluation process. However, these methods are not limited to evaluating reliability estimates. The proposed measurement model for non-cognitive assessments is a unique approach for modeling and accounting for potentially invalid participant responses. Developing approaches to account for potential threats to the validity of inferences gained from assessments is an active area of research for educational and psychological assessment (Bowling et al., 2021; Curran, 2016; Huang et al., 2012; Huang et al., 2015; Ulitzsch et al., 2021). The growing literature has identified a variety of response styles that are prototypical of careless or insufficient effort (C/IE) responding (Curran, 2016). The literature on response effort has been seen as a special case of the literature for identifying and modeling response styles (Alessandri et al., 2010; Horan et al., 2003) and method effects more generally (Bradburn et al., 1978; Campbell & Fiske, 1959; Podsakoff et al., 2003).

A major difference between those methods and the misclassification-based approach proposed in this project is how the analysis accounts for the responses. This project focused on differentially weighting responses that show low effort through a low response time. The specification of the relationship between response time and measurement error is discussed in more detail in the following section. Directly weighting responses contrast with the methods for identifying response styles using mixture models (Cernat & Vandenplas, 2020)), which can be difficult to estimate. The benefit of the methods proposed in this project is the need to estimate a complex mixture

model is traded-off to estimate misclassification parameters that may not be "identified" from the survey responses alone. The trade-off in complexity may be defensible when the purpose is to evaluate the reliability of validity evidence associated with a measurement tool. But, the mixture modeling approach may be more useful when researchers need a way to identify respondents that do not demonstrate effort.

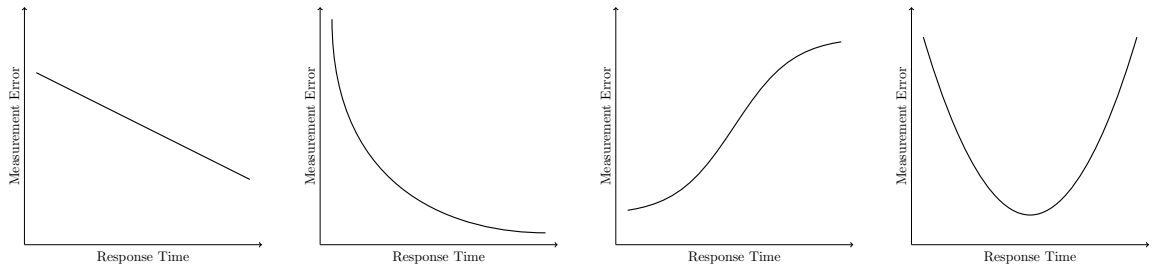
Expanding the flexibility to incorporate a wide range of measurement scenarios will be useful for a variety of psychometric applications such as scale construction, fraud detection, construct evaluation, etc. When constructing a measurement tool is of primary interest, researchers have many analytic choices concerning the combination of models to use. The models may utilize response time which has a growing component to understanding the measurement of personality constructs (Molenaar et al., 2021; Ranger, 2013). Investigating how the personality or psychosocial construct relates to response time can aid item development and evaluation (De Boeck & Jeon, 2019; Ranger & Kuhn, 2012) and develop flexible approaches for combining models. For example, the bivariate generalized linear IRT (B-GLIRT) framework provides researchers with a flexible approach to jointly model responses and response time (Molenaar, Tuerlinckx, & van der Maas, 2015b). The misclassification approach complements the B-GLIRT approach by providing researchers a mechanism for modeling item-level measurement error using external sources of information (e.g., response time). However, researchers are not limited to using response time but can utilize any information they theorize to be relevant. Additional information on measurement error may be gathered from expert judgment about the construct under investigation (Groves, 2004). For example, error may increase when items ask respondents to recall past events, judge or evaluate opinions, attitudes or "nonattitude" towards the topic, motivation, or response tendency (e.g., socially desirable responses, nah-saying, etc.) which all are threats to the validity of the data we obtain (Groves, 2004, p.407-441).

Accounting for potential threats to the validity of inferences drawn from surveys are major benefit of the methods developed in this work.

Response Time and Measurement Error

A point of contention in the methods developed in this project is constructing the relationship between response time and measurement error. In this project, I limited the scope to the measurement of psychosocial constructs such as personality where theory suggests that response time is related to the latent trait through a distance function (Ferrando, 2006; Ferrando & Lorenzo-Seva, 2007b; Holden & Kroner, 1992; Kuncel, 1973; Meng et al., 2014; Molenaar et al., 2021; Ranger, 2013; Ranger & Ortner, 2011). Additionally, van der Maas et al. (2011) discussed how measurement error increases as response time decreases on items as an individual's distance to item changes. Therefore, developing a modeling approach for specifying how the response time relates to measurement error is part of the novelty of this project. Due to the novelty, the relationship was specified such that a shorter response time implies more measurement error which is similar to how methods for identifying low-effort responses uses a low response time as an indicator (Meade & Craig, 2012; Rios & Soland, 2021; Wise & Demars, 2006). However, other specifications are possible. For example, response time could be non-monotonically related to measurement error where the functional form of the relationship could take on a parabolic structure, similar to ideal-point IRT models. Four possible representations of how response time could relate to measurement error are shown in Figure 6.1.

The model developed as part of this project can incorporate one's hypothesis of the relationship between response time and measurement error. The relationship is modeled as part of the link function $g_E(\eta_p; \gamma)$ (see Equations 3.6 and 3.7). Deciding among different representations of this relationship is possible by estimating the model under each hypothesis and deciding which model best fits the data using



(a) Linear *(b)* Decreasing *(c)* Increasing *(d)* Parabolic

Figure 6.1. Functional form of response time-measurement error relationship. (a) Measurement error linearly decreases as response time increases; (b) Measurement error monotonically decreases as response time increases−similar to form theorized in this study; (c) measurement error monotonically increases and response time increases; and (d) measurement error is lowest at a particular "point" of response time leading to a parabolic relationship. *Note.* The scale is omitted for simplicity of discussion.

information criteria such as the deviance information criteria (DIC; Spiegelhalter et al., 2002); and this procedure was done by Molenaar, Tuerlinckx, and van der Maas (2015b) to test different representations of the person-item distance function the joint measurement model with response time. The choice of how the trait relates to response time is potentially non-trivial. The choice may influence how one should model the relationship between response time and measurement error. For example, if the person-item distance is strongly related to response time, then correctly specifying that relationship is needed to compute the effective latent response time accurately. However, if there is no strong relationship, the difference between weighting using the effective latent response time versus the observed response time may be negligible. The importance of this decision is likely application-specific.

The application-specific nature of joint models for response and response time is a feature and a limitation. As a feature, the unique nature of individual applications allows for a great breadth of utility. For example, the models allow for evaluation of cheating and fraud detection (Becker et al., 2021; Holden & Kroner, 1992; van der Linden, 2009), increased precision of estimation (Ranger & Kuhn, 2012; van der Linden et al., 2010; Wise & Demars, 2006), and unique insights into the trait (Molenaar et al.,

2021). However, due to the uniqueness of the models to each application, researchers may find deciding among many potential modeling decisions to be intractable. For example, Molenaar et al. (2021) discussed how modeling the person-item distance relationship (or "inverted-U effect") depends in part on the type of trait. The authors described how the trait might be bipolar (e.g., dependent-independent), leading to individuals having separate prototypes underlying their representation. A bipolar trait may then be expected to yield different relationships between response time and the trait depending on the pole an individual falls near. The authors made a similar argument for traits and decisions made under a diffusion process (Tuerlinckx et al., 2016; Tuerlinckx & De Boeck, 2005). However, a researcher may not believe their subject of study aligns well under either representation, and they then need to develop their hypothesis for how to relate the construct and response time. These decisions make nuanced assumptions about the response process, and may be difficult to conceptualize how these decisions are represented in the model specified. The methods developed in this project may help researchers find a way to model their hypothesis by providing a mechanism for relating the construct, response time, and measurement error.

Model Evaluation

In the previous section, I discussed how evaluating among models is possible for helping decide among a set of competing hypotheses of the person-item distance relationship and the link between response time and measurement error. Previous researchers in this context have used information criteria such as the DIC to help decide which representation best fits their data (Molenaar et al., 2021; Molenaar, Tuerlinckx, & van der Maas, 2015b). Molenaar et al. (2021) found in a simulation study that DIC can differentiate between correctly and incorrectly specified models. However, the conditions were limited to two separate models, and additional simulation studies would help to clarify if the DIC can differentiate among the various

different distance functions and measurement error link functions. Additional model selection methods may be useful as well, such as the widely applicable information criteria (WAIC), leave-one-out cross-validation (LOO), and Pareto-smoothed important sampling (PSIS) (Linde & van der Linde, 2012; Vehtari et al., 2017; Watanabe & Opper, 2010). An evaluation of these approaches to model selection would provide additional evidence for which model(s) researchers should focus their attention.

Related to model selection is evaluating the fit of an individual model. Within the framework developed in this project, the measurement error link function works to inform the misclassification matrices. The misclassification matrix posits that the observed data are inherently biased. The inherent bias is what the models try to model through the misclassification matrix; however, this poses a dilemma when evaluating the model-data fit. One approach to checking the fit is evaluating where an observed summary statistic (mean, item category proportions, etc.) is captured by the posterior-predictive distribution (Gelman et al., 2013; Gelman et al., 1996). A dilemma occurs when we believe the observed summary statistic is biased. Does capturing the biased observed summary statistics in the posterior predictive distribution mean our model predicts the bias, or would that imply that the model is not accounting for the bias? Evaluating the fit of the model under such circumstances is an open area, but I have yet to identify literature explicitly addressing this issue.

Despite the limitations of evaluating model fit, progress is being made. For example, Garnier-Villarreal and Jorgensen (2020) discussed how the commonly utilized fit measures (CFI, RMSEA, etc.) can be adapted and estimated within a Bayesian framework. Yet, they caution that these indices descriptively evaluate "the degree to which their model fails to reproduce the observed data; they were not developed to be test statistics." (Garnier-Villarreal & Jorgensen, 2020, p. 67). Additionally, other authors have been working to help develop and describe how to evaluate the fit of models within a Bayesian context (Ariyo et al., 2021; Levy, 2011; Zhang et al., 2022).

Levy (2011) discussed a variety of approaches to evaluate the data-model globally using familiar indices such as SRMR and fit for components of a model using residuals (i.e., local model fit). Adapting these model evaluation approaches to the scenario when the observed data are theorized to be biased would be directly useful for the methods developed in this study.

Discussion of Simulation Results

The results from the two simulation studies point to different aspects for how the results can be influenced depending on the analytic approach taken. In simulation study 1, factor loadings were consistently underestimated across all models. However, the degree of underestimation depended on whether item-level misclassification was modeled. Modeling item-level misclassification allows researchers to investigate how sensitive model results are to measurement error. In particular, modeling individual measurement error becomes possible to test whether inferences about a specific aspect of a model depend highly on the modeling approach. The results of simulation study 1 suggested that the proposed approach can provide a useful sensitivity analysis of results to effects of item-level measurement error.

The results of simulation study 2 suggested that the results of the sensitivity analyses should be taken with a grain of salt. Evaluation of the estimation of the full model found that the proposed approach to estimate the model may not be adequate for the system hypothesized. Estimating Bayesian item factor models is complex and only recently available in open source software (Merkle & Rosseel, 2018). I utilized a "latent response variable" formulation of the underlying item response process in this project. Still, my implementation was non-trivially different from the approach used in blavaan and presumably commercial software. The code used to estimate the full model is likely too verbose in how data were augmented relative to other approaches to the latent response formulation. The verbosity potentially results in less efficient sam-

pling from the posteriors of parameters of interest, such as reliability. For instance, the approach used in this project maps the latent response variable back to the probability scale as shown in the model specification diagrams (see Figure 3.2 and Figure 3.4). However, other implementations have been proposed (Albert & Chib, 1993; Chib & Greenberg, 1998; Gelman & Hill, 2006). The alternative implementations use the item threshold parameters to truncate the latent response variables and use the observed observations to define the regions of an individual's latent response. Using truncated latent response variables does not immediately illuminate how item-level misclassification could be incorporated, which is why the current implementation of this project was developed. The lack of congruence between the methods employed here for item factor analysis and recommended implementations by other experts of item factor analysis potentially limits the applicability of these methods. However, despite the lack of congruence, the underlying idea of representing item-level measurement error as a misclassified response has potential.

One potential area is for modeling misclassification in a broader nomological network (i.e., a structural equation model with multiple latent constructs). Unfortunately, the use of misclassification in such an application has not been conducted to the best of my knowledge. However, misclassification methods have been explored in univariate regression models (Goldstein et al., 2018; Goldstein et al., 2008; Gustafson & Le Nhu, 2002; Richardson & Gilks, 1993). For instance, predictors measured with error generally result in biased regression parameters (Fuller, 1987). However, expanding the results from such investigations into a broader latent variable modeling context has been less explored. Therefore, an open question is whether regressions among latent variables are biased when those factors' indicators are measured with error.

Recommendations for Analysts

Factor analytic methods are commonly used in social science for scale development or construct validation. The uses of factor analytic methods are not without potential downsides due to inherent uncertainty in measuring observed traits, which has been shown to influence inferences (Rigdon et al., 2019; Rigdon et al., 2020). Accounting for uncertainty in the effort or attention respondents give to item responses was the purpose of the current work. Accounting for inattentive responses helps increase the validity of inferences from a factor-analytic method. The validity of inferences also partially depends on how well the hypothesized model can be estimated with given statistical methods. In the two simulation studies of this work, I have shown that traditional factor models estimated to data simulated with misclassification can severely underestimate model parameters and estimates of reliability. We recommend that researchers using factor analytic methods investigate the sensitivity of model results to potential sources of item-level measurement error. The methods discussed in this project are one approach to probing those measurement error effects.

Discussion of Real Data Results

Two applied data examples were used in this study. The data for these two examples were unidimensional scales, which is simpler than most applied studies reviewed. However, the simplification helps demonstrate the misclassification's applicability in the proposed item factor analysis model. In the NAEP Data example, the results were fairly straightforward in understanding the reliability of Math Identity as measured by a set of five items selected by researchers at the American Institutes for Research. Modeling misclassification resulted in tremendous gains in the magnitude of estimates of McDonald's ω. The use of response time to inform the misclassification also resulted in a posterior distribution of ω with even higher values of reliability on average. The increase in reliability estimates suggests that differentially weighting

item-level responses based on response time can strengthen the relationship between the underlying factor and the observed scores. Therefore, the methods developed in this work could prove extremely useful for researchers utilizing surveys as all observed data can be used while accounting for individuals that may be less attentive to their responses. However, the approach to accounting for non-effortful responses contrasts with many methods for accounting for inattentive respondents. Other approaches focus on identifying such responses (Bowling et al., 2021; Meade & Craig, 2012; Niessen et al., 2016; Rios & Soland, 2021; Wise, 2017). The connection to response effort is explored further in the general discussion below.

For the results of the applied examples, the observed data characteristics suggested that individuals often took significantly longer to respond to the first item relative to the remaining items. In the NAEP data analysis, this occurred more clearly when the average response time for item 1 was about 10 seconds (2.3 on a log scale), but the remaining items decreased in average response time: 5 seconds, 4 seconds, 3.7 seconds, and 3 seconds, respectively. This trend was not observed in Extroversion Data. Still, the documentation for those data was not clear on item order, how the data were cleaned, or if all items from the survey were included in the open-access data file. An increased response time for the first item is likely due to the time needed to read the instructions for the item then this text is ignored on future screens. Knowing the first item is likely to result in a longer response time could be used in scale development. For example, scale developers could organize items such that an item is believed to require the most thought or most related to the underlying construct, which aligns well with using a reference indicator for scaling the latent variable in a unidimensional scale.

Choice of scaling method for latent variables impacts the prior specification. For instance, a reference indicator approach specified a fixed factor loading but a free factor variance. Therefore, researchers may need to decide which parameterization to

use based on what prior knowledge available to create informative priors. Defining defensible informative priors is a crucial step to Bayesian analysis, especially in smaller sample sizes (Koenig et al., 2022; Smid et al., 2020). In the Extroversion example, the priors chosen significantly impacted results. The priors for the factor loadings were especially influentially on the posterior for reliability. However, contrary to the recommendations for restricting the factor loadings to be positive (Levy & Mislevy, 2016), the sensitivity analysis results showed that the posterior for reliability was essentially unaffected by this decision. More work on how to define informative priors for factor loadings is needed.

The sensitivity analysis conducted as part of the extroversion data analysis was difficult to interpret. The sensitivity analysis used 40 separate model specifications defined by varying only the factor loading and tuning parameter prior values (see Table 4.2). The results can be summarized as the posterior distribution for reliability (McDonald's ω) was sensitive to the prior specification for factor loadings but less sensitive to the prior specification for the tuning parameter (see Figure 5.7). When the factor loading prior was the theoretically defined $\lambda \sim \mathrm{N}^{+}(0, 0.44)$, the posterior reliability was nearly the same as the induced prior on reliability. This result indicated that restricting the factor loadings to positive values (similar to IRT discrimination parameters) resulted in induced priors that were highly informative. The highly informative nature of the induced prior was especially true when the factor loading prior was most diffuse with precision set to 0.01 in the second column of the plot. When the prior precision for factor loadings was the largest at 10 (fourth column), the induced priors pulled the posterior of reliability towards the lower end of the possible values. The two extremes for diffuse and precise priors for the factor loadings resulted in posteriors that were barely updated from the data. The least informative prior on reliability was $\lambda \sim \mathrm{N}(0, 0.44)$, which resulted in posteriors that were negligibly different than when the prior was restricted to the positive reals. One interpretation

of the sensitivity analysis is that reasonably precision priors for the factor loadings (precision between 0.44-1) and most values for the tuning parameters will not impact inferences. The inferences under these specifications would be that the scale is reliable enough as all posteriors were above 0.70. The generally accepted lower threshold for the reliability of scales constructed for social science research in low-stake scenarios.

Additionally, the inference above is conditional on the specification of a reasonable precision for the prior on the factor loading. One potential way to account for the uncertainty in what the prior precision would be to specify a hyper-prior for the prior precision (see Gelman et al., 2013, p. 442-444 for an example). Gelman et al. (2013) discussed how specifying hyper-priors for unknown parameters can be a way to evaluate the sensitivity of posterior inferences to uncertainty in the prior specification. Gelman and colleagues varied the degrees of freedom of the t-distribution using a uniform prior to obtain a posterior that accounts for the uncertainty in the choice of prior. A similar idea can be done by varying the prior for the variance component of the normal priors for the factor loadings. For instance, the prior specification for factor loadings could be changed to

$$\lambda \sim \text{Normal}(0, \tau_\lambda)$$

$$\tau_\lambda \sim \text{Uniform}(0.01, 10).$$

The posterior of reliability could then be summarized as a bivariate density plot of the posterior for reliability by the posterior for the factor loading prior precision (as shown in Figure 6.2). The posterior summary demonstrates how the posterior of ω is insensitive the prior over a certain range of priors. However, the posterior was found to be sensitive when the precision was fixed to the boundaries (0.01 or 10). By accounting for uncertainty in the specification of the prior precision, inferences about the posterior reliability are less conditional on the specific values chosen in the initial sensitivity analysis. Hyper-priors other than uniform as are possible, such as

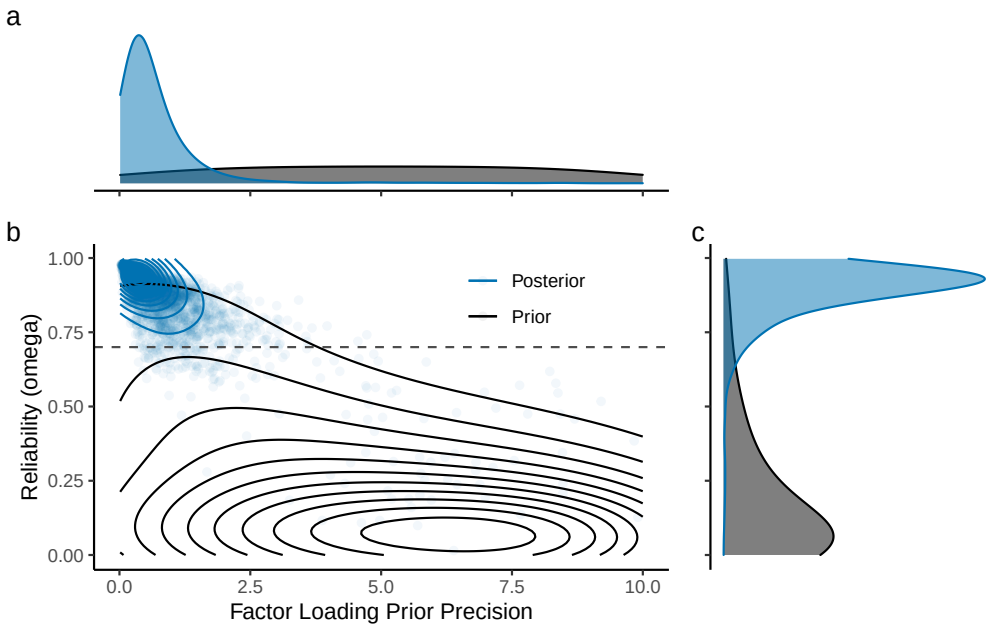


Figure 6.2. Expanded sensitivity analysis using bivariate density plots. (a) Marginal density of hyper-prior parameter τ_λ for factor loading prior precision; (b) Bivariate posterior density plot of hyper-prior and reliability; and (c) marginal density of reliability distribution with simulated induced prior on reliability. The MCMC samples for are plotted with alpha-shading in (b) to demonstrate how disperse some draws from the posterior were. The dashed line at 0.70 represents the common minimal acceptable level of reliability for scales in the social sciences.

the half-Cauchy distribution also used by Gelman et al. (2013) in a latter part of their analysis. This alternative specification for the hyper-prior is presented in the Appendix, and the conclusion is similar to the results of the uniform hyper-prior.

