A Second-Order Growth Model for Longitudinal Item Response Data

Daniel Serrano

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology: L.L. Thurstone Psychometric Laboratory.

Chapel Hill
2010

Approved by

Patrick J. Curran, Ph.D.
Robert C. MacCallum, Ph.D.
David Thissen, Ph.D.
Daniel J. Bauer, Ph.D.
Eric Youngstrom, Ph.D.

ABSTRACT

Daniel Serrano: A Second-Order Growth Model for Longitudinal Item Response Data
(Under the direction of Patrick J. Curran)

This Dissertation explores the unique issues related to specifying and fitting a second-order growth model to longitudinal item response data. The model examined is a hybrid of the logistic IRT model for binary item responses and the latent growth curve model for repeated measures. Attention centered around parameterization, identification, estimation, and issues related to estimation of the model such as convergence, improper solutions, bias and root mean squared error (RMSE). Two variations on the proposed model: one with correlated errors (model 2) and one without (model 1). In each model two types of estimator were examined: full and limited information estimators. Two sample size conditions were examined, one with $N = 750$ observations, and another with $N = 3000$. In addition, two item parameter sets were examined, one having a wide range of difficulty and the other, narrow. Comparing analyses stratified across model, findings indicated greater rates of improper solutions and bias for model 2 versus model 1. Limited information estimators of model 1 performed worse than full information estimators, while the opposite was true for model 2. Bias and convergence issues were greatest when difficulty had a wide range. Lastly, sample size appeared to play a negligible role in bias and RMSE, though it did affect convergence issues and improper solutions. Based on empirical results presented in this simulation the proposed model appears to be a logical statistical framework for modeling longitudinal item responses.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Patrick J. Curran, without whose guidance and keen insights none of this would have been possible. I would like to thank my committee members, who have aided this project through feedback throughout the process. Many thanks are due to Taehun Lee, whose friendship and interest in discussing this project greatly contributed to the successful completion. Lastly, I would like to thank the institution of the L. L. Thurstone Psychometric Lab for providing the education, environment, and support necessary to complete this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Over the past 30 years psychological research has gradually shifted away from cross sectional designs and toward longitudinal or repeated sampling designs, under which scales are repeatedly administered to a sample of subjects tracked over multiple assessment occasions. The field's origins in experimental research likely account for the heavy reliance on cross sectional designs. However, as research began to focus on developmental processes of both normative and pathological behavior, longitudinal designs became prominent (Nesselroade & Baltes, 1979). Advances in statistical theory and computing have facilitated the adoption of this approach within the field (Bollen & Curran, 2006; Mehta & West, 2000). However, it is rare for such studies to incorporate measurement models to account for the measurement error inherent in the study of such unobservable psychological constructs as internalizing or externalizing behavior.

In the case of most measurement studies, single stage sampling is employed and results in the random sampling of respondents and their responses to a scale composed of k items. This historical design resulted in the development of many useful statistical models for assessing measurement properties in cross section, but renders the analysis of longitudinal responses complex. The primary obstacle to specification of a latent variable model characterizing the measurement properties of the sampled responses is the specification of a distribution for the latent variable and a distribution for the item-responses conditional upon the latent variable. In the case of uni-dimensional scales, only a univariate distribution need be specified for the latent variable. In contrast, consider a two

stage sampling design of a uni-dimensional psychopathology scale. In the first stage of sampling respondents are randomly sampled with equal probability of selection. In the second stage of sampling, responses to the items composing the uni-dimensional psychopathology scale are sampled at T measurement occasions. This design induces an T dimensional joint distribution for the latent variables underlying the uni-dimensional scale at each occasion. Thus the model specification now requires characterizing a T dimensional distribution for the latent variables, increasing the complexity of the conditional response distributions. When item responses are discrete and we wish to estimate item parameters or obtain latent variable scores while simultaneously taking into account the parameters characterizing the T dimensional latent variable distribution, there are two classes of methods: Factor analysis and item response theory procedures.