The prior-to-posterior sensitivity analysis above was restricted to parameters directly related to the posterior feature (reliability) of interest. The posterior distribution of reliability may be sensitive to other parameters in the model such as factor variances. Incorporating the prior for the (co)variance components would results in a more robust approach to the sensitivity analysis.

Evaluating the prior-to-posterior sensitivity is one consideration when investigating the results. Another consideration is the model specification. The model is specified using theory to guide how measurement error is incorporated. However,

the model specification can be evaluated. The section on *Model Evaluation* discussed model selection, but consideration of how inferences (e.g., how reliability is this scale) change depending on how the model is specified. Any model specified simplifies the data generating process to a degree, and by evaluating whether similar conclusion can be drawn under different degrees of model complexity or hypotheses of how measurement error may influence results lends more evidence to the decision we try to make using these data. For example, evaluating the reliability of measuring math identity using the NAEP data, the inferences about reliability could be drastically different depending on what level of reliability we set is needed. If we only need reliability to be at 0.70 for a low-stakes decision, then any of the models used meet the criteria. But, if we are plan to make more high-stakes policy-related decisions using the math identity construct as part of the model, we may say we need reliability to be at least 0.9 or higher. Under those conditions, we would need to be careful as to which model specification is used and be cautious of the appropriateness of which model specification(s) to use to help inform decisions.

Recommendations for Analysts

The choice of priors is nontrivial, and relying on default priors for complex models such as the one proposed in this project is not sufficient to decide based on the model. The specification of hyper-priors may be a useful way to help summarize results of sensitivity analyses. In the analysis conducted with the extroversion data, the initial sensitivity analysis varied among priors specified by other software and those suggested in the literature. However, the vast number of models results in the conclusion that the posterior for reliability was sensitivity to the prior for the factor loadings. Therefore, I recommend trying to identify over what priors the results are insensitive. The hyper-prior on the precision can be used in such a way to capture the range of theoretically defined priors plus all the values in-between to evaluate the

posterior of reliability relative to the precision. This approach can be used with any parameter of interest in the model to evaluate the sensitivity of the results to a range of priors with a similar structure.

Future Research Directions

One of the major contentions of the methods developed in this work is the specification of the relationship between response time and measurement error. Future work can focus on modeling multiple specifications and evaluating which specification is most likely for a given application. Model selection can be accomplished using the model comparison approaches discussed in the *Model Evaluation* section. Another approach is to develop a way to average over different specifications of measurement error. Averaging over different model specifications may be done using Bayesian model average (BMA; Hinne et al., 2020). BMA could provide a useful way to simultaneously account for multiple possible specifications of the relationship between response time and measurement error. However, the implementation of such an approach may be far down the line as more work is needed to evaluate the estimation of the individual models.

The implementation for item-factor analysis in this study should be compared to alternative parameterizations and implementations. For example, the results should be compared to the model estimated using the approach by Merkle and Rosseel (2018). The latent response formulation approach to conceptualizing latent variable models with categorical indicators is well documented (Kamata & Bauer, 2008; Muthén, 1984). The Bayesian approach to latent response variables, also known as data- augmentation, is well used in various contexts (Albert & Chib, 1993; Chib & Greenberg, 1998; Gelman & Hill, 2006; Naranjo et al., 2019). However, each implementation reviewed has been slightly different, resulting in the synthesized approach used in this study. More work on developing the estimation properties would provide more

evidence to utilize the method across contexts. Similarly, more understanding of efficiently parameterizing and estimating these models would make developing an open-source software package more user-friendly. For example, the motivation for developing an R package is that the software would provide a utility for these models to be more easily estimated by other users. The development of an easy to use software for such purposes would benefit the evaluation of more complex models within a wider structural equation modeling framework.

The use of misclassification methods developed in this project to evaluate the effects in a broader structural equation model could provide useful evidence as to the extent such item-level measurement error influences inferences about latent variables. In a model that specifies latent regressions, how are these structural parameters affected by the error of the indicators? This open question could be investigated using a Monte Carlo simulation study. In the study, the goal would be to evaluate the bias in the latent regressions. The study could be conducted using a similar process as simulation study 2, but focusing on the structural parameters. One of the simulation conditions would also be model specification whether correctly or incorrectly specified misclassification component. Additionally, the investigation would benefit from incorporating the prior as a condition to evaluate the performance over different degrees of prior informativeness.

Incorporating the prior specification into the Monte Carlo simulation study would provide evidence as to a potentially useful "initial" prior specification for other researchers. Recommending default priors for complex models has potential to unduly influence results in scenarios where the defaults overwhelm the data. However, providing guidance as to a prior specification that is relatively uninformative would be of utility to researchers. Similar, developing an approach for researchers to derive informative priors would be advantageous. For instance, Veen et al. (2020) used a panel of 14 experts to generate informative priors for a latent growth curve model.

Developing approaches to incorporate expert knowledge into measurement models would provide a valuable test of the developed measurement model.

In this study, I focused primarily on the use of McDonald's ω as defined in his text (McDonald, 1999, p. 88). The rational for the applicability of this formulation of ω is due to the use of latent response variables so that estimates of reliability are based on the continuous latent responses and not the categorical observed indicators. Another approach to investigate in future research is the use of categorical ω (Green & Yang, 2009). Categorical ω is computed utilizing the information captured in the thresholds and bivariate relationships among all items distinct from the commonly used McDonald's ω for linear factor analysis. The use of categorical ω is now recommended for estimating reliability in nonlinear factor models (Yang & Xia, 2019). However, an interesting comparison may be possible when comparing the results of categorical ω with the posterior of McDonald's ω from the methods proposed in this project that account for misclassification. The use of categorical ω instead of the traditional formula may help account for the last bit of under-estimation we found in nearly all Monte Carlo simulation study conditions.

Similar to extending the evaluation of the misclasification item-factor analysis methods and categorical-ω, the methods of this study can take an information approach to reliability. Reliability is not a scale feature evaluated within an IRT framework due to the use of the information function to evaluate which level of trait the scale provides the most information on. The information function can be complex as the model grows in complexity and deriving the item and test information functions would be a useful alternative to point estimates of reliability. The differences between the information functions of a traditional IRT model versus the effort-moderated IRT model was a primary outcome from Wise and Demars (2006). Extending the information approach to assessing the relative informativeness of the scale conditional on the

level of the trait would be a useful facet to consider for evaluating the measurement of a construct. Future work should be devoted to deriving these functions.

Concluding Remarks

In conclusion, the answers to the three research questions are

(1) Yes, modeling item-level misclassification can change inferences regarding scale reliability as measurement by McDonald's ω. How the inferences are effected by item-level misclassification depends on how misclassification is model (*Simulation Study 1*), characteristics of the data (number of categories, number of items, and sample size; see *Simulation Study 2*), and prior specification (see *Sensitivity Analysis*).

(2) The relative bias of point estimates and coverage rates of the credible intervals under the correctly specified model were adequate on average for most parameters and conditions. . In conditions with higher response categories (five and seven), the performance was worse in conditions with fewer items or a lower sample size. Additionally, coverage rates tended to decrease as sample size increases for parameters such as factor loadings, residual variances, person-item distance relationship, and reliability.

(3) Yes, using response time to differentially weight responses as informative priors for the misclassification rates changed inferences about the reliability of the NAEP math identity scale and the Extroversion scale. Similar to question one, the degree to which inferences were affected depended on how misclassification was specified.

Lastly, throughout this project, I have discussed a different perspective on modeling measurement error in survey items. The alternative perspective brings in a new way of thinking of the individual item-responses to highlight how external information will inform item-level measurement error. The methods developed in this study will

help researchers have more confidence in research findings when making inferences about underlying constructs.

APPENDICES

APPENDIX A

Instruments

NAEP Math Identity

Due to restriction on the NAEP process data in place by NCES, the amount of information available for sharing is limited. The summary tables provided below is intended for illustrative use only and is not necessarily representative of the nation.

Table A.1

NAEP data description and summary statistics of responses and response times

NAEP ID	TDDC ID	Item statement	Item Response	Response Time
M831501	VH269049	I want other students to think I am good at math	3.3 (1.3)	2.3 (0.6)
M831502	VH269050	I want to show others that my math schoolwork is easy for me	2.9 (1.3)	1.6 (0.8)
M831503	VH269053	I want to look smart in comparison to the other students in my math class	2.9 (1.5)	1.4 (0.8)
M831504	VH269059	I want to learn as much as possible in my math class	3.9 (1.1)	1.3 (0.8)
M831505	VH269056	I want to become better in math this year	4.3 (0.9)	1.1 (0.7)

Note. Item stem - *How much does each of the following statements describe a person like you? Select* ***one*** *answer choice on each row.*; item responses were recorded on a five-point Likert-type scale with response options *Not at all like me* (1), *A little bit like me* (2), *Somewhat like me* (3), *Quite a bit like me* (4), and *Exactly like me* (5); response time is reported in log seconds; estimates reported at the unweighted mean and standard deviation for item responses and response times, respectively.

Extroversion Data

A summary of all item in the extroversion dataset are given in Table A.2. Each item is simply a statement (habit) and each respondent was asked whether they think that statement describes them. The descriptors are translated from Dutch (Molenaar, Tuerlinckx, & van der Maas, 2015a).

Table A.2

Extroversion data item description

Item	Proportion Yes	Response Time Average (SD)
active	.74	1.49 (0.80)
noisy	.53	1.36 (0.65)
energetic	.85	1.12 (0.63)
enthusiastic	.95	1.00 (0.66)
impulsive	.92	1.30 (0.70)
jovial	.53	1.26 (0.68)
viable	.93	1.14 (0.54)
eupeptic	.95	1.09 (0.63)
communicative	.82	1.73 (0.75)
spontaneous	.86	0.99 (0.53)

APPENDIX B

Simulation Study 1

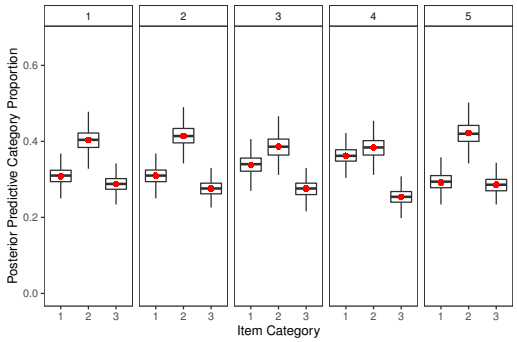
Data Conditions Simulated

The values for the parameters simulated are

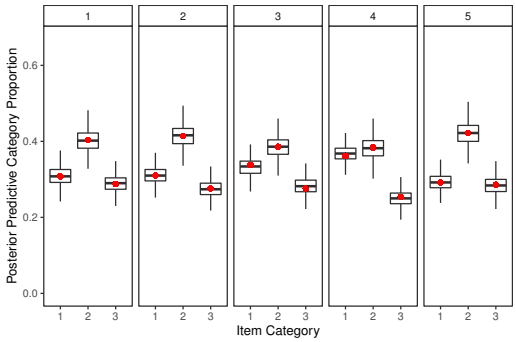
- $\lambda = 0.7$
- $\tau = \begin{bmatrix} -0.82 & 0.78 \\ -0.75 & 0.88 \\ -0.62 & 0.83 \\ -0.39 & 1.03 \\ -0.78 & 0.88 \end{bmatrix}$
- $\sigma_i = 1$
- $\mathbb{V}(\eta_1) = 1$
- McDonald's-$\omega = 0.83$
- $\beta = 1.5$
- $\sigma_{lrt} = 0.25$
- $\sigma_s = 0.1$
- $\rho = 0.1$.

R Code for Generating Item Threshold Parameters

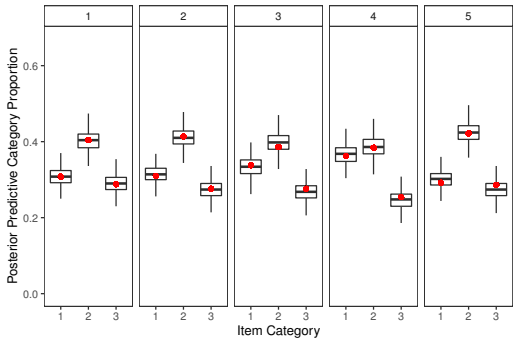
```
  getTau <- function(N_cat, N_items, seed=1){
  set.seed(seed)
  if(N_cat == 2){
    tau <- matrix(
      runif(N_items, -0.5, 0.5),
      ncol=N_cat - 1,
      nrow=N_items,
      byrow=T
    )
    tau = -tau
  }
  if(N_cat > 2 & N_cat < 7){
    tau <- matrix(ncol=N_cat-1, nrow=N_items)
    for(c in 1:(N_cat-1)){
      if(c == 1){
        $\tau_{,1] <- runif(N_items, -1, -0.33)
      }
      if(c > 1){
        $\tau_{,c] <- $\tau_{,c-1] + runif(N_items, 0.25, 1)
      }
    }
  }
  if(N_cat == 7){
        tau <- matrix(ncol=N_cat-1, nrow=N_items)
        for(c in 1:(N_cat-1)){
      if(c == 1){
        $\tau_{,1] <- runif(N_items, -2, -1.33)
      }
      if(c > 1){
        $\tau_{,c] <- $\tau_{,c-1] + runif(N_items, 0.33, 1.0)
      }
    }
  }
  tau = tau*lambda
  return(tau)
}
```



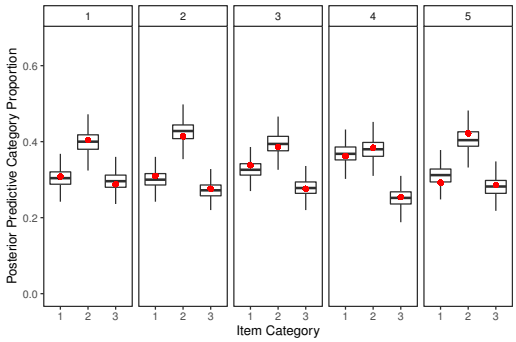
(a) Model 1



(b) Model 2



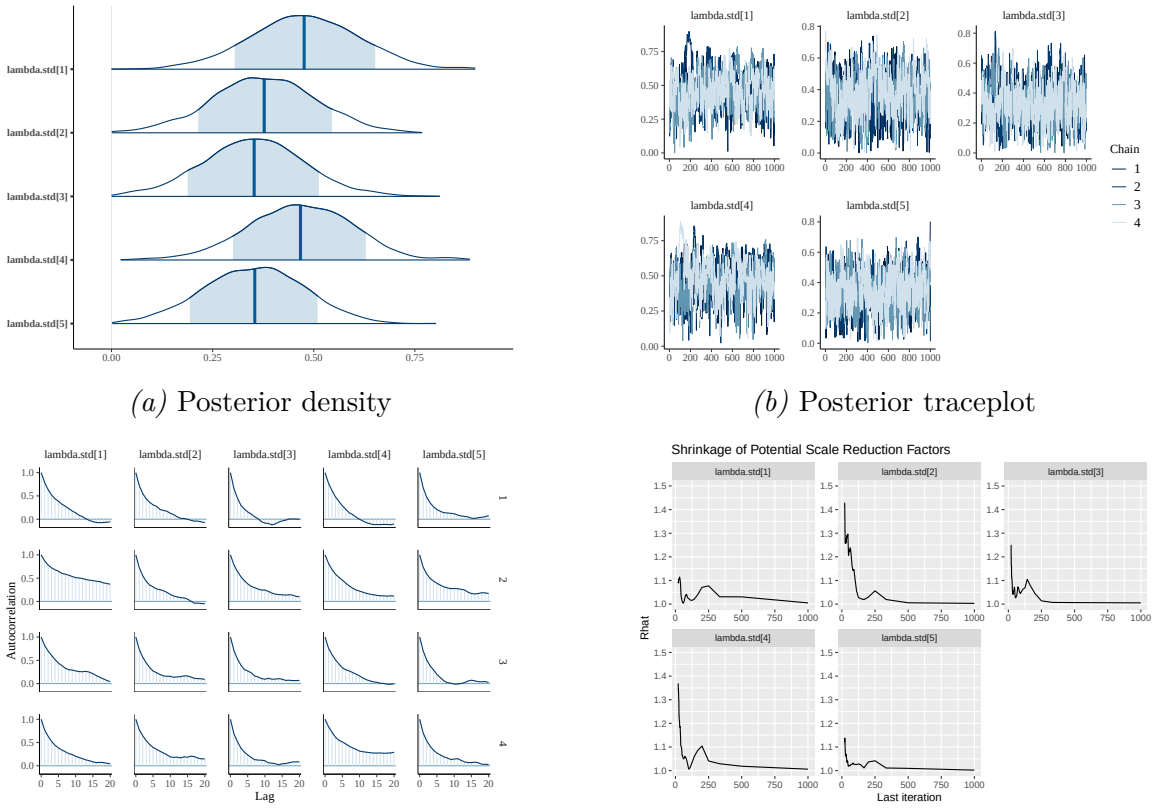
(c) Model 3



(d) Model 4

Figure B.1. Simulation study 1 posterior predictive distributions

Parameters in Model 1



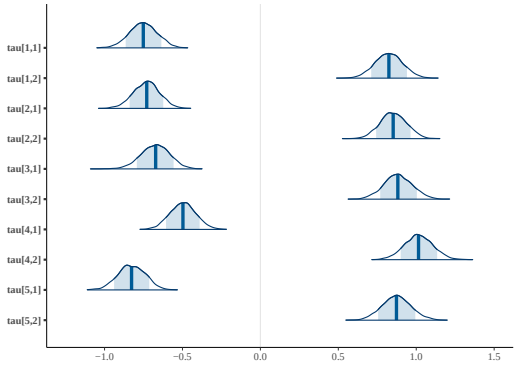
(a) Posterior density

(b) Posterior traceplot

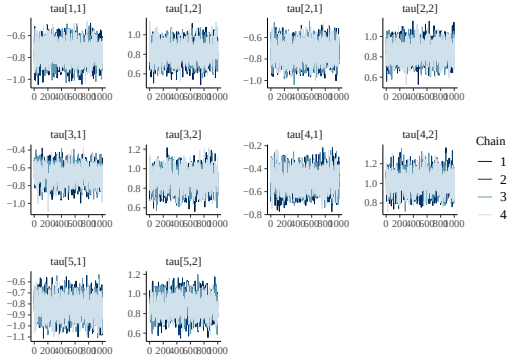
(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

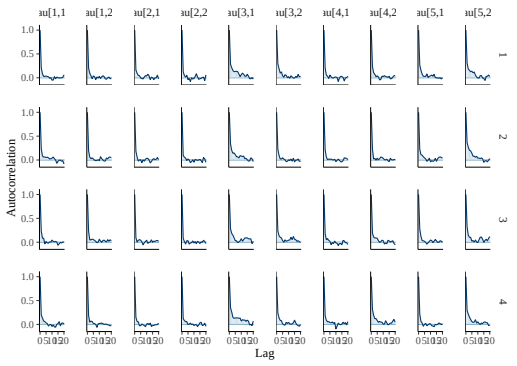
Figure B.2. Study 1 model 1: Posterior convergence evidence for factor loadings (standardized).



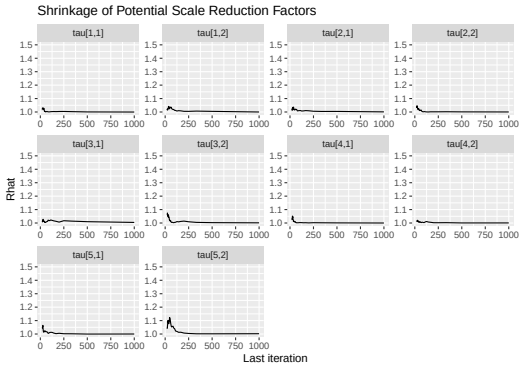
(a) Posterior density



(b) Posterior traceplot

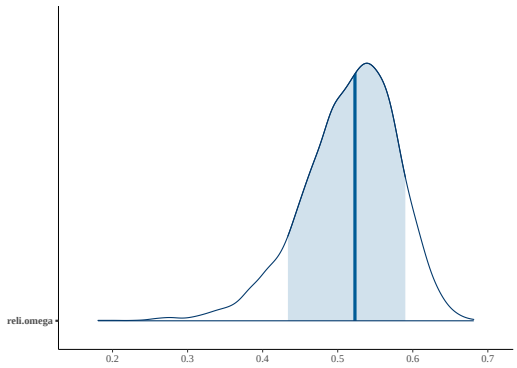


(c) Autocorrelation plot

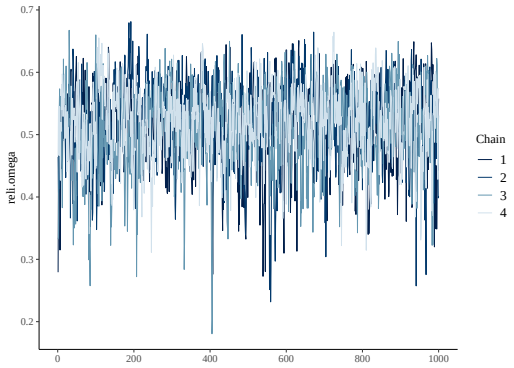


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

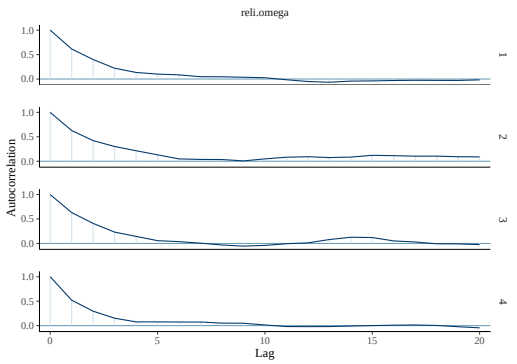
Figure B.3. Study 1 model 1: Posterior convergence evidence for item thresholds.