Within the factor-analytic tradition several models have been developed for embedding a measurement model for discrete item responses within a longitudinal model for the joint distribution of the repeated latent-variable measures. While this approach permits specification of a general model for longitudinal measurement models, it is limited to few items. In addition, it may not be amenable to application to psychopathology scales. Such scales often contain items which are rarely endorsed within a sample; examples would include items assessing suicidal ideation on a depression inventory. Such behaviors are rare enough in the general population that item endorsement rates are generally low. The factor analytic approach performs poorly with such items. In contrast the item response theory (IRT) model easily handles large item sets and rarely endorsed items. However, the traditional item response theory (IRT) model (Birnbaum, 1968; Bock & Lieberman, 1970; Bock & Aitkin, 1981; Bartholomew & Knott, 1999), is not easily amenable to accounting for the joint distribution of latent variables arising from repeated measures. Often, application of the IRT model to longitudinal responses requires modification of the data to fit existing architecture.

For example, in order to approximate latent variable scores Curran et al. (2008)

modified longitudinal item response data through a random sampling selection process converting the longitudinal data into representative cross sectional data thus permitting item evaluation within existing measurement model architecture. Provisional on the obtained pseudo-cross-sectional item parameters, latent variable scores were obtained for each subject at each time-point. However, because the scores were not obtained from a model which explicitly accounted for the repeated measures, the degree of precision of the resulting scores was likely distorted. The extent of this distortion is unclear, and would depend on the representativeness of the sample resulting from the random selection process. Nonetheless, some imprecision is induced under such an approach and exists for two reasons.

First, the item parameters on which the scores are conditioned do not characterize any of the time-specific item response patterns. Thus we can not determine how the scale performs at any given assessment. In addition, because the item parameters do not map on to any one assessment, the time-specific latent variable scores that can be estimated may not be conditioned on the correct item parameters. Consequently there may be some misspecification in the latent variable scores. Second, latent variable scores of this type are obtained using a shrinkage estimator. By that I mean scores are concentrated toward areas of highest precision and extreme scores are pulled in toward the mean. Because of the random sampling process, a given random draw could result in a preponderance of cases from a single time-point. In so doing, scores would be shrunken towards those characterizing a specific time-point rather than the time-specific latent variable which we desire. These two issues are but a few which serve to motivate the longitudinal IRT model.

Part of the reason that a more comprehensive model was not implemented by Curran et al. (2008) is that until recently computational limitations have prevented the specification and estimation of such models. In fact, such models remain in their infancy and the field is full of opportunities for study and development. Several longitudinal item

response models have been proposed, In this dissertation I develop and study a general model for longitudinal binary item responses. Specifically, I examine the performance of second-order growth models in the analysis of longitudinal item response data. I consider the application of the model to designs commonly encountered in the study of psychopathology: multiple repeated measures of a small item pool. This is in contrast to the item pools common to educational testing in which large item sets are repeatedly sampled few times. The second-order growth model has yet to be explicated or evaluated in the case of longitudinal item response data.

Though the literature on longitudinal IRT models is diverse, I focus here on the most prominent and notable work developing general statistical frameworks for estimating item parameters and scoring. Embedding a measurement model within a generalized linear mixed model, Johnson and Raudenbush (2006) specified a repeated measures Rasch model for two longitudinal assessments. Because of the Rasch specification prevented estimation of a distinct slope for each item, slopes were constrained to equality across items, the estimates then for the constrained slope and the threshold parameters were estimated as the average of the parameters between the two repeated measures. Justifications for employing the restricted Rasch model included estimation limitations, simplicity of interpretation of scale, and an empirical evaluation indicating that the Rasch model fit the data considered in application better than a two parameter logistic (2pl) model. Because the model given by Johnson and Raudenbush (2006) specifies item parameters as averages of the time-specific item parameters, the framework does not provide a means for testing invariance.

In a recent dissertation, Hill (2006) developed a 2pl IRT bi-factor model for two repeated measures. Hill (2006) employed the testlet-factors approach to specification, augmented by a general factor at each time. The testlet factors were specified by giving each repeated item a constant loading on the testlet, and constraining the mean and variance of the testlet factor to be $0, 1$ respectively. The error correlation for each pair of

items was then given by the square of the testlet loading. The bi-factor model given by (Hill, 2006) was identified by constraining $\theta_1 \sim N(0,1)$ and freely estimating the mean and variance of θ_2. Though a simulation design was not employed, Hill (2006) examined the performance of her model under single replications of several simulated data sets in order to examine item-set and sample size effects.

Alternative models have been proposed by Gibbons and Hedeker (1997), Liu and Hedeker (2006), and Liu (2008). These authors have specified IRT models in which the repeated measures are modeled at the item response level and not the level of θ. Specifically, these authors have fit standard IRT models at each assessment while simultaneously fitting a random effect for time to the observed item-responses. Because the time trend is modeled at the level of the item responses and not θ, one cannot disentangle the effects of growth from non-invariance. In contrast, the second-order growth parameterization models the repeated measures at the latent ability (θ) level. Though I defer thoroughly examining measurement invariance to future work, the model provides a general framework for the testing of invariance. The contribution of this work is therefore the explication of the proposed model and assessment of its performance in simulation.

1.1 Proposed Model

The proposed model is based on a second-order growth model where the lower order measurement models are repeated samples of the IRT model over time. These IRT models are parameterized using the threshold parameterization for the kth item:

$$\varphi_k = a_k(\theta - b_k), \tag{1.1}$$

where a_k is the slope (discrimination in IRT parlance) and b_k the threshold, and θ is the latent variable measured by the items. In this model the item-parameters to be estimated are a_k and b_k. In the case of binomial family response distributions the model will employ the logit inverse link function:

$$P(item_k = 1|\theta) = \frac{1}{1 + e^{-\varphi_k}}. \tag{1.2}$$

1.1.1 A Model for Correlated $\boldsymbol{\theta}$

To account for the correlated nature of the T longitudinal assessments, a second-order growth model will be specified for the repeated θ. The second-order growth model is simply a mixed effects model for $\boldsymbol{\theta}$, which accounts for the trend in the mean and covariance structure observed among the repeated measures in $\boldsymbol{\theta}$. Let $\boldsymbol{\eta}$ be an R dimensional matrix of unobserved random effects:

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta}, \end{bmatrix} \tag{1.3}$$

and let $\boldsymbol{\eta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\eta}}, \boldsymbol{\tau}_{\boldsymbol{\eta}})$. The fixed effects, contained in $\boldsymbol{\mu}_{\boldsymbol{\eta}}$, characterize the mean trend structure of $\boldsymbol{\theta}$, while the $R \times R$ dimensional random effects covariance matrix, $\boldsymbol{\tau}_{\boldsymbol{\eta}}$, characterizes the trend in the covariance structure of $\boldsymbol{\theta}$. The conditional model relating $\boldsymbol{\eta}$ to the $1 \times T$ dimensional vector $\boldsymbol{\theta}$, whose T^{th} element, $\boldsymbol{\theta}_T$, is the latent variable value for a single respondent at the Tth assessment, may be expressed as a function of a mixed-effect polynomial trend for time:

$$\boldsymbol{\theta} = \boldsymbol{\lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{1.4}$$

The model of interest employs a linear polynomial trend to parameterize $\boldsymbol{\lambda}$, though this model permits the parameterization of more complex trends. For example,

$$\boldsymbol{\lambda} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & T-1 \end{bmatrix}. \tag{1.5}$$

The time-specific disturbance terms, contained in $\boldsymbol{\epsilon}$, are assumed $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\psi})$, where $\boldsymbol{\psi}$ is a $T \times T$ dimensional diagonal matrix of time-specific variances, the Tth element of which is the variance of $\boldsymbol{\theta}_T$ conditioned upon the growth model.