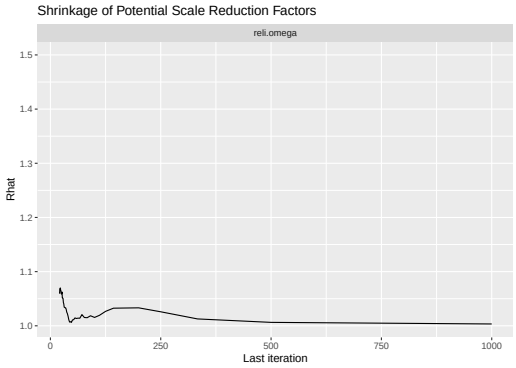
(a) Posterior density



(b) Posterior traceplot



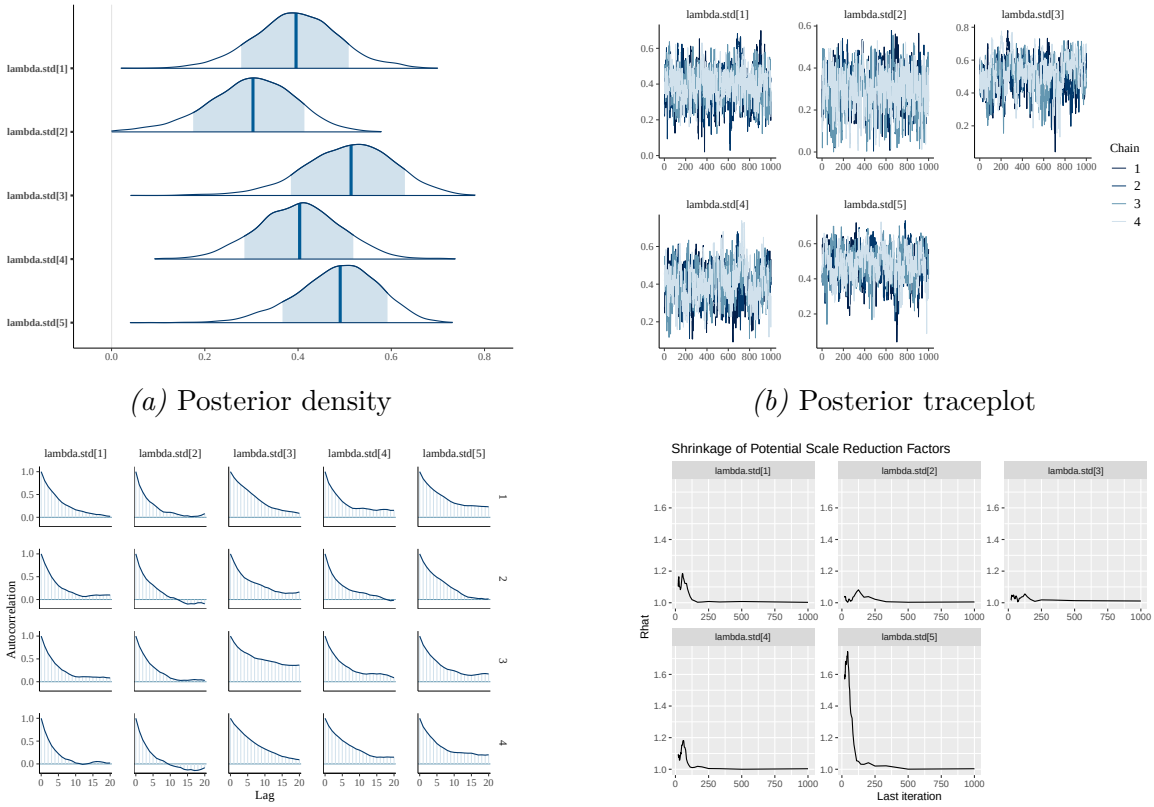
(c) Autocorrelation plot



(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.4. Study 1 model 1: Posterior convergence evidence for reliability (ω).

Parameters in Model 2



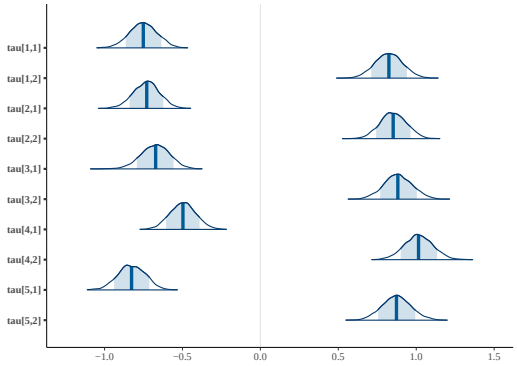
(a) Posterior density

(b) Posterior traceplot

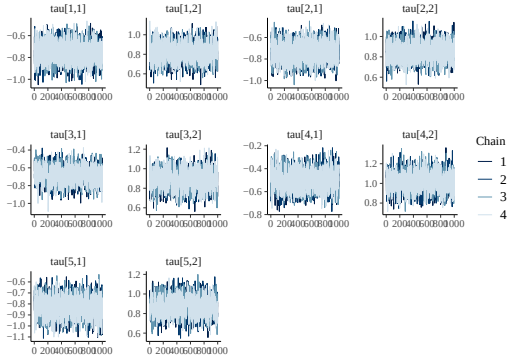
(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

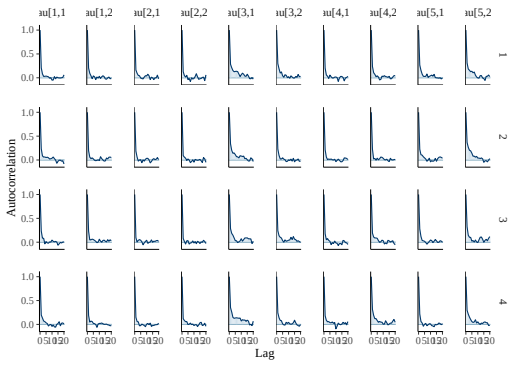
Figure B.5. Study 1 model 1: Posterior convergence evidence for factor loadings (standardized).



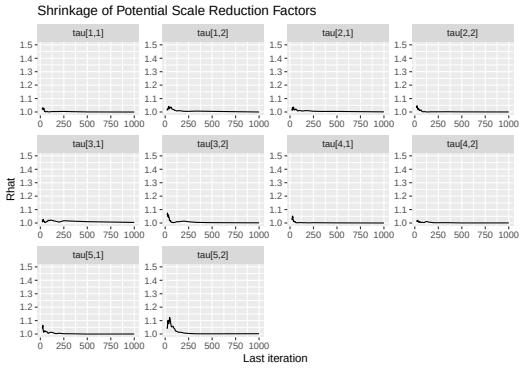
(a) Posterior density



(b) Posterior traceplot

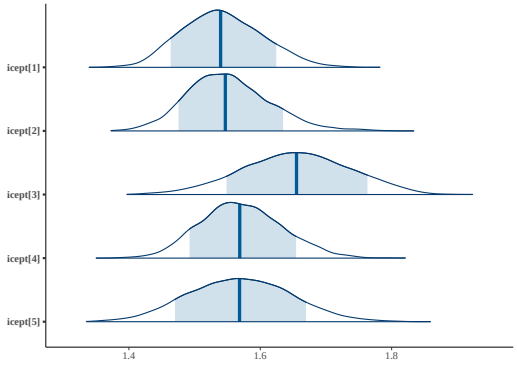


(c) Autocorrelation plot

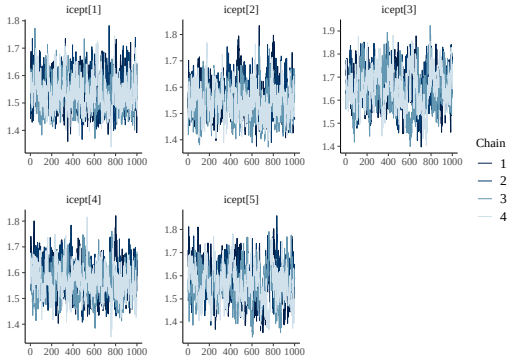


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

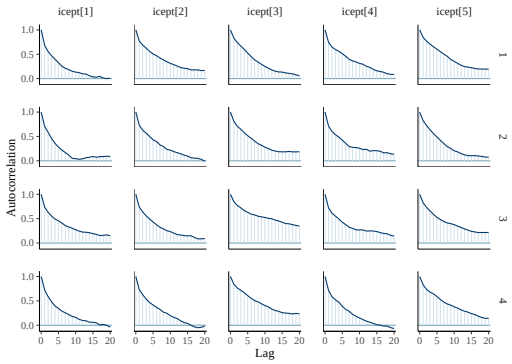
Figure B.6. Study 1 model 2: Posterior convergence evidence for item thresholds.



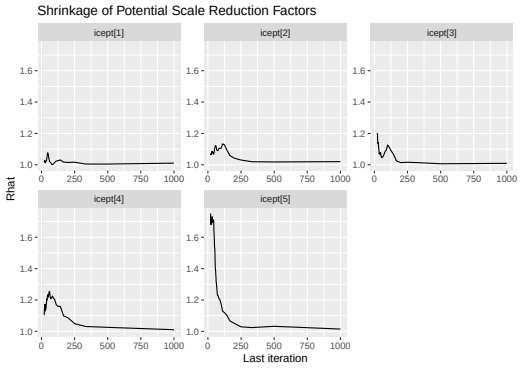
(a) Posterior density



(b) Posterior traceplot

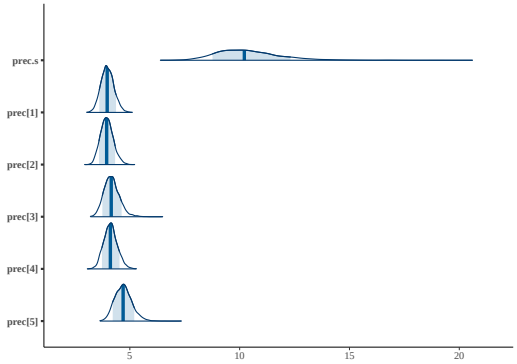


(c) Autocorrelation plot

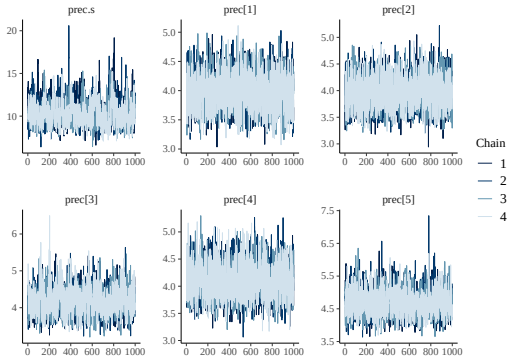


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

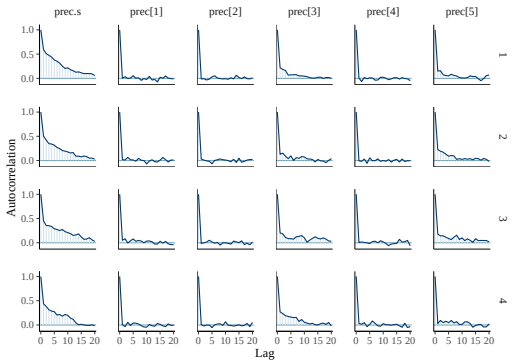
Figure B.7. Study 1 model 2: Posterior convergence evidence for response time intercepts.



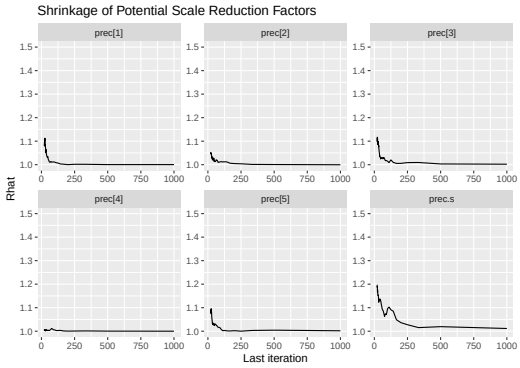
(a) Posterior density



(b) Posterior traceplot

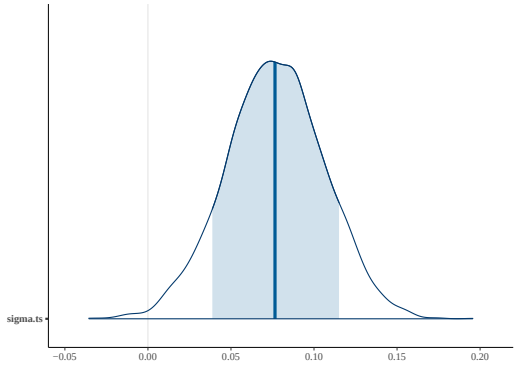


(c) Autocorrelation plot

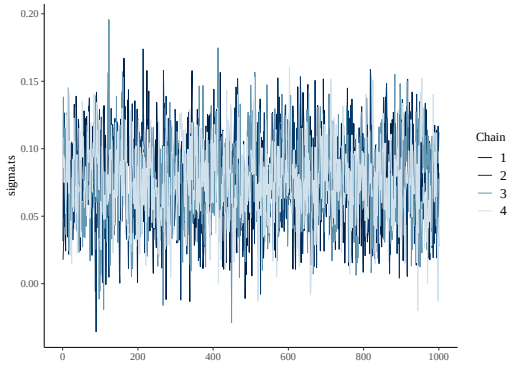


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

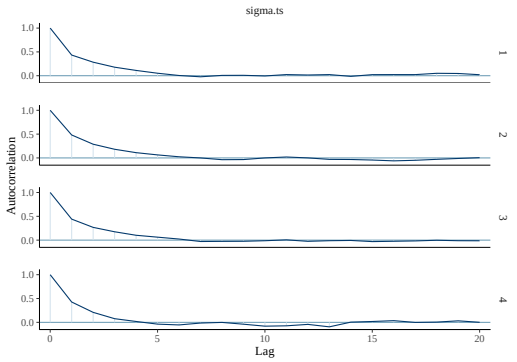
Figure B.8. Study 1 model 2: Posterior convergence evidence for response time item and factor precision.



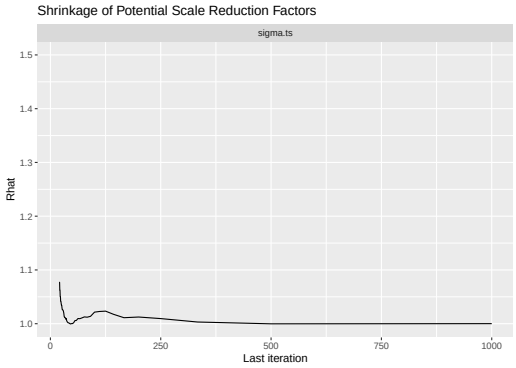
(a) Posterior density



(b) Posterior traceplot

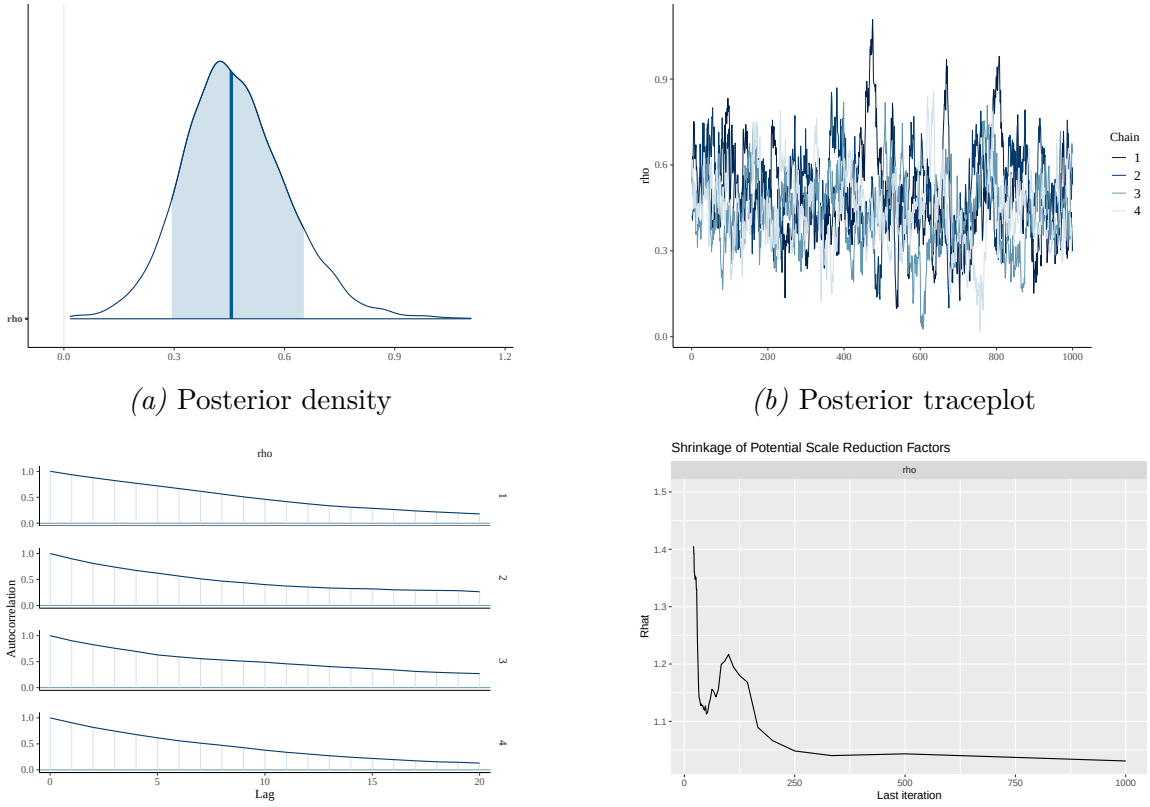


(c) Autocorrelation plot



(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.9. Study 1 model 2: Posterior convergence evidence for factor covariance.



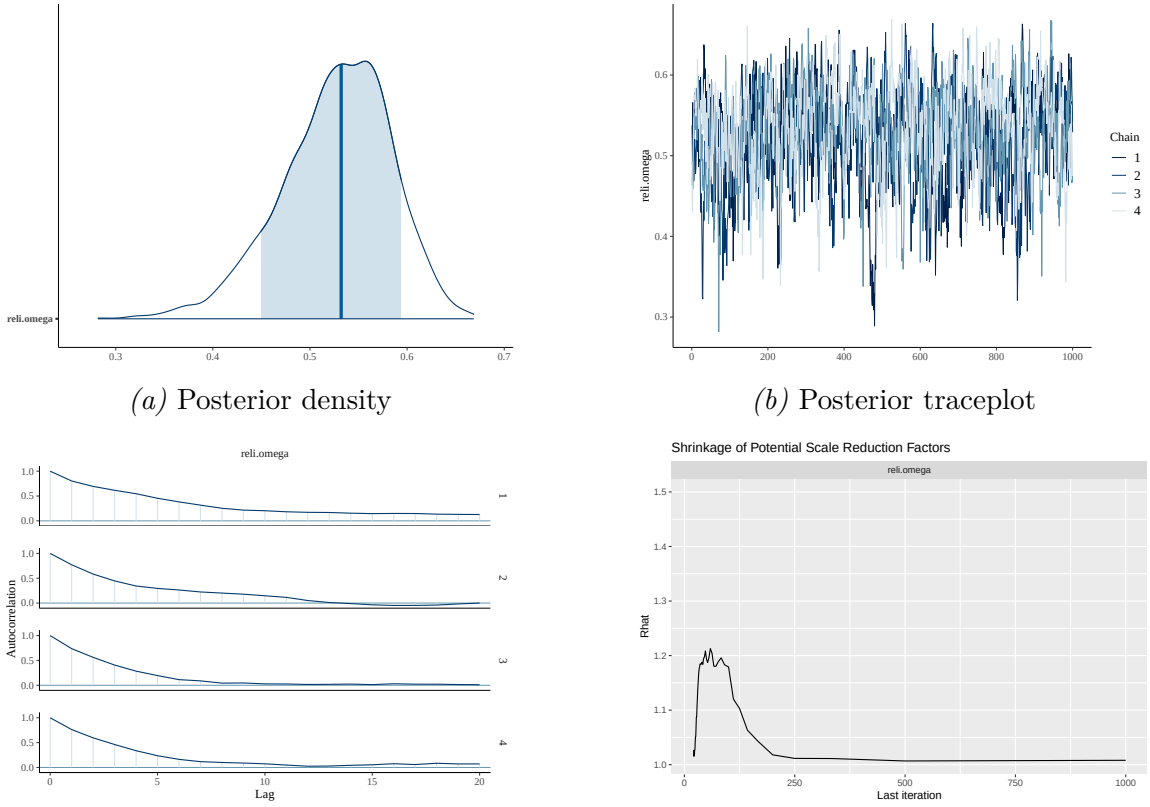
(a) Posterior density

(b) Posterior traceplot

(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.10. Study 1 model 2: Posterior convergence evidence for ρ.