The distribution for $\boldsymbol{\theta}$ may be derived using properties of linear combinations of Gaussian variates from the conditional model. Specifically, $\boldsymbol{\theta} \sim N(\boldsymbol{\lambda}\boldsymbol{\mu}_{\boldsymbol{\eta}}, \boldsymbol{\lambda}\boldsymbol{\tau}_{\boldsymbol{\eta}}\boldsymbol{\lambda}' + \boldsymbol{\psi})$.

This model provides estimates of parameters characterizing the distribution of the time-specific $\boldsymbol{\theta}_T$. However, the parameters are not freely estimated, but instead are implied by the linear trend specified in the model. Thus the model explicitly assumes that ability increases or decreases via a linear polynomial over time.

1.1.2 Advantages of this Approach

There is one main advantages of this approach relative to a multivariate longitudinal model which jointly estimates time-specific IRT models and characterizes the change over time by freely estimating the saturated mean and covariance structure of the matrix-valued $\boldsymbol{\theta}$. The advantage relates to the dimensions of integration required for the model relative to the number of parameters which summarize the moments of the distribution of $\boldsymbol{\theta}$. Under a binary response model both models estimate $2k$ item parameters, consequently the primary difference between the two models relates to the number of parameters estimated to characterize the distribution of $\boldsymbol{\theta}$. The multivariate model for T repeated measures estimates $\frac{(T)(T+1)}{2}$ parameters characterizing the covariance matrix $\boldsymbol{\theta}$, and T means. Defining the dimensions of the second-order growth component matrix $(\boldsymbol{\eta})$ as $R = dim(\boldsymbol{\eta})$, the number of estimated parameters under the second-order model may be expressed as $\frac{(R)(R+1)}{2}$ parameters characterizing the covariance matrix of $\boldsymbol{\eta}$, R means, and T time-specific error variances. With four repeated measures the saturated multivariate model estimates 14 total parameters, while the second-order model with linear polynomial trend estimates a total of 9 parameters. The difference in parameters required to summarize the distribution of $\boldsymbol{\theta}$, is a common result of invoking any structural model, however, what is unexpected is that this can be accomplished without increasing the burden of numerical integration.

We can express the marginal likelihood generally as

$$\int\int p(y|\boldsymbol{\theta})\ p(\boldsymbol{\theta}|\boldsymbol{\eta})\ \boldsymbol{\phi}(\boldsymbol{\eta})\ \partial\boldsymbol{\eta}\ \boldsymbol{\phi}(\boldsymbol{\theta})\partial\boldsymbol{\theta}, \tag{1.6}$$

which may be partitioned into:

$$\int p(y|\boldsymbol{\theta}) \left[\int p(\boldsymbol{\theta}|\boldsymbol{\eta})\ \boldsymbol{\phi}(\boldsymbol{\eta})\ \partial\boldsymbol{\eta}\right]\ \boldsymbol{\phi}(\boldsymbol{\theta})\ \partial\boldsymbol{\theta} \tag{1.7}$$

And the component of the integrals required to express the conditional model relating $\boldsymbol{\theta}$ to $\boldsymbol{\eta}$:

$$\int p(\boldsymbol{\theta}|\boldsymbol{\eta})\ \boldsymbol{\phi}(\boldsymbol{\eta})\ \partial\boldsymbol{\eta} \tag{1.8}$$

has a closed form solution, meaning that this model only requires T dimensions of integration and not $T + R$ dimensions of integration. Thus, the second-order growth model can characterize the distribution of $\boldsymbol{\theta}$ with as many dimensions of integration as a multivariate model with fewer parameters. Consequently, the apparent added complexity of the second-order model results in parametric savings with no additional estimation burden.