(a) Posterior density

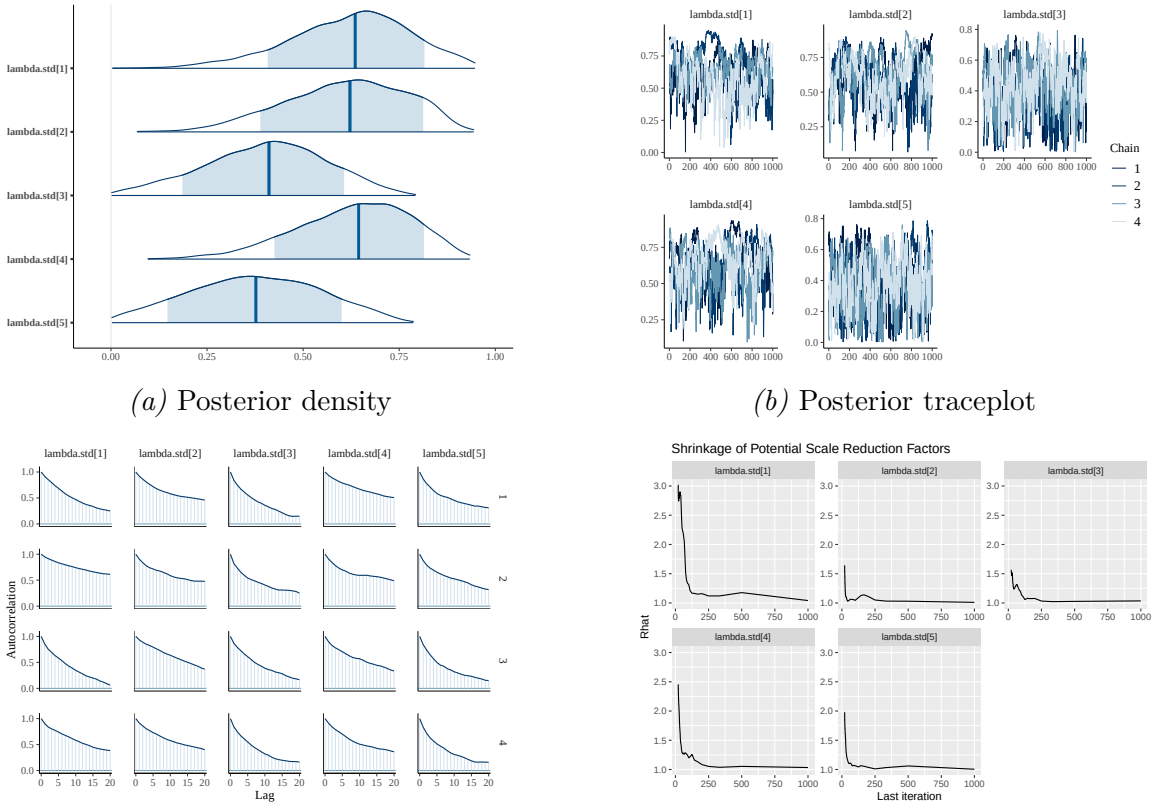
(b) Posterior traceplot

(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.11. Study 1 model 2: Posterior convergence evidence for reliability (ω).

Parameters in Model 3



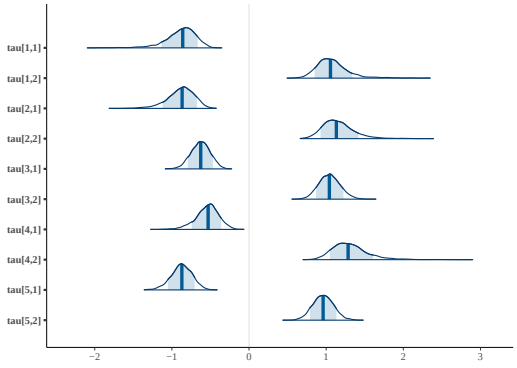
(a) Posterior density

(b) Posterior traceplot

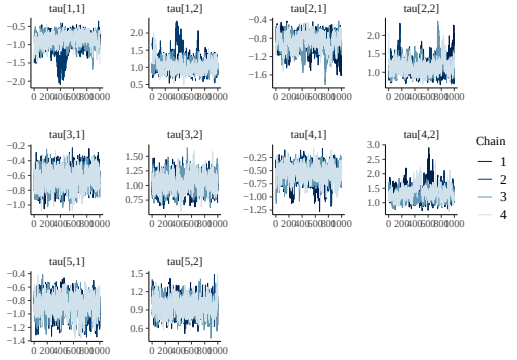
(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

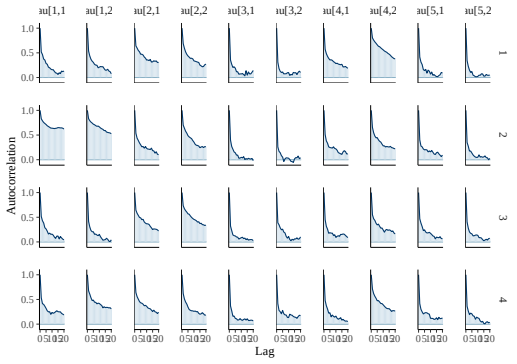
Figure B.12. Study 1 model 3: Posterior convergence evidence for factor loadings (standardized).



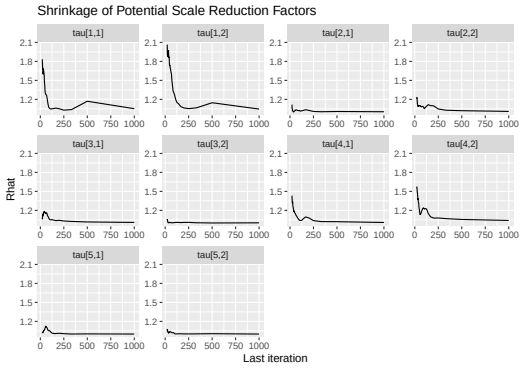
(a) Posterior density



(b) Posterior traceplot

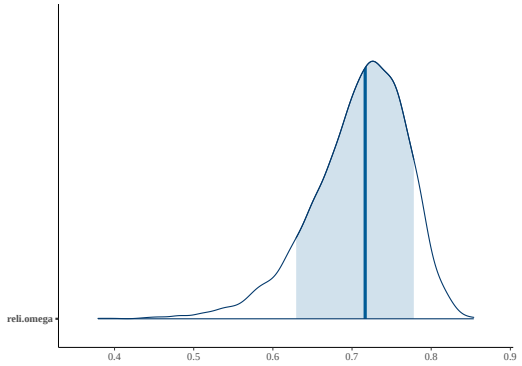


(c) Autocorrelation plot

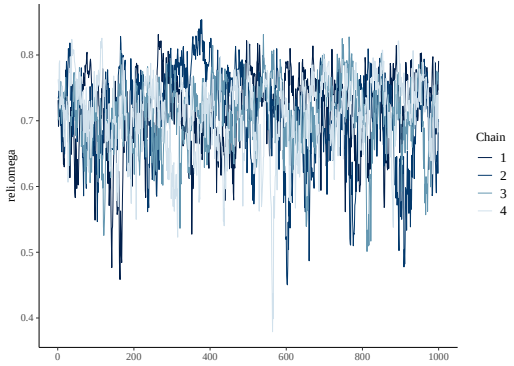


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

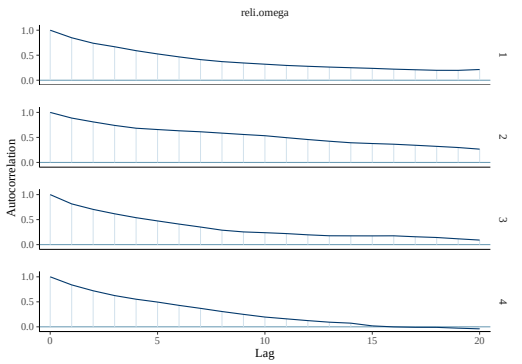
Figure B.13. Study 1 model 3: Posterior convergence evidence for item thresholds.



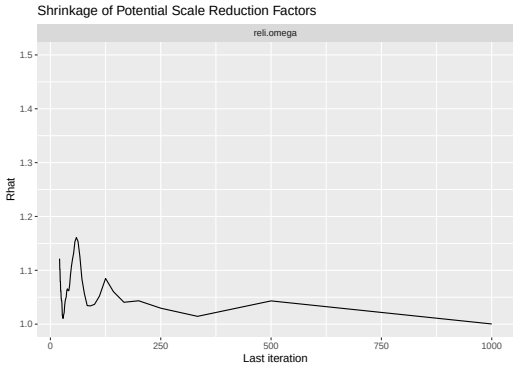
(a) Posterior density



(b) Posterior traceplot



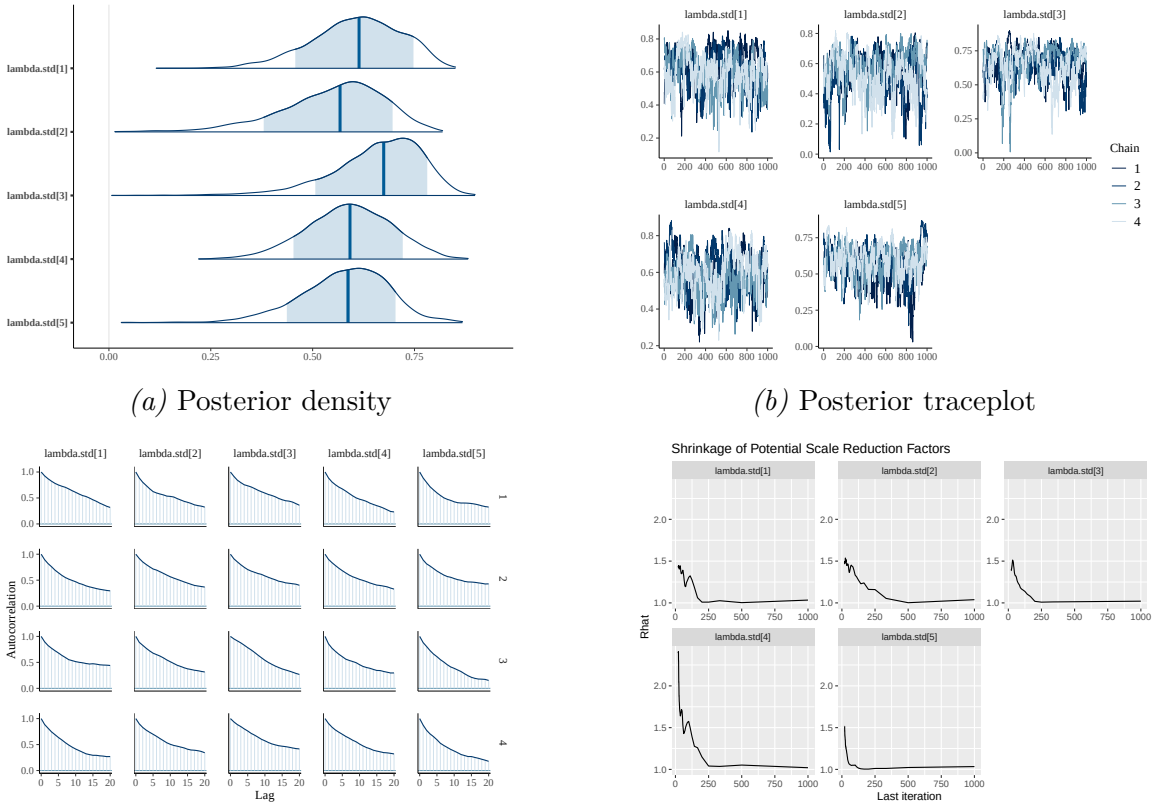
(c) Autocorrelation plot



(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.14. Study 1 model 3: Posterior convergence evidence for reliability (ω).

Parameters in Model 4



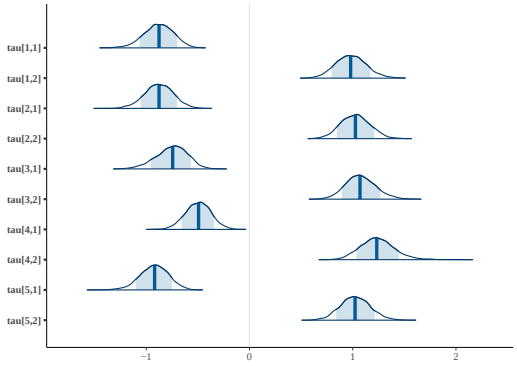
(a) Posterior density

(b) Posterior traceplot

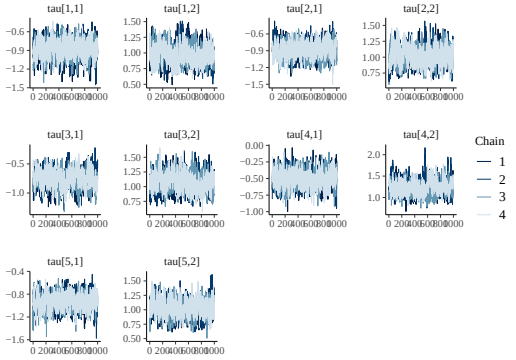
(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

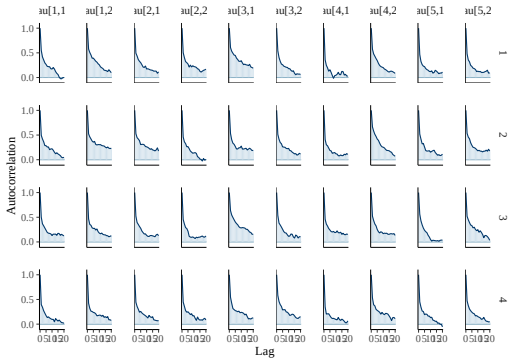
Figure B.15. Study 1 model 4: Posterior convergence evidence for factor loadings (standardized).



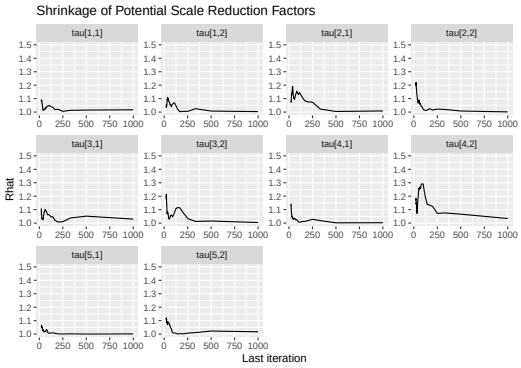
(a) Posterior density



(b) Posterior traceplot

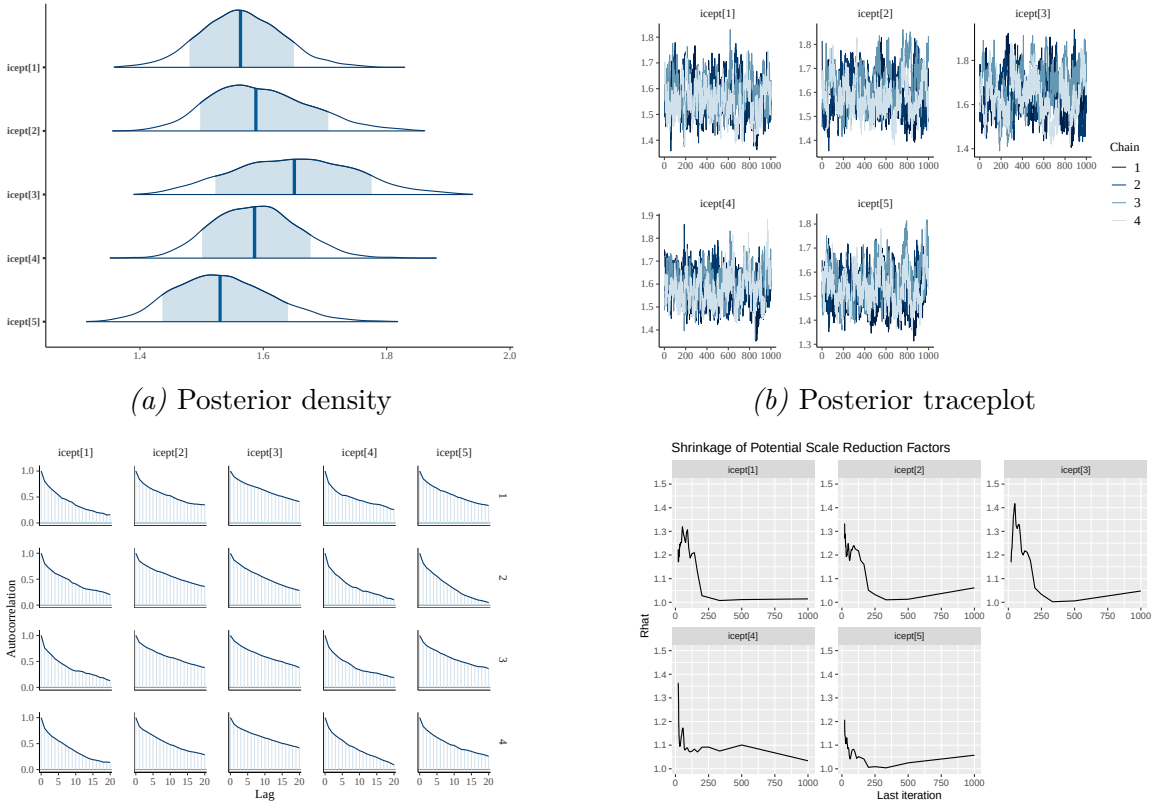


(c) Autocorrelation plot



(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.16. Study 1 model 4: Posterior convergence evidence for item thresholds.



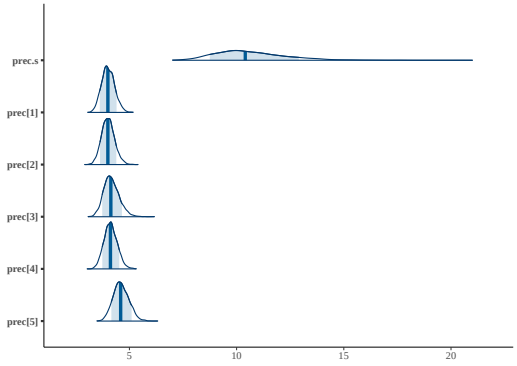
(a) Posterior density

(b) Posterior traceplot

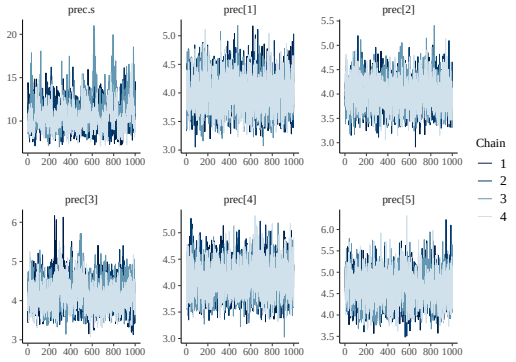
(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

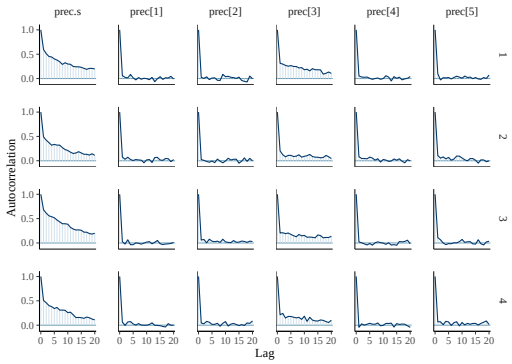
Figure B.17. Study 1 model 4: Posterior convergence evidence for response time intercepts.



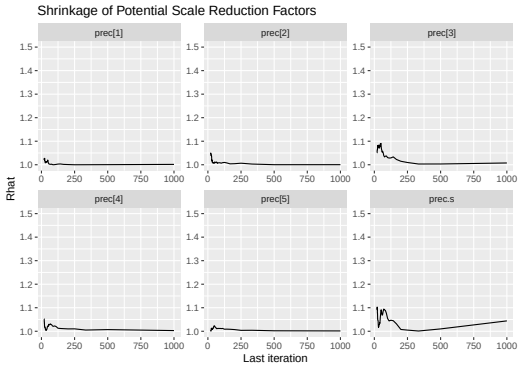
(a) Posterior density



(b) Posterior traceplot

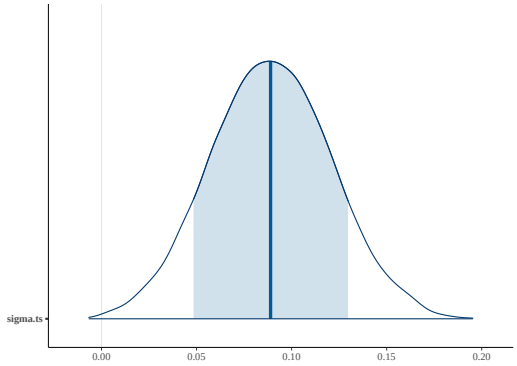


(c) Autocorrelation plot

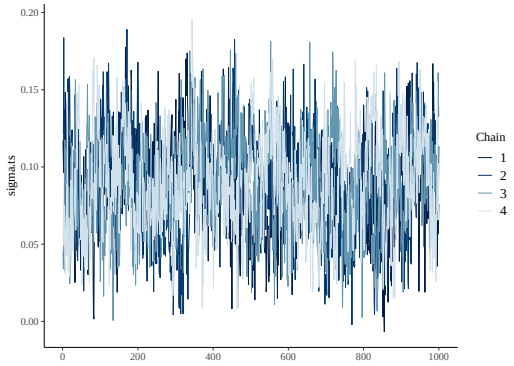


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

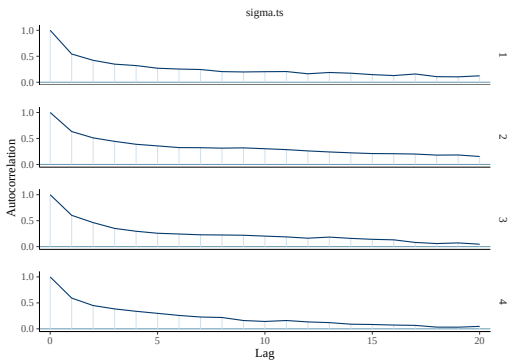
Figure B.18. Study 1 model 4: Posterior convergence evidence for response time item and factor precision.