1.1.3 Identification

In traditional IRT sampling situations $i = 1 \ldots L$ subjects are sampled and each responds to $k = 1 \ldots N$ items. Under such sampling we identify the IRT model by assuming that $\theta \sim N(0, 1)$. However, the model of interest accounts for the effect of time on $\boldsymbol{\theta}$ through parameterization of the mean and covariance matrix of $\boldsymbol{\theta}$. Consequently, a different identification procedure is required, one that permits estimation of the parameters characterizing the distribution of $\boldsymbol{\theta}$. One alternative, employed by Hill (2006), identifies the model by constraining $\theta_1 \sim N(0, 1)$, constraining the item parameters corresponding to the first item to equality, and freely estimating all other item and structural parameters. However, consider imposing the constraint that the variance of $Var(\boldsymbol{\theta}_1) = 1$ in the context of the growth model. Under the proposed model, $Var(\boldsymbol{\theta}_1) = \tau_\alpha + \psi_{1,1}$, consequently, in order to employ the identification constraint employed by Hill (2006), the linear constraint $\tau_\alpha + \psi_{1,1} = 1$ would have to be imposed on the second order model.

Instead, I advocate identification based on the alternative reference item identification constraint. Under this constraint, the slope and threshold corresponding to the kth

item are constrained to 1 and 0 respectively at each time-point and the mean and variance for $\boldsymbol{\theta}_T$ are freely estimated at each repeated measure. The constraint on the item parameters implies moments for the distribution of $\boldsymbol{\theta}$. Specifically, omitting the inequalities used to derive these implied moments, $\mu = -a_k^\star b_k^\star$ and $\sigma^2 = (a_k^\star)^2$. It is important to observe that selection of the reference item has implications for the moments of the latent variable, because when estimation is sensitive to the location and dispersion of the distribution of the latent variable, poor choices of the reference item could lead to estimation difficulties. Thus when employing this identification constraint, it is important to keep this issue in mind in the event that estimation troubles are encountered. Nonetheless, it is a common and easily implemented identification constraint. In addition, the reference item is invariant to time, and therefore, changes in items over time must be attended to in order to understand potential estimation troubles.

1.1.4 Measurement Invariance

In order to model trends in $\boldsymbol{\theta}$, be it externalizing behavior, or educational ability, across repeated measures, we must assume that the definition of $\boldsymbol{\theta}$ does not change over time. In other words, to be able to describe change over time in $\boldsymbol{\theta}$ we must be able to demonstrate that $\boldsymbol{\theta}$ is changing, rather than the *definition* of $\boldsymbol{\theta}$ changing. While this notion exists within both the factor analytic and response theoretic frameworks, this concept of constant definition of $\boldsymbol{\theta}$ can generally be referred to as factorial invariance. There is more than one way of establishing factorial invariance; in this project I define invariance by holding the item parameters constant within item over time, yet permitting the mean and variance to change over time as detailed in the preceding section. Though invariance is likely to be violated within any given application of the proposed model, a comprehensive examination of the proposed model under *both* invariance and non-invariance is beyond the scope of this project. Consequently, in this dissertation all models examined will be parameterized consistent with invariance assumptions.

1.1.5 A Model For Correlated Errors

The probability model given in equation 1.2 expresses a relationship between the true score and item response under the assumption of independent and standardized errors. The error model corresponds to the standardized logistic distribution, with error variance $\frac{\pi^2}{3}$, and is a necessary restriction imposed for model identification (McCullough & Nelder, 1989). Given a set of repeated item responses one does not know whether the the growth model explicated is sufficient to model the correlated nature of the repeated responses. In some circumstances, repeated measurement may induce correlation in *both* the true scores and the errors. In that case one must augment the proposed model with correlations among the errors. However, constraining the error model for identification does not permit flexible parameterization of the error covariances as is common in linear models. Instead, one could consider testlet factors to account for the item-specific correlation. Whereas $\boldsymbol{\theta}_T$ is a constant person effect at the Tth measurement occasion which accounts for the correlation in item responses, the testlet factor accounts for the correlation within a specific item induced by repeated administration of that item. The testlet factor model may be expressed for a given subject by augmenting equation 1.1:

$$\varphi_{kt} = a_{kt}(\theta_t - b_{kt}) + \lambda_k \theta_k, \tag{1.9}$$

where λ_k is the loading that item k has on the item-specific testlet factor θ_k. For identification purposes I set $\theta_k \sim N(0, 1)$. When item k is measured repeatedly, each realized item response is a function of the testlet factor θ_k. Given constant loading on the testlet factor across items, we can determine the within-item error-correlation resulting from repeated administration by $\boldsymbol{\lambda}\boldsymbol{\lambda}'$. For example, with three repeated measures of a given item and constant loading estimated to be .5477 the implied residual item-correlation is .3, given by $\lambda_k^2 = .5477^2 = \rho_k = .3$. This is identical to the method employed by Hill (2006) for handling correlated errors. In some cases it may be reasonable to expect a decaying correlation pattern among more temporally distal measurements of the same

item. In such circumstances, $\boldsymbol{\lambda}$ may be structured in order to fit patterned correlation structures. Nonlinear constraints could be employed in order for λ_k to produce an autoregressive structure of user-specified order. A substantial problem with the specification of an unrestricted error-correlation model is related to computational burden. Whereas the model for correlated true scores increases by one dimension for every added repeated measure, the unrestricted model for correlated errors with N items and M repeated measures consists of a minimum of $N + M$ dimensions. And that is under a model with constant error-correlation within item. If error correlations are non-constant within item over time the dimension of the unrestricted model for correlated errors can have as many as

$$\left(\frac{M(M-1)}{2}\right) N + M \tag{1.10}$$

dimensions. Thus an unrestricted model may be computationally impractical for more than a few items or a few repeated measures, unless an estimation routine can be employed which can handle very high-dimensional models.

1.2 Estimation

Two primary estimation methods exist for latent variable models with Bernoulli item responses: Full and limited information estimation. The advantage of the full information approach is that it takes advantage of the raw response patterns and is rooted in the established theory of maximum likelihood (Bock & Lieberman, 1970). Historical disadvantages of the full information approach have included heavy computing burden and the restriction of the approach to a single factor model and few items. These disadvantages motivated the development of limited information techniques that permitted the fitting of multi-factor models to larger sets of items with reduced computing time (Christofferson, 1975; Muthén, 1978; Olsson, 1979). The limited information approach avoids the use of the raw response patterns and instead uses first-order and second-order marginal proportions obtained from contingency tables in order to fit the model. This

approach has gained wide appeal among factor analysts because it serves as an analog in discrete indicators to the weighted least squares estimation of the common factor model under continuous indicators (Bartholomew & Knott, 1999). A substantial limitation of this approach is the rate at which the weight matrix grows, which has historically restricted the method to the analysis of no more than 25 items. The ascension of the limited information estimator did not limit work refining the full information estimator. The modern full-information approach to item parameter estimation is based on the work of Bock and Aitkin (1981) who employed an EM-type algorithm to facilitate the approximation to the integrals that bogged down the computations in Bock and Lieberman (1970). The advantage of the Bock and Aitkin (1981) method is that it increased the number of items that could be analyzed from a maximum of 12 to nearly 100, making it an exceedingly useful method for estimating single-factor models with large item-sets. These two approaches to estimation are discussed in the following sections.

1.2.1 Limited Information Estimation

Drawing on the strengths of the factor analytic model, Christofferson (1975), Muthén (1978), and Olsson (1979) developed a general framework for fitting models to discretely distributed item responses. Parameter estimates obtained within this framework may be translated to IRT parameters. This framework serves as an elegant complement to the item response model. Because the parameters can be compared, and the proposed model has been estimable within this framework for years, I am interested in examining the limited information estimator and its performance relative to the full information estimator of IRT.

Pearson (1901) pioneered an approach to the analysis of Bernoulli item responses that has enjoyed widespread favor among psychometricians, primarily because of the fact that it serves as a heuristic analog in discrete indicators to the widely employed, and well understood, method of fitting the common factor model to the correlation matrix of continuous indicators. Following Lord and Novick (1968), for pairs of discrete items

$\{k, l\}$, the joint Gaussian density defined through an auxiliary threshold model gives conditional proportions:

$$\pi_{kl} = \int_{\tau_k}^{\infty} \int_{\tau_l}^{\infty} \Phi(\theta_k, \theta_l, \rho_{kl}) \partial\theta_k \partial\theta_l, \tag{1.11}$$

where θ_k and θ_l are unobserved Gaussian variates dichotomized at τ_k and τ_l respectively to define the observed dichotomous indicators y_k and y_l, and where $y_k = 1 \iff \theta_k \geq \tau_k$ and $y_k = 0 \iff \theta_k \leq \tau_k$. The unobserved Gaussian variates are correlated ρ_{kl}. The conditional proportion given above in equation 1.11 corresponds to $y_k = 1$ and $y_l = 1$, the complement can be obtained by inversion of the integration limits, and the remaining proportions can be obtained by conditional inversion of the integration limits. The marginal proportions for y_k and y_l defined by the auxiliary threshold model are given by

$$\pi_k = \int_{\tau_k}^{\infty} \int_{-\infty}^{\infty} \Phi(\theta_k, \theta_l, \rho_{kl}) \partial\theta_k \partial\theta_l, \tag{1.12}$$

$$\pi_l = \int_{-\infty}^{\infty} \int_{\tau_l}^{\infty} \Phi(\theta_k, \theta_l, \rho_{kl}) \partial\theta_k \partial\theta_l, \tag{1.13}$$

If the proportions, π_k, π_l, π_{kl} are known then the parameters, τ_k, τ_l, and ρ_{kl} defining the joint Gaussian distribution characterizing θ_k and θ_l are uniquely defined by the given functions. However, in the absence of the population proportions, the sample estimates must be used, resulting in the estimation of a sample tetrachoric correlation coefficient. When, for all pairs of items, the sample tetrachoric correlation coefficient is computed, the tetrachoric correlation matrix can be populated. The previously mentioned heuristic method relies on the sample tetrachoric correlation matrix as input to the fitting function. An unfortunate property of the sample tetrachoric correlation matrix is that even when the population tetrachoric correlation matrix is well defined, the sample tetrachoric correlation matrix can be degenerate, particularly as $|\hat{\rho_{kl}}| \to 1$ (Lord & Novick, 1968).

Modern methods of limited information estimation are similar to the method proposed by Pearson (1901), where each of three related methods seek to simplify the com-

puting burden inherent in the approach taken by Pearson (1901) and early full information methods (Bock & Lieberman, 1970). The first limited information estimator was derived by Christofferson (1975) who sought to find some approximation to the full information in the data that would permit easier model estimation of more complex models for more items. Rather than employing the response pattern, Christofferson (1975) chose to employ the first and second-order marginal proportions for model estimation. The population first-order proportions have been defined in equations 1.12 and 1.13 while the second-order proportions are given in equation 1.11. Christofferson (1975) defined the sample realizations of these proportions as $\hat{P}_k = \pi_k + \epsilon_k$, $\hat{P}_l = \pi_l + \epsilon_l$, and $\hat{P}_{kl} = \pi_{kl} + \epsilon_{kl}$.

While the first-order sample proportions were estimated by Christofferson (1975) via standard routines, the second-order sample proportions had to be approximated via the tetrachoric expansion. The population and sample proportions can be stacked in the vectors $\boldsymbol{\pi}$ and $\mathbf{P}$ respectively. Stacking the errors, defined as $\mathbf{P} - \boldsymbol{\pi}$, into the vector $\boldsymbol{\epsilon}$ and then computing the corresponding covariance matrix of these errors yields $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ would produce the covariance matrix of the errors, permitting weighted minimization of the squared errors, $\boldsymbol{\epsilon}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{\epsilon}'$. Unfortunately, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ is not known. Christofferson (1975) derived a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ based on third and fourth-order moments which produced an efficient generalized least squares (GLS) estimator. In fact, in comparison to the results obtained by Bock and Lieberman (1970), both the point estimates and the standard errors for the GLS estimator were nearly identical. Compared to the solution given by Bock and Lieberman (1970), this approach reduced the computing burden, increased the analyzable number of items to 25, and allowed for the fitting of multi-factor models. However, it required the evaluation of the integrals contained in equations 1.11− 1.13 at each stage of iteration.