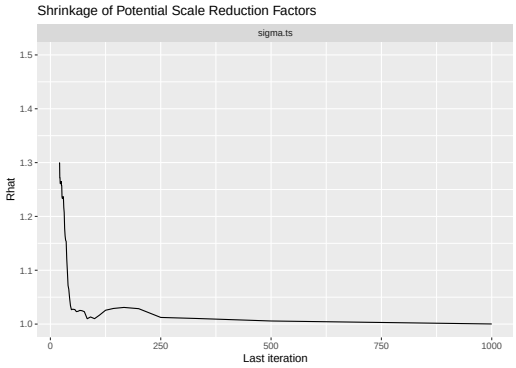
(a) Posterior density



(b) Posterior traceplot

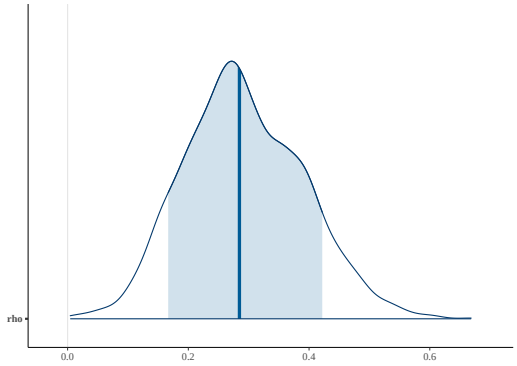


(c) Autocorrelation plot

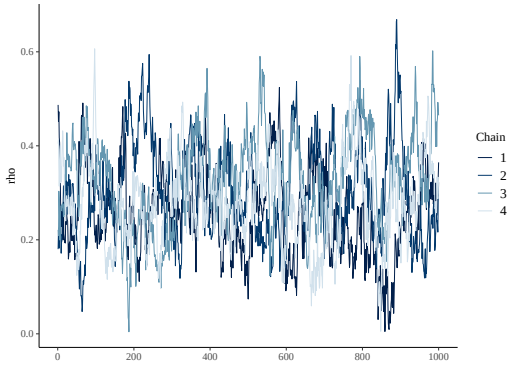


(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

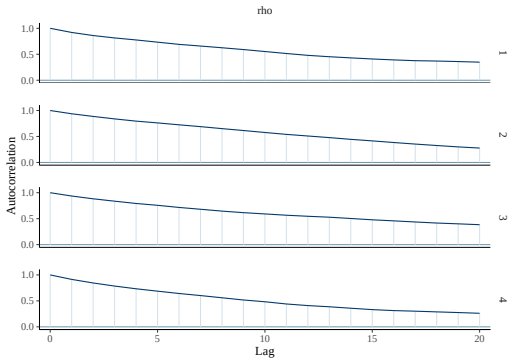
Figure B.19. Study 1 model 4: Posterior convergence evidence for factor covariance.



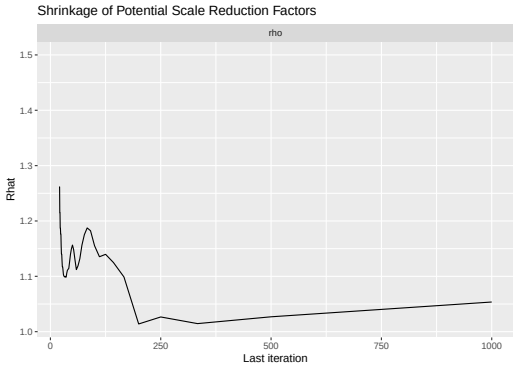
(a) Posterior density



(b) Posterior traceplot

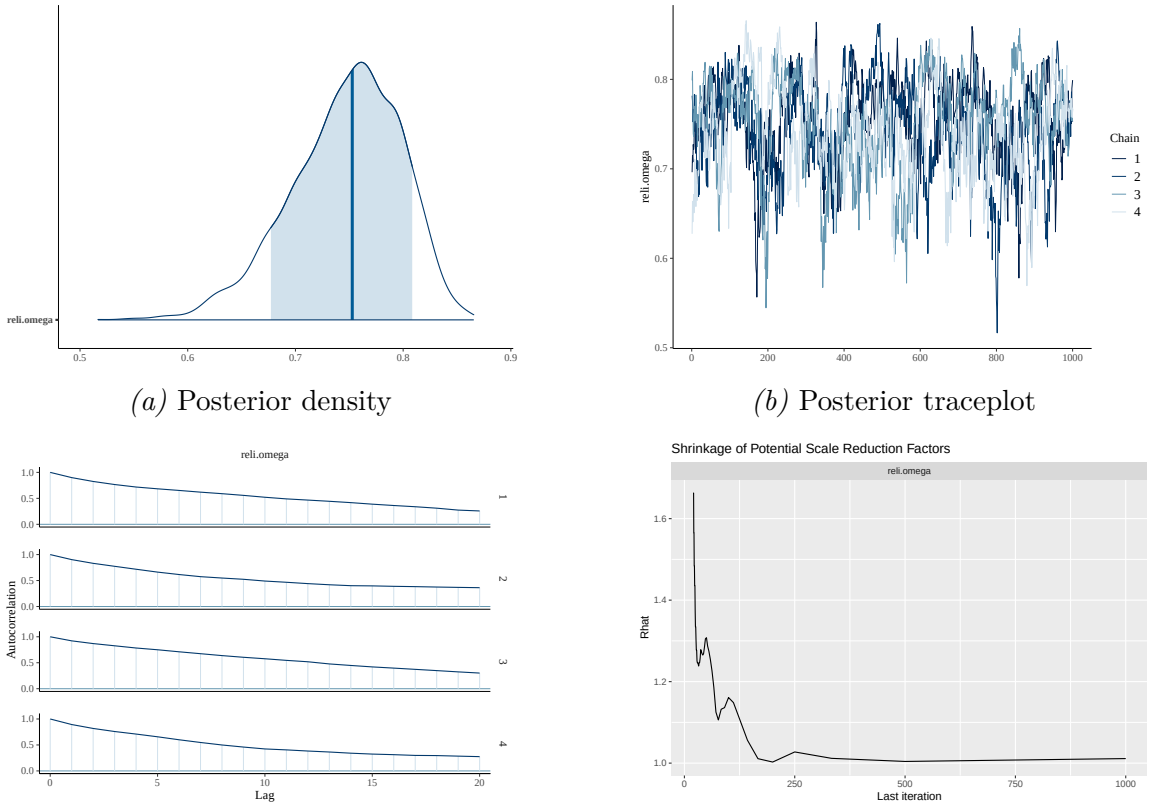


(c) Autocorrelation plot



(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.20. Study 1 model 4: Posterior convergence evidence for ρ.



(a) Posterior density

(b) Posterior traceplot

(c) Autocorrelation plot

(d) Gelem-Rubin-Brooks ($\hat{R}$ Shrinkage)

Figure B.21. Study 1 model 4: Posterior convergence evidence for reliability (ω).

Model 1: Item Factor Analysis

```
model {
  for(p in 1:N){
    for(i in 1:nit){
      y[p,i] ~ dcat(omega[p,i, ])  # data model
      ystar[p,i] ~ dnorm(lambda[i]*eta[p], 1)# Latent Response
         Variable
      pi[p,i,3] = phi(ystar[p,i] - tau[i,2]) # Pr(nu = 3)
      pi[p,i,2] = phi(ystar[p,i] - tau[i,1]) - phi(ystar[p,i] - tau[
         i,2]) # Pr(nu = 2)
      pi[p,i,1] = 1 - phi(ystar[p,i] - tau[i,1]) # Pr(nu = 1)
    }
  }
  ### Priors
  for(p in 1:N){
    eta[p] ~ dnorm(0, 1) # latent ability
  }
  for(i in 1:nit){
    tau[i, 1] ~ dnorm(0.0,0.1) # Thresholds
    tau[i, 2] ~ dnorm(0, 0.1)T(tau[i, 1],)
    lambda[i] ~ dnorm(0, .44)T(0,)# loadings
    theta[i] = 1 + pow(lambda[i],2) # LRV total variance
    lambda.std[i] = lambda[i]/pow(theta[i],0.5)# standardized
       loading
  }
  # compute omega
  lambda_sum[1] = lambda[1]
  for(i in 2:nit){
    lambda_sum[i] = lambda_sum[i-1]+lambda[i]
  }
  reli.omega = (pow(lambda_sum[nit],2))/(pow(lambda_sum[nit],2)+nit)
}
```

Model 2: Joint Item Factor Analysis with Response Time Model

```
model {
  for(p in 1:N){
    for(i in 1:nit){
      y[p,i] ~ dcat(omega[p,i, ])  # data model
      ystar[p,i] ~ dnorm(lambda[i]*eta[p], 1)# Latent Response
         Variable
      pi[p,i,3] = phi(ystar[p,i] - tau[i,2]) # Pr(nu = 3)
      pi[p,i,2] = phi(ystar[p,i] - tau[i,1]) - phi(ystar[p,i] - tau[
         i,2]) # Pr(nu = 2)
      pi[p,i,1] = 1 - phi(ystar[p,i] - tau[i,1]) # Pr(nu = 1)
      dev[p,i]<-lambda[i]*(eta[p] - (tau[i,1]+tau[i,2])/2)
      mu.lrt[p,i] <- icept[i] - speed[p] - rho * abs(dev[p,i])
      lrt[p,i] ~ dnorm(mu.lrt[p,i], prec[i])
    }
  }
  ### Priors
  for(p in 1:N){
    eta[p] ~ dnorm(0, 1) # latent ability
    speed[p]~dnorm(sigma.ts*eta[p],prec.s)  # latent speed
  }
  sigma.ts ~ dnorm(0, 0.1)
  prec.s ~ dgamma(.1,.1)
  rho~dnorm(0,.1)I(0,)
  sigma.t = pow(prec.s, -1) + pow(sigma.ts, 2) # speed variance
  cor.ts = sigma.ts/(pow(sigma.t,0.5)) # LV correlation
  for(i in 1:nit){
    icept[i]~dnorm(0,.1) # lrt parameters
    prec[i]~dgamma(.1,.1)
    tau[i, 1] ~ dnorm(0.0,0.1) # Thresholds
    tau[i, 2] ~ dnorm(0, 0.1)T(tau[i, 1],)
    lambda[i] ~ dnorm(0, .44)T(0,)# loadings
    theta[i] = 1 + pow(lambda[i],2) # LRV total variance
    lambda.std[i] = lambda[i]/pow(theta[i],0.5)# standardized
       loading
  }
  # compute omega
  lambda_sum[1] = lambda[1]
  for(i in 2:nit){
    lambda_sum[i] = lambda_sum[i-1]+lambda[i]
  }
  reli.omega = (pow(lambda_sum[nit],2))/(pow(lambda_sum[nit],2)+nit)
}
```

Model 3: Item Factor Analysis with Misclassification Informed Directly by RT

```
model {
  for(p in 1:N){
    for(i in 1:nit){
      y[p,i] ~ dcat(omega[p,i, ])  # data model
      ystar[p,i] ~ dnorm(lambda[i]*eta[p], 1)# Latent Response
         Variable
      pi[p,i,3] = phi(ystar[p,i] - tau[i,2]) # Pr(nu = 3)
      pi[p,i,2] = phi(ystar[p,i] - tau[i,1]) - phi(ystar[p,i] - tau[
         i,2]) # Pr(nu = 2)
      pi[p,i,1] = 1 - phi(ystar[p,i] - tau[i,1]) # Pr(nu = 1)
      # MISCLASSIFICATION MODEL
      for(c in 1:ncat){
        # generate informative prior for misclassificaiton
        for(ct in 1:ncat){
          alpha[p,i,ct,c] <- ifelse(c == ct,
                                    ilogit(lrt[p,i]),
                                    (1/(ncat-1))*(1-ilogit(lrt[p,i])
                                       )
        )
        }
        # sample misclassification parameters using the informative
           priors
        gamma[p,i,c,1:ncat] ~ ddirch(alpha[p,i,c,1:ncat])
        # observed category prob (Pr(y=c))
        omega[p,i, c] = gamma[p,i,c,1]*pi[p,i,1] +
          gamma[p,i,c,2]*pi[p,i,2] +
          gamma[p,i,c,3]*pi[p,i,3]
      }
    }
  }
  ### Priors
  for(p in 1:N){
    eta[p] ~ dnorm(0, 1) # latent ability
  }
  for(i in 1:nit){
    tau[i, 1] ~ dnorm(0.0,0.1) # Thresholds
    tau[i, 2] ~ dnorm(0, 0.1)T(tau[i, 1],)
    lambda[i] ~ dnorm(0, .44)T(0,)# loadings
    theta[i] = 1 + pow(lambda[i],2) # LRV total variance
    lambda.std[i] = lambda[i]/pow(theta[i],0.5)# standardized
       loading
  }
  lambda_sum[1] = lambda[1]
  for(i in 2:nit){
    lambda_sum[i] = lambda_sum[i-1]+lambda[i]
  }
  reli.omega = (pow(lambda_sum[nit],2))/(pow(lambda_sum[nit],2)+nit)
}
```

Model 4: Joint Model with Misclassification Informed by ELRT

```
model {
  for(p in 1:N){
    for(i in 1:nit){
      y[p,i] ~ dcat(omega[p,i, ])  # data model
      ystar[p,i] ~ dnorm(lambda[i]*eta[p], 1)# Latent Response
      pi[p,i,3] = phi(ystar[p,i]-tau[i,2]) # Pr(nu = 3)
      pi[p,i,2] = phi(ystar[p,i]-tau[i,1])-phi(ystar[p,i]-tau[i,2])
      pi[p,i,1] = 1-phi(ystar[p,i]-tau[i,1]) # Pr(nu = 1)
      dev[p,i]<-lambda[i]*(eta[p] - (tau[i,1]+tau[i,2])/2)
      mu.lrt[p,i] <- icept[i] - speed[p] - rho * abs(dev[p,i])
      lrt[p,i] ~ dnorm(mu.lrt[p,i], prec[i])
      elrt[p,i] <- (icept[i] - speed[p])/(rho*abs(dev[p,i])) # ELRT
      # MISCLASSIFICATION MODEL
      for(c in 1:ncat){
        # generate informative prior for misclassificaiton
        for(ct in 1:ncat){
          alpha[p,i,ct,c] <- ifelse(c == ct,  ilogit(exp(elrt[p,i]))
              , (1/(ncat-1))*(1-ilogit(exp(elrt[p,i]))))
        }
      gamma[p,i,c,1:ncat] ~ ddirch(alpha[p,i,c,1:ncat])
      # observed category prob (Pr(y=c))
      omega[p,i,c] = gamma[p,i,c,1]*pi[p,i,1] +
       gamma[p,i,c,2]*pi[p,i,2] + gamma[p,i,c,3]*pi[p,i,3]
      }
    }
  }
  for(p in 1:N){
    eta[p] ~ dnorm(0, 1) # latent ability
    speed[p]~dnorm(sigma.ts*eta[p],prec.s)  # latent speed
  }
  sigma.ts ~ dnorm(0, 0.1)
  prec.s ~ dgamma(.1,.1)
  rho~dnorm(0,.1)I(0,)
  sigma.t = pow(prec.s, -1) + pow(sigma.ts, 2) # speed variance
  for(i in 1:nit){
    icept[i]~dnorm(0,.1) # lrt parameters
    prec[i]~dgamma(.1,.1)
    tau[i, 1] ~ dnorm(0.0,0.1) # Thresholds
    tau[i, 2] ~ dnorm(0, 0.1)T(tau[i, 1],)
    lambda[i] ~ dnorm(0, .44)T(0,)# loadings
    theta[i] = 1 + pow(lambda[i],2) # LRV total variance
    lambda.std[i] = lambda[i]/pow(theta[i],0.5)# standardize
  }
  lambda_sum[1] = lambda[1]
  for(i in 2:nit){
    lambda_sum[i] = lambda_sum[i-1]+lambda[i]
  }
  reli.omega = (pow(lambda_sum[nit],2))/(pow(lambda_sum[nit],2)+nit)
}
```

APPENDIX C

Simulation Study 2

Results when Indicators are Trichotomous (Three Ordered Categories)

Table C.1

Posterior convergence by $\hat{R}$ of three category items

		Avg $\hat{R}$			SD $\hat{R}$			Prop. Converge		
Parameter	N	5	10	20	5	10	20	5	10	20
λ	500	1.06	1.03	1.02	0.03	0.01	0.00	87	98	99
	2500	1.04	1.02	1.02	0.01	0.01	0.00	95	99	100
τ	500	1.02	1.01	1.01	0.01	0.00	0.00	98	100	100
	2500	1.01	1.01	1.01	0.00	0.00	0.00	100	100	100
θ	500	1.07	1.03	1.02	0.04	0.01	0.00	80	98	99
	2500	1.04	1.02	1.02	0.02	0.01	0.00	93	99	100
β_{lrt}	500	1.03	1.02	1.01	0.02	0.01	0.01	96	100	100
	2500	1.03	1.02	1.01	0.02	0.01	0.01	99	100	100
σ_{lrt}	500	1.00	1.00	1.00	0.00	0.00	0.00	100	100	100
	2500	1.00	1.00	1.00	0.00	0.00	0.00	100	100	100
σ_s	500	1.01	1.00	1.00	0.01	0.01	0.00	100	100	100
	2500	1.01	1.00	1.00	0.01	0.01	0.00	100	100	100
σ_{st}	500	1.01	1.01	1.01	0.01	0.01	0.01	100	100	100
	2500	1.01	1.01	1.00	0.01	0.01	0.01	100	100	100
ρ	500	1.03	1.03	1.02	0.04	0.02	0.02	97	99	99
	2500	1.03	1.02	1.02	0.02	0.02	0.01	98	100	100
ω	500	1.05	1.02	1.02	0.04	0.01	0.01	89	100	100
	2500	1.02	1.02	1.01	0.02	0.01	0.02	100	100	99

Note. Number of items 5, 10, 20 are represented along the columns of this table. Avg $\hat{R}$ = Average $\hat{R}$ across replications; SD $\hat{R}$ = standard deviation of $\hat{R}$ estimates across replications; % Converge = percent of replications with $\hat{R} < 1.10$.

Table C.2

Posterior bias of three category items

		Average Relative Bias			Average Bias		
Parameter	N	5	10	20	5	10	20
λ	500	-14.18	-13.31	-12.47	-0.13	-0.12	-0.11
	2500	-15.38	-15.69	-15.88	-0.14	-0.14	-0.14
τ	500	58.04	-50.96	-67.29	0.12	0.11	0.11
	2500	54.39	-48.43	-52.75	0.12	0.11	0.11
θ	500	-7.28	-8.71	-8.87	-0.13	-0.16	-0.16
	2500	-11.89	-12.49	-12.77	-0.22	-0.23	-0.23
β_{lrt}	500	0.35	0.05	-0.78	0.01	0.00	-0.01
	2500	1.43	0.75	0.06	0.02	0.01	0.00
σ_{lrt}	500	0.96	0.27	0.08	0.04	0.01	0.00
	2500	0.51	0.21	0.12	0.02	0.01	0.00
σ_s	500	8.40	6.76	6.08	0.84	0.68	0.61
	2500	8.63	8.05	6.18	0.86	0.81	0.62
σ_{st}	500	-12.05	0.22	9.01	-0.01	0.00	0.01
	2500	0.55	0.38	2.51	0.00	0.00	0.00
ρ	500	36.35	26.41	8.04	0.04	0.03	0.01
rho	2500	63.92	51.16	36.93	0.06	0.05	0.04
ω	500	-3.69	-3.16	-1.67	-0.03	-0.03	-0.02
	2500	-6.82	-4.21	-2.35	-0.05	-0.04	-0.02

Results when Indicators are Polytomous (Seven Ordered Categories)

Table C.3

Posterior convergence by $\hat{R}$ of seven category items

		Avg $\hat{R}$			SD $\hat{R}$			Prop. Converge		
Parameter	N	5	10	20	5	10	20	5	10	20
λ	500	1.07	1.05	1.05	0.03	0.02	0.01	78	85	88
	2500	1.06	1.05	1.04	0.02	0.01	0.01	83	90	92
τ	500	1.03	1.02	1.02	0.01	0.00	0.00	95	99	99
	2500	1.02	1.02	1.02	0.01	0.00	0.00	99	100	100
θ	500	1.08	1.06	1.05	0.03	0.02	0.01	75	84	86
	2500	1.06	1.05	1.04	0.03	0.01	0.01	82	90	92
β_{lrt}	500	1.03	1.02	1.01	0.02	0.01	0.01	96	99	100
	2500	1.03	1.02	1.01	0.02	0.01	0.01	95	99	100
σ_{lrt}	500	1.00	1.00	1.00	0.00	0.00	0.00	100	100	100
	2500	1.00	1.00	1.00	0.00	0.00	0.00	100	100	100
σ_s	500	1.01	1.00	1.00	0.01	0.01	0.00	100	100	100
	2500	1.01	1.00	1.00	0.01	0.01	0.00	100	100	100
σ_{st}	500	1.02	1.01	1.01	0.02	0.01	0.01	100	100	100
	2500	1.02	1.01	1.01	0.01	0.01	0.01	100	100	100
ρ	500	1.04	1.03	1.02	0.04	0.02	0.01	87	98	100
	2500	1.04	1.02	1.02	0.03	0.02	0.01	90	98	100
ω	500	1.07	1.04	1.03	0.06	0.03	0.03	81	92	98
	2500	1.04	1.03	1.02	0.04	0.02	0.02	86	98	100

Note. Number of items 5, 10, 20 are represented along the columns of this table. Avg $\hat{R}$ = Average $\hat{R}$ across replications; SD $\hat{R}$ = standard deviation of $\hat{R}$ estimates across replications; % Converge = percent of replications with $\hat{R} < 1.10$.

Table C.4

Posterior bias of seven category items

		Average Relative Bias			Average Bias		
Parameter	N	5	10	20	5	10	20
λ	500	-14.66	-13.48	-12.76	-0.13	-0.12	-0.11
	2500	-14.93	-14.61	-14.40	-0.13	-0.13	-0.13
τ	500	-3.81	-6.67	-22.63	0.00	-0.00	0.01
	2500	-8.11	-9.11	-16.54	-0.02	-0.01	-0.01
θ	500	-9.39	-9.83	-9.50	-0.17	-0.18	-0.17
	2500	-11.81	-11.84	-11.76	-0.21	-0.21	-0.21
β_{lrt}	500	0.53	-0.30	-0.82	0.01	-0.00	-0.01
	2500	1.42	0.57	0.39	0.02	0.01	0.01
σ_{lrt}	500	1.00	0.33	0.31	0.04	0.01	0.01
	2500	0.15	0.08	0.13	0.01	0.00	0.01
σ_s	500	10.07	7.25	6.22	1.01	0.73	0.62
	2500	8.09	6.86	5.71	0.81	0.69	0.57
σ_{st}	500	-6.78	-6.13	-4.97	-0.00	-0.00	-0.00
	2500	-1.90	-3.69	-1.87	-0.00	-0.00	-0.00
ρ	500	35.60	21.37	10.46	0.04	0.02	0.01
	2500	65.98	46.40	40.98	0.07	0.05	0.04
ω	500	-4.44	-3.17	-1.69	-0.04	-0.03	-0.02
	2500	-6.57	-3.85	-2.08	-0.05	-0.03	-0.02

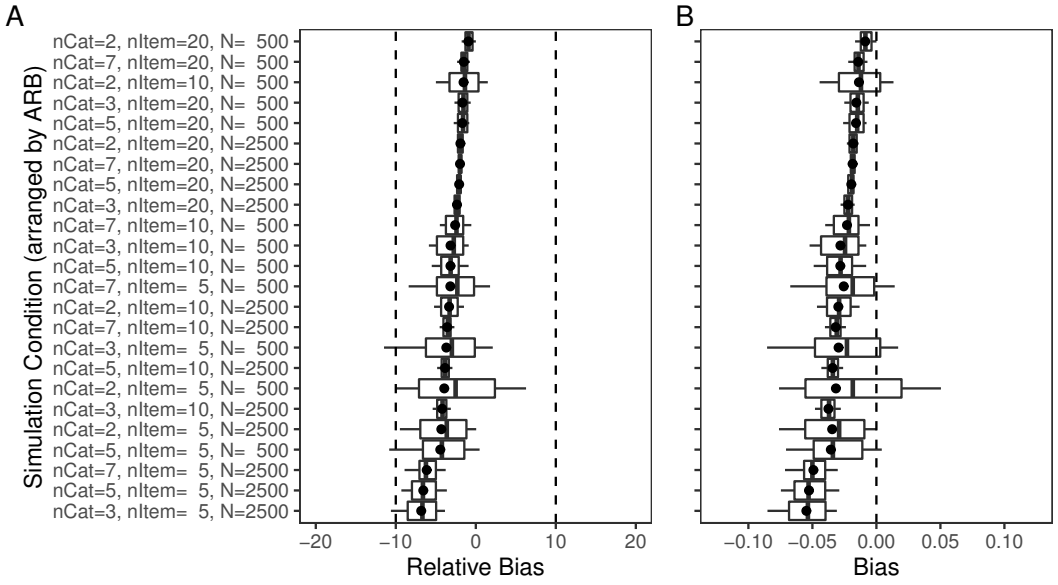


Figure C.1. Bias of posterior median estimate of factor reliability. (A) relative bias estimates where dashed lines represent ±10% relative bias, (B) average bias of posterior median. ARB = average relative bias.

Table C.5

Credible interval coverage rate

		Dichotomous			3-Cat			5-Cat			7-Cat		
	N	5	10	20	5	10	20	5	10	20	5	10	20
λ	500	88	91	92	89	86	84	87	88	87	88	88	84
	2500	83	85	78	68	56	43	74	61	49	70	57	44
τ	500	93	95	95	92	93	92	78	81	81	90	91	88
	2500	96	95	95	79	80	78	52	50	52	71	69	66
θ	500	88	90	92	89	86	84	87	89	88	88	88	84
	2500	83	84	79	69	57	44	75	62	50	69	58	46
β_{lrt}	500	98	95	93	97	96	93	97	97	96	96	97	96
	2500	94	93	97	90	96	97	93	94	97	91	97	98
σ_{lrt}	500	96	95	95	95	93	95	98	96	94	96	96	95
	2500	96	94	95	94	94	96	95	96	96	96	96	95
σ_s	500	87	95	89	90	91	88	92	92	90	95	90	90
	2500	75	64	74	74	69	72	72	54	69	80	72	71
σ_{st}	500	97	93	96	97	90	94	94	94	93	96	92	98
	2500	95	98	94	93	96	96	90	91	92	92	98	89
ρ	500	100	95	98	92	95	93	93	93	97	97	92	98
	2500	86	56	37	62	51	25	60	49	50	53	47	45
ω	500	94	97	92	89	46	33	92	66	45	88	59	33
	2500	79	24	1	5	0	0	16	0	0	5	0	0

Table C.6

Summary of credible interval width across conditions and replications

		Avg. Width						Width SD					
		3-Cat			7-Cat			3-Cat			7-Cat		
	N	5	10	20	5	10	20	5	10	20	5	10	20
λ	500	1.07	0.79	0.66	0.81	0.60	0.51	0.32	0.14	0.08	0.20	0.07	0.04
	2500	0.45	0.34	0.28	0.34	0.26	0.22	0.06	0.03	0.02	0.03	0.02	0.01
τ	500	0.54	0.50	0.49	0.56	0.52	0.51	0.08	0.04	0.03	0.12	0.07	0.06
	2500	0.23	0.22	0.21	0.24	0.23	0.22	0.02	0.01	0.01	0.03	0.02	0.02
θ	500	2.04	1.36	1.10	1.47	0.99	0.83	1.39	0.62	0.38	0.87	0.30	0.21
	2500	0.71	0.52	0.42	0.53	0.41	0.34	0.21	0.10	0.06	0.10	0.06	0.04
β_{lrt}	500	0.23	0.18	0.15	0.20	0.16	0.13	0.05	0.03	0.02	0.04	0.02	0.01
	2500	0.12	0.09	0.07	0.10	0.08	0.06	0.02	0.01	0.01	0.01	0.01	0.00
σ_{lrt}	500	1.23	1.10	1.04	1.17	1.09	1.04	0.18	0.09	0.08	0.12	0.08	0.07
	2500	0.53	0.49	0.46	0.52	0.48	0.46	0.03	0.02	0.02	0.02	0.02	0.02
σ_s	500	5.50	3.88	3.23	4.98	3.59	3.15	1.59	0.59	0.31	1.04	0.48	0.25
	2500	2.51	1.78	1.46	2.19	1.62	1.38	0.33	0.14	0.08	0.20	0.10	0.06
σ_{st}	500	0.13	0.09	0.08	0.11	0.08	0.07	0.01	0.01	0.00	0.01	0.00	0.00
	2500	0.06	0.04	0.03	0.05	0.04	0.03	0.00	0.00	0.00	0.00	0.00	0.00
ρ	500	0.26	0.18	0.13	0.22	0.16	0.11	0.06	0.03	0.02	0.06	0.02	0.01
	2500	0.16	0.10	0.07	0.13	0.09	0.06	0.02	0.01	0.00	0.02	0.01	0.00
ω	500	0.18	0.07	0.03	0.13	0.06	0.02	0.05	0.01	0.00	0.03	0.01	0.00
	2500	0.08	0.03	0.01	0.06	0.03	0.01	0.01	0.00	0.00	0.01	0.00	0.00

APPENDIX D

NAEP Math Identity Analysis and Posterior Investigation

Posterior Summaries

Table D.1

NAEP mathematics identity model 1 posterior distribution summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.94	0.01	0.92	0.93	0.94	0.94	0.95	1.08	43
λ_2	0.93	0.01	0.91	0.92	0.93	0.93	0.94	1.02	160
λ_3	0.91	0.01	0.89	0.90	0.91	0.91	0.92	1.01	230
λ_4	0.66	0.03	0.60	0.64	0.66	0.68	0.71	1.01	590
λ_5	0.68	0.03	0.62	0.66	0.68	0.70	0.73	1.00	780
$\tau_{1,1}$	-3.50	0.19	-3.90	-3.60	-3.50	-3.40	-3.20	1.00	99
$\tau_{2,1}$	-2.40	0.14	-2.70	-2.50	-2.40	-2.30	-2.10	1.00	130
$\tau_{3,1}$	-1.90	0.11	-2.10	-1.90	-1.90	-1.80	-1.70	1.00	300
$\tau_{4,1}$	-3.00	0.13	-3.20	-3.10	-3.00	-2.90	-2.70	1.00	1500
$\tau_{5,1}$	-3.50	0.16	-3.90	-3.60	-3.50	-3.40	-3.20	1.00	4000
$\tau_{1,2}$	-2.01	0.14	-2.20	-2.10	-2.00	-1.90	-1.70	1.00	170
$\tau_{2,2}$	-0.84	0.11	-1.10	-0.91	-0.84	-0.76	-0.64	1.00	210
$\tau_{3,2}$	-0.65	0.10	-0.85	-0.71	-0.64	-0.58	-0.46	1.00	220
$\tau_{4,2}$	-1.91	0.08	-2.10	-2.00	-1.90	-1.90	-1.80	1.00	1200
$\tau_{5,2}$	-2.62	0.11	-2.80	-2.70	-2.60	-2.50	-2.40	1.00	1600
$\tau_{1,3}$	0.21	0.11	-0.01	0.14	0.21	0.29	0.44	1.00	560
$\tau_{2,3}$	1.14	0.11	0.87	1.00	1.10	1.10	1.30	1.00	1100
$\tau_{3,3}$	0.72	0.10	0.52	0.65	0.72	0.78	0.91	1.00	270
$\tau_{4,3}$	-0.71	0.07	-0.84	-0.75	-0.71	-0.66	-0.58	1.00	870
$\tau_{5,3}$	-1.53	0.08	-1.70	-1.60	-1.50	-1.50	-1.40	1.00	1100
$\tau_{1,4}$	2.42	0.17	2.10	2.30	2.40	2.50	2.70	1.00	84
$\tau_{2,4}$	3.01	0.16	2.70	2.90	3.00	3.10	3.30	1.00	410
$\tau_{3,4}$	2.00	0.12	1.80	2.00	2.00	2.10	2.30	1.00	340
$\tau_{4,4}$	0.41	0.06	0.28	0.36	0.41	0.45	0.53	1.00	1100
$\tau_{5,4}$	-0.27	0.06	-0.40	-0.32	-0.27	-0.23	-0.15	1.00	2900
ω	0.77	0.01	0.75	0.76	0.77	0.78	0.79	1.02	130

Note. Reported factor loadings are standardized.

Table D.2

NAEP mathematics identity model 2 posterior distribution summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.94	0.01	0.92	0.93	0.94	0.94	0.95	1.04	64
λ_2	0.93	0.01	0.91	0.92	0.93	0.93	0.94	1.01	280
λ_3	0.91	0.01	0.89	0.90	0.91	0.92	0.93	1.01	190
λ_4	0.66	0.03	0.60	0.64	0.66	0.68	0.71	1.01	440
λ_5	0.68	0.03	0.62	0.66	0.68	0.70	0.73	1.01	530
$\tau_{1,1}$	-3.42	0.18	-3.81	-3.54	-3.41	-3.29	-3.08	1.04	66
$\tau_{2,1}$	-2.39	0.14	-2.70	-2.48	-2.38	-2.29	-2.13	1.00	950
$\tau_{3,1}$	-1.86	0.11	-2.08	-1.94	-1.86	-1.79	-1.64	1.00	630
$\tau_{4,1}$	-2.97	0.12	-3.23	-3.06	-2.97	-2.89	-2.74	1.00	2700
$\tau_{5,1}$	-3.51	0.16	-3.85	-3.62	-3.51	-3.41	-3.21	1.00	690
$\tau_{1,2}$	-1.89	0.13	-2.16	-1.98	-1.89	-1.80	-1.64	1.03	89
$\tau_{2,2}$	-0.82	0.11	-1.04	-0.89	-0.82	-0.74	-0.61	1.00	1700
$\tau_{3,2}$	-0.62	0.09	-0.80	-0.68	-0.62	-0.56	-0.45	1.01	530
$\tau_{4,2}$	-1.91	0.08	-2.07	-1.96	-1.91	-1.85	-1.75	1.00	850
$\tau_{5,2}$	-2.56	0.11	-2.78	-2.64	-2.56	-2.49	-2.37	1.01	450
$\tau_{1,3}$	0.26	0.11	0.04	0.18	0.26	0.34	0.48	1.00	1400
$\tau_{2,3}$	1.12	0.11	0.92	1.05	1.12	1.2	1.33	1.01	230
$\tau_{3,3}$	0.76	0.09	0.58	0.70	0.76	0.83	0.95	1.01	300
$\tau_{4,3}$	-0.69	0.06	-0.82	-0.74	-0.69	-0.65	-0.57	1.01	460
$\tau_{5,3}$	-1.51	0.08	-1.66	-1.56	-1.51	-1.46	-1.36	1.00	620
$\tau_{1,4}$	2.41	0.16	2.12	2.29	2.40	2.52	2.74	1.01	190
$\tau_{2,4}$	3.09	0.16	2.77	2.98	3.09	3.20	3.40	1.01	200
$\tau_{3,4}$	2.12	0.12	1.89	2.03	2.11	2.20	2.35	1.01	280
$\tau_{4,4}$	0.42	0.06	0.30	0.38	0.42	0.46	0.55	1.00	910
$\tau_{5,4}$	-0.26	0.06	-0.39	-0.30	-0.26	-0.21	-0.13	1.00	780
$\beta_{lrt,1}$	11.08	0.39	10.33	10.82	11.07	11.33	11.9	1.00	1700
$\beta_{lrt,2}$	5.12	0.15	4.84	5.02	5.12	5.22	5.41	1.00	1200
$\beta_{lrt,3}$	4.16	0.11	3.95	4.08	4.16	4.24	4.37	1.00	1300
$\beta_{lrt,4}$	3.80	0.11	3.60	3.73	3.80	3.87	4.02	1.00	1200
$\beta_{lrt,5}$	3.30	0.12	3.07	3.22	3.29	3.37	3.54	1.00	1100
σ_s	3.16	0.17	2.85	3.05	3.16	3.28	3.49	1.00	2600
σ_{st}[a]	0.12	0.04	0.05	0.09	0.12	0.15	0.19	1.00	2500
ρ	0.03	0.01	0.01	0.02	0.03	0.03	0.04	1.00	770
ω	0.77	0.01	0.75	0.76	0.77	0.78	0.79	1.01	470

Note. Reported factor loadings are standardized. [a] Standardize to be a correlation.

Table D.3

NAEP mathematics identity model 3 posterior distribution summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.98	0.00	0.97	0.97	0.98	0.98	0.99	1.10	22
λ_2	0.97	0.00	0.96	0.97	0.97	0.97	0.98	1.10	57
λ_3	0.98	0.00	0.97	0.97	0.98	0.98	0.98	1.40	11
λ_4	0.64	0.04	0.55	0.61	0.64	0.67	0.71	1.00	100
λ_5	0.70	0.04	0.61	0.67	0.70	0.73	0.77	1.00	290
$\tau_{1,1}$	-5.50	0.51	-6.60	-5.90	-5.50	-5.20	-4.60	1.20	18
$\tau_{2,1}$	-3.60	0.30	-4.20	-3.80	-3.60	-3.40	-3.10	1.00	150
$\tau_{3,1}$	-3.50	0.42	-4.50	-3.80	-3.50	-3.20	-2.80	1.20	16
$\tau_{4,1}$	-3.80	0.44	-4.80	-4.00	-3.70	-3.50	-3.20	1.00	950
$\tau_{5,1}$	-5.90	1.30	-9.30	-6.50	-5.60	-5.00	-4.30	1.00	1300
$\tau_{1,2}$	-3.10	0.32	-3.70	-3.30	-3.00	-2.80	-2.50	1.20	21
$\tau_{2,2}$	-1.20	0.20	-1.60	-1.30	-1.20	-1.10	-0.83	1.00	500
$\tau_{3,2}$	-1.10	0.23	-1.60	-1.20	-1.10	-0.94	-0.66	1.10	39
$\tau_{4,2}$	-2.40	0.15	-2.70	-2.50	-2.40	-2.30	-2.10	1.00	1800
$\tau_{5,2}$	-3.80	0.37	-4.60	-4.00	-3.70	-3.50	-3.20	1.00	870
$\tau_{1,3}$	0.50	0.20	0.13	0.36	0.49	0.62	0.92	1.00	290
$\tau_{2,3}$	1.60	0.20	1.30	1.50	1.60	1.70	2.10	1.00	78
$\tau_{3,3}$	1.30	0.23	0.91	1.20	1.30	1.50	1.80	1.10	41
$\tau_{4,3}$	-1.00	0.09	-1.20	-1.10	-1.00	-0.97	-0.85	1.00	2900
$\tau_{5,3}$	-2.40	0.16	-2.80	-2.50	-2.40	-2.30	-2.10	1.00	980
$\tau_{1,4}$	4.00	0.44	3.30	3.70	4.00	4.30	5.00	1.10	26
$\tau_{2,4}$	4.50	0.33	3.90	4.30	4.50	4.70	5.20	1.10	61
$\tau_{3,4}$	3.60	0.38	2.90	3.30	3.60	3.90	4.40	1.30	14
$\tau_{4,4}$	0.19	0.08	0.03	0.13	0.19	0.25	0.36	1.00	3600
$\tau_{5,4}$	-0.69	0.09	-0.86	-0.74	-0.69	-0.63	-0.51	1.00	1700
ω	0.90	0.01	0.88	0.89	0.90	0.91	0.92	1.10	31

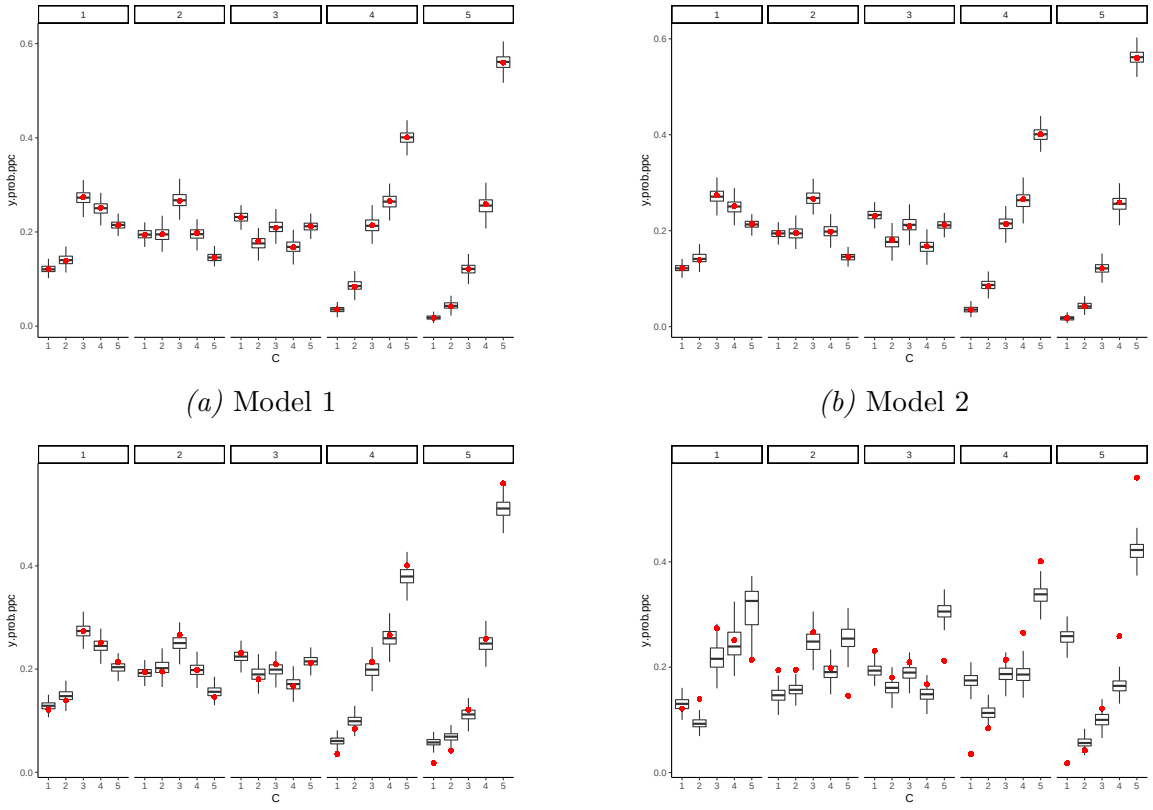
Note. Reported factor loadings are standardized.

Table D.4

NAEP mathematics identity model 4 posterior distribution summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.99	0.01	0.96	0.99	0.99	0.99	0.99	1.30	19
λ_2	0.97	0.01	0.95	0.96	0.97	0.97	0.98	1.10	29
λ_3	0.96	0.01	0.94	0.95	0.96	0.96	0.97	1.10	51
λ_4	0.95	0.01	0.93	0.94	0.95	0.96	0.96	1.10	50
λ_5	0.94	0.01	0.91	0.92	0.94	0.95	0.95	1.10	27
$\tau_{1,1}$	-2.10	0.58	-3.40	-2.60	-1.90	-1.70	-1.30	2.90	5
$\tau_{2,1}$	-1.70	0.13	-2.00	-1.80	-1.70	-1.60	-1.50	3.40	5
$\tau_{3,1}$	-0.98	0.12	-1.30	-1.00	-0.97	-0.90	-0.77	1.40	12
$\tau_{4,1}$	-0.59	0.11	-0.82	-0.64	-0.56	-0.51	-0.42	3.10	5
$\tau_{5,1}$	0.34	0.09	0.13	0.28	0.35	0.39	0.50	1.40	14
$\tau_{1,2}$	-0.93	0.44	-1.70	-1.20	-0.85	-0.59	-0.34	6.40	4
$\tau_{2,2}$	-0.35	0.17	-0.68	-0.47	-0.38	-0.21	-0.05	1.60	8
$\tau_{3,2}$	0.11	0.11	-0.11	0.04	0.13	0.19	0.32	1.30	12
$\tau_{4,2}$	0.13	0.10	-0.07	0.07	0.14	0.22	0.3	1.20	24
$\tau_{5,2}$	0.49	0.06	0.38	0.45	0.49	0.53	0.62	1.50	9
$\tau_{1,3}$	1.40	0.39	0.69	1.10	1.30	1.70	2.10	1.20	17
$\tau_{2,3}$	1.50	0.21	1.00	1.30	1.50	1.60	1.80	1.20	26
$\tau_{3,3}$	1.50	0.09	1.30	1.50	1.50	1.60	1.60	1.10	150
$\tau_{4,3}$	1.30	0.11	1.10	1.20	1.40	1.40	1.50	1.70	7
$\tau_{5,3}$	0.90	0.06	0.79	0.86	0.90	0.95	1.00	1.10	55
$\tau_{1,4}$	5.10	1.10	3.70	4.20	4.90	5.60	7.00	7.00	4
$\tau_{2,4}$	4.00	0.32	3.50	3.80	4.00	4.30	4.60	2.80	5
$\tau_{3,4}$	2.80	0.15	2.60	2.70	2.80	2.90	3.20	2.10	6
$\tau_{4,4}$	2.50	0.10	2.30	2.50	2.60	2.60	2.70	1.60	8
$\tau_{5,4}$	1.70	0.10	1.50	1.60	1.70	1.70	1.80	1.10	31
$\beta_{lrt,1}$	8.30	1.7 0	4.00	7.30	8.80	9.60	10.00	1.30	16
$\beta_{lrt,2}$	4.50	0.30	3.90	4.20	4.60	4.70	4.90	1.00	77
$\beta_{lrt,3}$	3.80	0.25	3.30	3.60	3.80	4.00	4.10	1.10	53
$\beta_{lrt,4}$	3.40	0.26	2.90	3.20	3.50	3.70	3.80	1.00	260
$\beta_{lrt,5}$	3.00	0.21	2.60	2.80	3.00	3.10	3.20	1.00	210
σ_s	5.40	0.88	4.30	4.800	5.10	5.80	7.60	1.60	9
ρ^{a}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	120
σ_{st}	0.47	0.09	0.34	0.40	0.44	0.53	0.67	2.81	5
ω	0.94	0.02	0.90	0.93	0.95	0.95	0.96	1.10	45

Note. Reported factor loadings are standardized. [a] rho (ρ) was nearly zero ($< .001$) across all samples. During the internship with AIR to examine these NAEP data, I did not have enough time and access to data to investigate the prior-posterior sensitivity for these models. This severely limits the generalizability of these results.



(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

Figure D.1. NAEP data analysis posterior predictive distributions.

APPENDIX E

Extroversion Inventory Analysis and Posterior Investigation

Posterior Summaries

Table E.1

Extroversion model 1 posterior distributions summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.43	0.16	0.10	0.32	0.44	0.55	0.71	1.01	430
λ_2	0.53	0.15	0.21	0.43	0.54	0.63	0.76	1.00	700
λ_3	0.87	0.08	0.68	0.84	0.89	0.92	0.96	1.02	330
λ_4	0.82	0.10	0.57	0.77	0.85	0.90	0.95	1.01	210
λ_5	0.52	0.15	0.18	0.43	0.54	0.63	0.76	1.01	590
λ_6	0.60	0.16	0.22	0.51	0.63	0.72	0.85	1.01	400
λ_7	0.68	0.16	0.26	0.60	0.71	0.79	0.90	1.01	630
λ_8	0.80	0.12	0.49	0.74	0.82	0.88	0.95	1.02	540
λ_9	0.57	0.15	0.23	0.47	0.59	0.68	0.82	1.00	710
λ_{10}	0.82	0.09	0.60	0.77	0.83	0.88	0.94	1.01	350
$\tau_{1,1}$	-0.96	0.18	-1.33	-1.07	-0.95	-0.83	-0.62	1.00	1400
$\tau_{2,1}$	-0.13	0.17	-0.46	-0.24	-0.13	-0.02	0.20	1.00	910
$\tau_{3,1}$	-2.50	0.58	-3.90	-2.82	-2.43	-2.08	-1.61	1.03	140
$\tau_{4,1}$	-3.06	0.64	-4.57	-3.40	-2.98	-2.61	-2.05	1.02	200
$\tau_{5,1}$	-0.15	0.17	-0.48	-0.26	-0.15	-0.04	0.17	1.00	4000
$\tau_{6,1}$	-2.13	0.30	-2.80	-2.31	-2.10	-1.91	-1.62	1.01	510
$\tau_{7,1}$	-2.80	0.46	-3.85	-3.06	-2.75	-2.48	-2.05	1.00	1600
$\tau_{8,1}$	-3.77	0.81	-5.77	-4.19	-3.62	-3.19	-2.55	1.00	2200
$\tau_{9,1}$	-1.48	0.23	-1.97	-1.62	-1.47	-1.33	-1.08	1.01	470
$\tau_{10,1}$	-2.32	0.46	-3.40	-2.58	-2.26	-2.00	-1.60	1.01	610
ω	0.92	0.02	0.88	0.91	0.93	0.94	0.95	1.01	370

Note. Reported factor loadings are standardized.

Table E.2

Extroversion model 2 posterior distributions summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.44	0.15	0.10	0.34	0.45	0.55	0.71	1.00	1200
λ_2	0.47	0.15	0.14	0.37	0.49	0.59	0.73	1.00	2700
λ_3	0.85	0.06	0.71	0.82	0.86	0.89	0.94	1.01	270
λ_4	0.85	0.07	0.66	0.82	0.86	0.90	0.94	1.01	230
λ_5	0.50	0.15	0.18	0.41	0.52	0.61	0.75	1.02	440
λ_6	0.57	0.15	0.23	0.48	0.59	0.68	0.80	1.00	2800
λ_7	0.72	0.11	0.45	0.67	0.74	0.80	0.88	1.01	450
λ_8	0.76	0.10	0.52	0.71	0.78	0.84	0.90	1.02	330
λ_9	0.59	0.13	0.28	0.52	0.61	0.69	0.80	1.01	690
λ_{10}	0.86	0.05	0.74	0.83	0.86	0.89	0.93	1.01	650
$\tau_{1,1}$	-0.95	0.17	-1.30	-1.06	-0.94	-0.83	-0.62	1.00	3700
$\tau_{2,1}$	-0.10	0.16	-0.41	-0.20	-0.09	0.01	0.22	1.00	3200
$\tau_{3,1}$	-2.24	0.40	-3.15	-2.46	-2.20	-1.96	-1.56	1.01	340
$\tau_{4,1}$	-3.18	0.59	-4.58	-3.52	-3.11	-2.77	-2.19	1.03	160
$\tau_{5,1}$	-0.18	0.16	-0.50	-0.29	-0.18	-0.07	0.13	1.00	1500
$\tau_{6,1}$	-2.07	0.28	-2.64	-2.24	-2.05	-1.89	-1.59	1.00	3600
$\tau_{7,1}$	-2.85	0.41	-3.77	-3.11	-2.81	-2.55	-2.13	1.02	150
$\tau_{8,1}$	-3.48	0.56	-4.75	-3.82	-3.43	-3.09	-2.57	1.01	540
$\tau_{9,1}$	-1.46	0.22	-1.91	-1.60	-1.45	-1.31	-1.06	1.00	660
$\tau_{10,1}$	-2.45	0.37	-3.25	-2.69	-2.42	-2.18	-1.80	1.01	1100
$\beta_{lrt,1}$	1.56	0.08	1.40	1.50	1.55	1.61	1.72	1.01	520
$\beta_{lrt,2}$	1.40	0.06	1.29	1.37	1.40	1.44	1.51	1.00	620
$\beta_{lrt,3}$	1.54	0.16	1.27	1.43	1.52	1.64	1.90	1.01	360
$\beta_{lrt,4}$	1.64	0.31	1.14	1.41	1.60	1.83	2.33	1.03	99
$\beta_{lrt,5}$	1.36	0.06	1.24	1.31	1.35	1.40	1.48	1.01	340
$\beta_{lrt,6}$	1.43	0.10	1.26	1.35	1.42	1.49	1.67	1.01	330
$\beta_{lrt,7}$	1.51	0.17	1.22	1.38	1.49	1.61	1.90	1.01	1200
$\beta_{lrt,8}$	1.57	0.24	1.20	1.39	1.53	1.72	2.12	1.02	160
$\beta_{lrt,9}$	1.86	0.09	1.71	1.80	1.86	1.92	2.05	1.00	830
$\beta_{lrt,10}$	1.45	0.17	1.16	1.33	1.45	1.56	1.81	1.02	130
$\sigma_{lrt,1}$	1.77	0.22	1.39	1.62	1.76	1.91	2.23	1.00	4000
$\sigma_{lrt,2}$	3.69	0.49	2.80	3.34	3.67	4.00	4.71	1.00	1000
$\sigma_{lrt,3}$	4.16	0.56	3.17	3.78	4.13	4.52	5.31	1.00	2200
$\sigma_{lrt,4}$	2.54	0.33	1.96	2.31	2.52	2.76	3.21	1.00	1700
$\sigma_{lrt,5}$	2.86	0.37	2.18	2.60	2.84	3.10	3.64	1.00	3200
$\sigma_{lrt,6}$	3.03	0.39	2.33	2.76	3.02	3.28	3.84	1.00	2500
$\sigma_{lrt,7}$	5.00	0.68	3.77	4.52	4.97	5.43	6.43	1.00	3800
$\sigma_{lrt,8}$	3.91	0.51	2.97	3.55	3.89	4.24	5.01	1.00	630
$\sigma_{lrt,9}$	2.61	0.34	2.00	2.37	2.59	2.84	3.29	1.00	4000
$\sigma_{lrt,10}$	6.81	1.01	5.04	6.10	6.75	7.45	9.01	1.00	4000
σ_s	9.69	1.61	6.93	8.58	9.56	10.65	13.27	1.00	760
σ_{ts}	0.00	0.06	-0.11	-0.03	0.00	0.04	0.11	1.02	120
rho	0.11	0.03	0.06	0.09	0.11	0.13	0.18	1.04	75
ω	0.92	0.02	0.88	0.91	0.92	0.93	0.95	1.01	200

Note. Reported factor loadings are standardized.

Table E.3

Extroversion model 3 posterior distributions summary

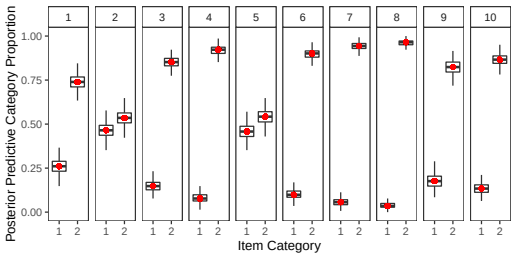
Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.77	0.19	0.24	0.70	0.83	0.91	0.97	1.06	160
λ_2	0.85	0.12	0.52	0.80	0.89	0.93	0.97	1.06	74
λ_3	0.85	0.12	0.49	0.83	0.89	0.93	0.96	1.09	200
λ_4	0.74	0.24	0.08	0.65	0.84	0.91	0.96	1.11	45
λ_5	0.75	0.17	0.32	0.67	0.79	0.88	0.95	1.06	94
λ_6	0.50	0.26	0.03	0.28	0.52	0.72	0.90	1.00	1500
λ_7	0.53	0.27	0.03	0.32	0.58	0.76	0.92	1.02	160
λ_8	0.53	0.27	0.03	0.31	0.55	0.76	0.92	1.03	130
λ_9	0.76	0.19	0.22	0.68	0.82	0.89	0.96	1.14	52
λ_{10}	0.81	0.17	0.25	0.77	0.87	0.92	0.96	1.13	65
$\tau_{1,1}$	-1.95	0.87	-4.33	-2.32	-1.75	-1.36	-0.87	1.08	66
$\tau_{2,1}$	0.03	0.44	-0.86	-0.24	0.02	0.30	0.90	1.02	200
$\tau_{3,1}$	-3.21	0.90	-5.20	-3.76	-3.11	-2.57	-1.77	1.01	590
$\tau_{4,1}$	-3.98	1.29	-7.17	-4.71	-3.74	-3.04	-2.15	1.14	27
$\tau_{5,1}$	-0.26	0.41	-1.12	-0.50	-0.24	0.00	0.50	1.05	60
$\tau_{6,1}$	-4.85	1.67	-8.79	-5.79	-4.57	-3.58	-2.48	1.02	120
$\tau_{7,1}$	-4.85	1.49	-8.32	-5.68	-4.59	-3.76	-2.70	1.02	140
$\tau_{8,1}$	-5.30	1.61	-9.22	-6.17	-5.01	-4.15	-2.98	1.01	250
$\tau_{9,1}$	-2.90	0.94	-5.14	-3.43	-2.74	-2.22	-1.51	1.04	69
$\tau_{10,1}$	-3.45	1.05	-5.89	-4.03	-3.30	-2.68	-1.86	1.02	150
ω	0.95	0.02	0.91	0.94	0.95	0.96	0.97	1.19	18

Table E.4

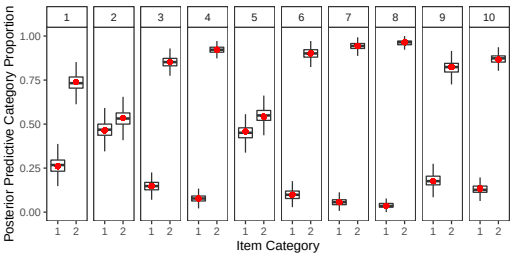
Extroversion model 4 posterior distributions summary

Parameter	Mean	SD	2.5%	25%	50%	75%	97.5%	$\hat{R}$	Neff
λ_1	0.82	0.13	0.47	0.78	0.86	0.91	0.96	1.09	63
λ_2	0.87	0.09	0.61	0.84	0.90	0.93	0.97	1.18	27
λ_3	0.90	0.06	0.74	0.88	0.92	0.94	0.97	1.08	78
λ_4	0.66	0.25	0.07	0.51	0.74	0.86	0.95	1.09	41
λ_5	0.71	0.20	0.20	0.62	0.76	0.86	0.94	1.05	84
λ_6	0.50	0.25	0.04	0.30	0.52	0.70	0.88	1.03	100
λ_7	0.44	0.24	0.02	0.24	0.45	0.63	0.84	1.00	1300
λ_8	0.47	0.25	0.03	0.27	0.49	0.67	0.87	1.02	160
λ_9	0.76	0.14	0.36	0.69	0.79	0.86	0.93	1.08	75
λ_{10}	0.85	0.10	0.58	0.82	0.87	0.91	0.95	1.15	61
$\tau_{1,1}$	-1.97	0.64	-3.45	-2.31	-1.87	-1.52	-0.99	1.01	310
$\tau_{2,1}$	-0.21	0.39	-0.97	-0.45	-0.22	0.03	0.61	1.04	67
$\tau_{3,1}$	-3.77	0.98	-5.88	-4.45	-3.67	-3.02	-2.13	1.05	66
$\tau_{4,1}$	-3.91	1.37	-7.45	-4.61	-3.58	-2.91	-2.07	1.22	18
$\tau_{5,1}$	-0.44	0.43	-1.40	-0.71	-0.41	-0.14	0.30	1.02	120
$\tau_{6,1}$	-4.67	1.59	-8.51	-5.59	-4.36	-3.44	-2.46	1.12	26
$\tau_{7,1}$	-4.70	1.45	-8.24	-5.49	-4.44	-3.64	-2.65	1.01	250
$\tau_{8,1}$	-5.22	1.56	-9.01	-6.09	-4.95	-4.08	-2.90	1.00	1100
$\tau_{9,1}$	-2.54	0.64	-4.01	-2.93	-2.45	-2.08	-1.55	1.02	180
$\tau_{10,1}$	-3.61	0.86	-5.51	-4.15	-3.52	-2.98	-2.14	1.01	350
$\beta_{lrt,1}$	1.69	0.16	1.45	1.58	1.66	1.77	2.07	1.02	180
$\beta_{lrt,2}$	1.46	0.09	1.31	1.40	1.45	1.51	1.67	1.04	75
$\beta_{lrt,3}$	1.61	0.31	1.16	1.38	1.56	1.78	2.34	1.07	43
$\beta_{lrt,4}$	1.27	0.28	0.96	1.09	1.19	1.36	2.03	1.13	30
$\beta_{lrt,5}$	1.38	0.09	1.23	1.32	1.37	1.42	1.60	1.00	710
$\beta_{lrt,6}$	1.45	0.21	1.21	1.31	1.39	1.55	2.00	1.06	51
$\beta_{lrt,7}$	1.30	0.15	1.10	1.19	1.26	1.37	1.68	1.02	210
$\beta_{lrt,8}$	1.27	0.23	1.03	1.13	1.20	1.33	1.91	1.04	180
$\beta_{lrt,9}$	1.93	0.17	1.70	1.82	1.90	2.01	2.36	1.01	430
$\beta_{lrt,10}$	1.36	0.23	1.00	1.18	1.33	1.50	1.88	1.05	67
$\sigma_{lrt,1}$	1.78	0.23	1.37	1.62	1.76	1.92	2.25	1.00	3900
$\sigma_{lrt,2}$	3.53	0.47	2.67	3.20	3.49	3.83	4.53	1.01	470
$\sigma_{lrt,3}$	4.26	0.60	3.20	3.84	4.21	4.64	5.57	1.01	240
$\sigma_{lrt,4}$	2.49	0.32	1.91	2.27	2.47	2.69	3.14	1.01	500
$\sigma_{lrt,5}$	2.86	0.37	2.18	2.60	2.84	3.09	3.62	1.00	2400
$\sigma_{lrt,6}$	3.03	0.39	2.31	2.75	3.01	3.28	3.83	1.00	4000
$\sigma_{lrt,7}$	4.92	0.65	3.74	4.47	4.88	5.33	6.33	1.00	4000
$\sigma_{lrt,8}$	3.91	0.50	2.99	3.56	3.89	4.22	4.95	1.00	4000
$\sigma_{lrt,9}$	2.57	0.32	1.99	2.35	2.56	2.78	3.25	1.00	3700
$\sigma_{lrt,10}$	6.63	0.97	4.97	5.94	6.58	7.27	8.72	1.00	730
σ_s	11.81	2.65	7.81	9.96	11.42	13.13	18.47	1.01	350
ρ	0.06	0.03	0.01	0.03	0.05	0.08	0.12	1.02	280
σ_{ts}	0.11	0.06	0.00	0.07	0.11	0.15	0.23	1.01	400
ω	0.95	0.02	0.90	0.94	0.95	0.96	0.97	1.18	22

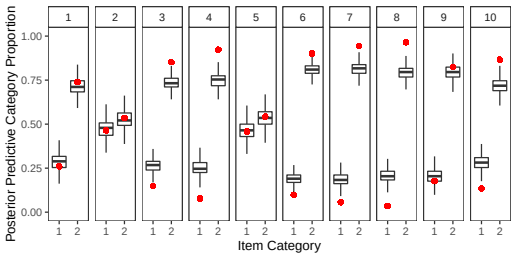
Note. Reported factor loadings are standardized.



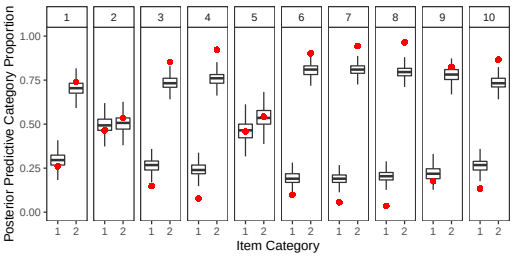
(a) Model 1



(b) Model 2



(c) Model 3



(d) Model 4

Figure E.1. Extroversion data analysis posterior predictive distributions.

Table E.5

Extroversion posterior sensitivity analysis

				Quantiles				
$\lambda \sim$	$\xi \sim$	Mean	SD	2.5%	25%	50%	75%	97.5%
$N^+(0, 0.44)$	1	0.94	0.02	0.89	0.94	0.95	0.96	0.97
	0.1	0.94	0.02	0.89	0.93	0.94	0.95	0.97
	10	0.94	0.02	0.89	0.93	0.95	0.96	0.97
	U(0.5, 1.5)	0.95	0.02	0.90	0.94	0.95	0.96	0.97
	G(1, 1)	0.95	0.02	0.90	0.94	0.95	0.96	0.98
$N^+(0, 0.01)$	1	0.99	0.00	0.98	0.98	0.99	0.99	0.99
	0.1	0.97	0.02	0.92	0.97	0.98	0.99	0.99
	10	0.99	0.01	0.97	0.99	0.99	0.99	0.99
	U(0.5, 1.5)	0.99	0.01	0.97	0.98	0.99	0.99	0.99
	G(1, 1)	0.99	0.01	0.97	0.98	0.99	0.99	0.99
$N^+(0, 1)$	1	0.90	0.03	0.83	0.89	0.91	0.92	0.95
	0.1	0.90	0.03	0.83	0.89	0.91	0.92	0.94
	10	0.91	0.03	0.85	0.90	0.92	0.93	0.96
	U(0.5, 1.5)	0.91	0.03	0.84	0.89	0.91	0.93	0.95
	G(1, 1)	0.91	0.03	0.84	0.89	0.91	0.93	0.95
$N^+(0, 10)$	1	0.51	0.10	0.29	0.45	0.52	0.59	0.69
	0.1	0.42	0.11	0.20	0.34	0.43	0.50	0.63
	10	0.52	0.10	0.30	0.46	0.53	0.60	0.70
	U(0.5, 1.5)	0.53	0.10	0.32	0.46	0.53	0.60	0.69
	G(1, 1)	0.52	0.10	0.31	0.46	0.53	0.60	0.69
$N(0, 0.44)$	1	0.91	0.07	0.70	0.89	0.93	0.95	0.97
	0.1	0.92	0.04	0.80	0.91	0.93	0.95	0.97
	10	0.92	0.03	0.84	0.91	0.93	0.94	0.96
	U(0.5, 1.5)	0.91	0.05	0.80	0.89	0.92	0.94	0.96
	G(1, 1)	0.92	0.03	0.84	0.91	0.93	0.95	0.97
$N(0, 0.01)$	1	0.98	0.01	0.95	0.98	0.99	0.99	0.99
	0.1	0.79	0.33	0.01	0.82	0.97	0.98	0.99
	10	0.98	0.01	0.94	0.97	0.98	0.99	0.99
	U(0.5, 1.5)	0.98	0.02	0.93	0.97	0.98	0.99	0.99
	G(1, 1)	0.98	0.01	0.96	0.98	0.98	0.99	0.99
$N(0, 1)$	1	0.86	0.08	0.62	0.85	0.88	0.91	0.94
	0.1	0.45	0.27	0.00	0.21	0.49	0.69	0.84
	10	0.87	0.06	0.72	0.85	0.89	0.91	0.94
	U(0.5, 1.5)	0.85	0.07	0.67	0.82	0.87	0.90	0.94
	G(1, 1)	0.87	0.06	0.73	0.84	0.88	0.91	0.94
$N(0, 10)$	1	0.25	0.18	0.00	0.10	0.24	0.39	0.60
	0.1	0.12	0.13	0.00	0.02	0.08	0.19	0.45
	10	0.19	0.17	0.00	0.04	0.14	0.31	0.54
	U(0.5, 1.5)	0.21	0.17	0.00	0.05	0.18	0.34	0.57
	G(1, 1)	0.20	0.18	0.00	0.04	0.15	0.32	0.58

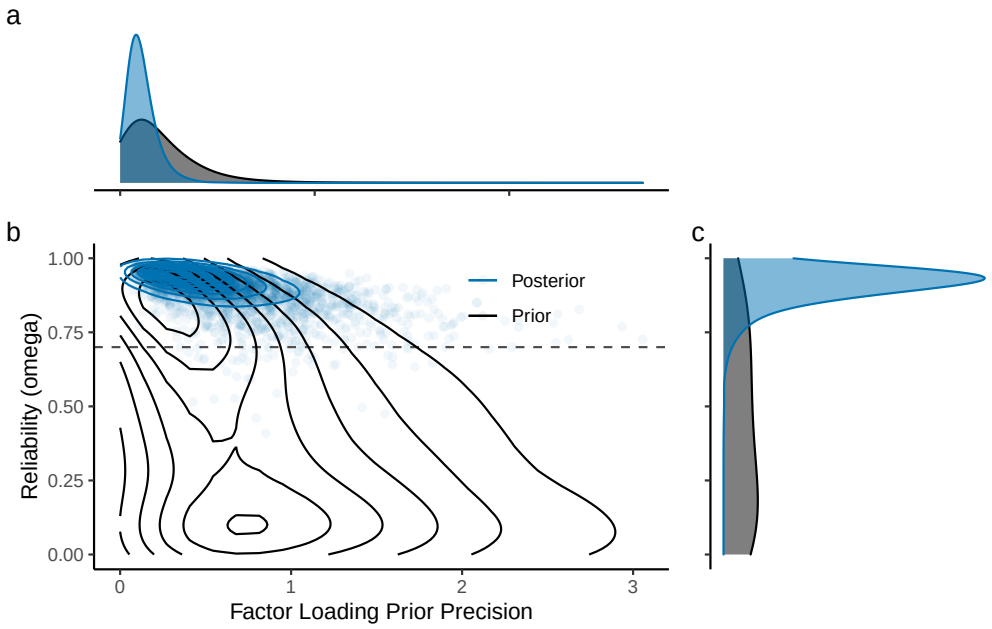


Figure E.2. Half-Cauchy hyper-prior for expanded sensitivity analysis using bivariate density plots. (a) Marginal density of hyper-prior parameter τ_λ for factor loading prior precision; (b) Bivariate posterior density plot of hyper-prior and reliability; and (c) marginal density of reliability distribution. The MCMC samples for are plotted with alpha-shading in (b) to demonstrate how disperse some draws from the posterior were.

REFERENCES

Agresti, A. (2010). *Modeling ordinal categorical data* (tech. rep.).

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*(422), 669–679. https://doi.org/10.1080/01621459.1993.10476321

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P. M., Barbaranelli, C., Medda, E., Nisticò, L., Stazi, M. A., & Caprara, G. V. (2010). Much more than model fitting? Evidence for the heritability of method effect associated with positively worded items of the life orientation test revised. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(4), 642–653. https://doi.org/10.1080/10705511.2010.510064

Ariyo, O., Lesaffre, E., Verbeke, G., Huisman, M., Heymans, M., & Twisk, J. (2021). Bayesian model selection for multilevel mediation models. *Statistica Neerlandica*. https://doi.org/10.1111/stan.12256

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 102–116. https://doi.org/10.1080/10705511.2014.859510

Becker, B., Debeer, D., Weirich, S., & Goldhammer, F. (2021). On the speed sensitivity parameter in the lognormal model for response times and implications for high-stakes measurement practice. *Applied Psychological Measurement*, 014662162110085. https://doi.org/10.1177/01466216211008530

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association1*, *45*(250), 164–180.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bollen, K. A., Bauer, D. J., Christ, S. L., & Edwards, M. C. (2010). Overview of structural equation models and recent extensions. *Statistics in the social sciences: Current methodological developments* (pp. 37–79). https://doi.org/10.1002/9780470583333.ch2

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Boughton, K., Smith, J., & Ren, H. (2017). Using response time data to detect compromised items and/or people. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of detecting cheating on tests* (pp. 177–190). Routledge.

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2021). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, 10944281211056520. https://doi.org/10.1177/10944281211056520

Bradburn, N. M., Sudman, S., Blair, E., & Stocking, C. (1978). Question threat and response bias. *Public Opinion Quarterly*, *42*(2), 221–234. https://doi.org/10.1086/268444

Brennan, R. L. (2005). Generalizability theory. *Educational Measurement Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second). The Guilford Press.

Burstyn, I., Yang, Y., & Robert Schnatter, A. (2014). Effects of non-differential exposure misclassification on false conclusions in hypothesis-generating studies. *International Journal of Environmental Research and Public health*, *11*(10), 10951–10966. https://doi.org/10.3390/ijerph111010951

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validity by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. Chapman; Hall/CRC.

Cernat, A., & Vandenplas, C. (2020). An evaluation of mixture confirmatory factor analysis for detecting social desirability bias. *Journal of Survey Statistics and Methodology*, *0*(smaa032), 1–27. https://doi.org/10.1093/jssam/smaa032

Chen, T. T., Timothy Chen, T., Hochberg, Y., & Tenenbein, A. (1984). Analysis of multivariate categorical data with misclassification errors by triple sampling schemes. *Journal of Statistical Planning and Inference*, *9*(2), 177–184. https://doi.org/10.1016/0378-3758(84)90018-1

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361. https://doi.org/10.1093/biomet/85.2.347

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart; Winston.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, *10*(102), 1–11. https://doi.org/10.3389/fpsyg.2019.00102

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Publications.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225–241. https://doi.org/10.1177/073428290502300303

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105–121. https://doi.org/10.1007/s10869-016-9479-0

Eickhoff, J. C., & Amemiya, Y. (2005). Latent variable models for misclassified polytomous outcome variables. *British Journal of Mathematical and Statistical Psychology*, *58*(Pt 2), 359–375. https://doi.org/10.1348/000711005X64970

Entink, R. H. K., Kuhn, J. T., Hornke, L. F., & Fox, J. P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, *14*(1), 54–75. https://doi.org/10.1037/a0014877

Falley, B. N., Stamey, J. D., & Beaujean, A. A. (2018). Bayesian estimation of logistic regression with misclassified covariates and response. *Journal of Applied Statistics*, *45*(10), 1756–1769. https://doi.org/10.1080/02664763.2017.1391182

Ferrando, P. J. (2006). Person-item distance and response time: An empirical study in personality measurement. *Psicologica*, *27*(1), 137–148.

Ferrando, P. J., Anguiano-Carrasco, C., & Demestre, J. (2013). Combining IRT and SEM: A hybrid model for fitting responses and response certainties. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 208–225. https://doi.org/10.1080/10705511.2013.769388

Ferrando, P. J., & Lorenzo-Seva, U. (2007a). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research*, *42*(4), 675–706. https://doi.org/10.1080/00273170701710247

Ferrando, P. J., & Lorenzo-Seva, U. (2007b). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*(6), 525–543. https://doi.org/10.1177/0146621606295197

Fox, J. P., Entink, R. K., & Van Der Linden, W. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*(7), 1–14. https://doi.org/10.18637/jss.v020.i07

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128

Fuller, W. A. (1987). *Measurement error models*. Wiley.

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46–70. https://doi.org/10.1037/met0000224

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). CRC press.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–807.

Goldstein, H., Browne, W. J., & Charlton, C. (2018). A bayesian model for measurement and misclassification errors alongside missing data, with an application to higher education participation in australia. *Journal of Applied Statistics*, *45*(5), 918–931. https://doi.org/10.1080/02664763.2017.1322558

Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling*, *8*(3), 243–261. https://doi.org/10.1177/1471082X0800800302

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*(1), 121–135. https://doi.org/10.1007/s11336-008-9098-4

Groves, R. M. (2004). *Survey errors and survey costs*. John Wiley & Sons.

Gustafson, P., & Le Nhu, D. (2002). Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics*, *58*(4), 878–887. https://doi.org/10.1111/j.0006-341x.2002.00878.x

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. https://doi.org/10.1177/2515245919898657

Hochberg, Y. (1977). On the use of double sampling schemes in analyzing categorical data with misclassification errors. *Journal of the American Statistical Association*, *72*(360a), 914–921. https://doi.org/10.1080/01621459.1977.10479983

Holden, R. R., & Kroner, D. G. (1992). Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology. *Psychological Assessment*, *4*(2), 170–173. https://doi.org/10.1037/1040-3590.4.2.170

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling an overview and a meta-analysis. *Sociological Methods and Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845. https://doi.org/10.1037/a0038510

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/BF02289343

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G., & Sörbom, D. (2015). LISREL 9.20 for Windows.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(1), 136–153. https://doi.org/10.1080/10705510701758406

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.

Koenig, C., Depaoli, S., Liu, H., & van de Schoot, R. (2022). *Moving beyond non-informative prior distributions: Achieving the full potential of bayesian methods for psychological research*. Frontiers Media SA.

Kroc, E., & Zumbo, B. D. (2020). A transdisciplinary view of measurement error models and the variations of X=T+E. *Journal of Mathematical Psychology*, *98*, 102372. https://doi.org/10.1016/j.jmp.2020.102372

Kuiper, N. A. (1981). Convergent evidence for the self as a prototype. *Personality and Social Psychology Bulletin*, *7*(3), 438–443. https://doi.org/10.1177/014616728173012

Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement*, *33*(3), 545–563. https://doi.org/10.1177/001316447303300302

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Chapman & Hall/CRC. https://doi.org/10.1201/9781315374604

Linde, A. v. d., & van der Linde, A. (2012). A bayesian view of model complexity. https://doi.org/10.1111/j.1467-9574.2011.00518.x

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

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, *80*(3), 366–380. https://doi.org/10.1037/0022-0663.80.3.366

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd). Routledge
Accession Number: 99847; OCLC: 53482692; Language: English.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press.

McGlothlin, A., Stamey, J. D., & Seaman, J. W. (2008). Binary regression with misclassified response and covariate subject to measurement error: A bayesian approach. *Biometrical Journal*, *50*(1), 123–134. https://doi.org/10.1002/bimj.200710402

McInturff, P., Johnson, W. O., Cowling, D., & Gardner, I. A. (2004). Modelling risk when binary outcomes are subject to error. *Statistics in Medicine*, *23*(7), 1095–1109. https://doi.org/10.1002/sim.1656

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meng, X. B., Tao, J., & Shi, N. Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *Journal of Statistical Computation and Simulation*, *84*(1), 1–21. https://doi.org/10.1080/00949655.2012.692368

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics*, *11*(1), 3–31. https://doi.org/10.3102/10769986011001003

Molenaar, D., Rózsa, S., & Kõ, N. (2021). Modeling asymmetry in the time–distance relation of ordinal personality items. *Applied Psychological Measurement*, *(online)*, 1–17. https://doi.org/10.1177/0146621621990756

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. J. (2015a). Fitting diffusion item response theory models for responses and response times using the R package diffIRT. *Journal of Statistical Software*, *66*(4), 1–34. http://www.jstatsoft.org/v66/i04/

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015b). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189. https://doi.org/10.1111/j.2044-8317.1985.tb00832.x

Naranjo, L., Pérez, C. J., Martín, J., Mutsvari, T., & Lesaffre, E. (2019). A Bayesian approach for misclassified ordinal response data. *Journal of Applied Statistics*, *46*(12), 2198–2215. https://doi.org/10.1080/02664763.2019.1582613

Nelson, T., Song, J. J., Chin, Y. M., & Stamey, J. D. (2018). Bayesian correction for misclassification in multilevel count data models. *Computational and Mathematical Methods in Medicine*, *2018*. https://doi.org/10.1155/2018/3212351

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Padgett, R. N., Jiang, S., Shero, L., & Kettler, T. (2022). Development of a perceptions of online learning scale to assess teachers' beliefs. *Journal of Psychoeducational Assessment (online)*. https://doi.org/10.31234/osf.io/ra5ms

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. *Proceedings of the 3rd international workshop on distributed statistical computing*, *124*(125.10), 1–10.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Press, S. James. (1968). Estimating from misclassified data. *Journal of the American Statistical Association*, *63*(321), 123–133. https://doi.org/10.1080/01621459.1968.11009227

Ranger, J. (2013). Modeling responses and response times in personality tests with rating scales. *Psychological Test and Assessment Modeling*, *55*(4), 361–382.

Ranger, J., & Kuhn, J. T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*(3), 214–231. https://doi.org/10.1177/0146621612439796

Ranger, J., & Kuhn, J.-T. (2018). Modeling responses and response times in rating scales with the linear ballistic accumulator. *Methodology*, *14*(3), 119–132. https://doi.org/10.1027/1614-2241/a000152

Ranger, J., & Ortner, T. M. (2011). Assessing personality traits through response latencies using item response theory. *Educational and Psychological Measurement*, *71*(2), 389–406. https://doi.org/10.1177/0013164410382895

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Richardson, S., & Gilks, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, *12*(18), 1703–1722. https://doi.org/10.1002/sim.4780121806

Rigdon, E. E., Becker, J. M., & Sarstedt, M. (2019). Factor indeterminacy as metrological uncertainty: Implications for advancing psychological measurement. *Multivariate Behavioral Research*, *54*(3), 429–443. https://doi.org/10.1080/00273171.2018.1535420

Rigdon, E. E., Sarstedt, M., & Becker, J. M. (2020). Quantify uncertainty in behavioral research. *Nature Human Behaviour*, *4*(4), 329–331. https://doi.org/10.1038/s41562-019-0806-0

Rios, J. A., & Soland, J. (2021). Investigating the impact of noneffortful responses on individual-level scores: Can the effort-moderated irt model serve as a solution? *Applied Psychological Measurement*, *45*(6), 391–406. https://doi.org/10.1177/01466216211013896

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In E. E. Roskam & R. Suck (Eds.), *Progress in mathematical psychology* (pp. 151–174). North-Holland.

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., & Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika*, *68*(4), 589–606. https://doi.org/10.1007/bf02295614

Roy, S., Banerjee, T., & Maiti, T. (2005). Measurement error model for misclassified binary responses. *Statistics in Medicine*, *24*(2), 269–283. https://doi.org/10.1002/sim.1886

Roy, S., & Banerjee, T. (2009). Analysis of misclassified correlated binary data using a multivariate probit model when covariates are subject to measurement error. *Biometrical Journal*, *51*(3), 420–432. https://doi.org/10.1002/bimj.200800127

Roy, S., Das, K., & Sarkar, A. (2013). Analysis of binary data with the possibility of wrong ascertainment. *Statistica Neerlandica*, *67*(3), 293–310. https://doi.org/10.1111/stan.12008

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In C. N. Mills, M. Potenza, J. J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237–266). Mahwah, NJ.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Taylor & Francis.

Smid, S. C., McNeish, D., & Miočević, R., Milica and van de Schoot. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201. https://doi.org/10.2307/1412107

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Sposto, R., Preston, D. L., Shimizu, Y., & Mabuchi, K. (1992). The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in a-bomb survivors. *Biometrics*, *48*(2), 605–617.

Su, Y.-S., & Yajima, M. (2020). *R2jags: Using r to run 'jags'* [R package version 0.6-1]. https://CRAN.R-project.org/package=R2jags

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association*, *65*(331), 1350–1361. https://doi.org/10.1080/01621459.1970.10481170

Thissen, D. (1983). Timed testing: An approach using item response theory. *New horizons in testing* (pp. 179–203). Elsevier. https://doi.org/10.1016/b978-0-12-742780-5.50019-6

Tuerlinckx, F., Molenaar, D., & van der Maas, H. L. J. (2016). Diffusion-based item response modeling. In W. J. van der Linden (Ed.), *Handbook of modern item response theory* (pp. 283–300). Chapman; Hall/CRC Press.

Tuerlinckx, F., & De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika*, *70*(4), 629–650. https://doi.org/10.1007/s11336-000-0810-3

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2021). A Response-Time-Based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*. https://doi.org/10.1007/s11336-021-09817-7

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, *34*(3), 378–394. https://doi.org/10.3102/1076998609332107

van der Linden, W. J., Klein Entink, R. H., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347. https://doi.org/10.1177/0146621609349800

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339–356. https://doi.org/10.1037/a0022749

Veen, D., Egberts, M. R., van Loey, N. E. E., & van de Schoot, R. (2020). Expert elicitation for latent growth curve models: The case of posttraumatic stress symptoms development in children with burn injuries. *Frontiers in Psychology*, *11*, 1197. https://doi.org/10.3389/fpsyg.2020.01197

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Verhelst, N. D., Verstralen, H. H. F. M., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). Springer New York. https://doi.org/10.1007/978-1-4757-2691-6_10

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12).

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. *Encyclopedia of statistics in behavioral science*.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. https://doi.org/10.1037/1082-989X.12.1.58

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & Demars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Yang, Y., & Xia, Y. (2019). Categorical omega with small sample sizes via bayesian estimation: An alternative to frequentist estimators. *Educational and Psychological Measurement*, *79*(1), 19–39. https://doi.org/10.1177/0013164417752008

Yiu, C.-F., & Poon, W.-Y. (2008). Estimating the polychoric correlation from misclassified data. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 49–74. https://doi.org/10.1348/000711006x131136

Zhang, J., Bohrnstedt, G., Park., J., Ikoma, S., Ogut., B., & Broer, M. (2021). *Mathematics motivation and its relationship with mathematics performance: Evidence from the NAEP-HSLS overlap sample* (tech. rep. [AIR-NAEP Working Paper #2021-03]). American Institutes for Research. Washington, DC.

Zhang, J., Templin, J., & Mintz, C. E. (2022). A model comparison approach to posterior predictive model checks in bayesian confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–11. https://doi.org/10.1080/10705511.2021.2012682