In order to circumvent the burdensome integration required by the GLS estimator proposed by Christofferson (1975), Muthén (1978) inverted the integrals in equations 1.11− 1.13. This inversion results in the definition of the thresholds and tetrachoric

correlations rather than the first-order and second-order marginal moments employed by Christofferson (1975). Reparameterizing the elements contained in $\boldsymbol{\epsilon}$ as a function of the distance between the sample and model-implied thresholds and tetrachoric correlation coefficients yielded a GLS estimator as efficient as that derived by Christofferson (1975). Muthén (1978) described the relationship between the methods proposed by himself and Christofferson (1975) to the heuristic method, noting that the heuristic method was analogous to the GLS method replacing the weight matrix with an identity matrix, resulting in an unweighted least squares (ULS) solution. The approaches proposed by Christofferson (1975) and Muthén (1978) provided a least squares theory for fitting the heuristic method, with the main added advantage being the provision of efficient standard errors. Because the fitting function employed in this approach measures discrepancy primarily as a function of the distance between the sample and model-implied correlation coefficients, this method serves as the dichotomous analog to the weighted least squares estimator for continuous indicators.

In the case of polytomous item responses, Olsson (1979) solved the problem of efficient computation of the polychoric correlation matrix. The polychoric correlation matrix is an extension of the tetrachoric correlation to multinomial item responses. Rather than attempting to evaluate the full multinomial distribution of the set of polytomous responses in order to estimate the polychoric correlations, Olsson (1979) provided an algorithm for deriving the correlations as a function of the bivariate marginals obtained from pairwise contingency tables. Focussing on pairwise evaluation of the conditional relations between the items to characterize the multivariate distribution underlying the set of contingency tables provided a practical solution to a formerly intractable computing problem. This approach forms the basis for modern approaches to weighted least squares estimates of item parameters given discrete item-responses. The algorithm implemented in most commercial software involves three steps of estimation. In the first stage, univariate marginal proportions are used to obtain estimates of the threshold parameters.

In the second stage, bivariate proportions are employed to obtain estimates of the poly or tetrachoric correlation (following Olsson, 1979) conditioned upon the threshold estimates obtained in stage 1. In the final stage of estimation the factor model is estimated by employing weighted least squares to the polychoric correlation matrix, where the weight matrix corresponds to the asymptotic covariance matrix of the polychoric correlations.

1.2.2 Full Information Estimation

For the case of both large and small item pools, no estimator is more optimal than the full information estimator historically employed within IRT. This estimator fits nicely within the generalized linear model theory (McCullough & Nelder, 1989) in which analysis is based on the expression and maximization of a likelihood for the raw item responses. Consequently, the full information estimator retains all the desirable properties associated with maximum likelihood theory which are not available within the limited information estimator.

Given the full response pattern rather than statistics summarizing the response pattern, estimation must employ a likelihood for the response given the latent variable. However, because the latent variable is not observed, in order to obtain estimates of the item parameters governing the responses to a given item it is reasonable to treat the latent variable as a nuisance over which to integrate in order to obtain item parameters marginal to the latent variable. This requires specification and evaluation of the marginal log likelihood function. This function may be expressed for respondent i as:

$$\ell_i = \sum_{k=1}^{N} \int_{-\infty}^{\infty} \log(g(item_k|\theta)) + \log(\phi(\theta))\partial\theta. \tag{1.14}$$

Here, $g(item_k|\theta)$ is a general expression for the conditional likelihood of the item response given θ, and $\phi(\theta)$ is the distribution assumed for θ, in most applications this is the standardized Gaussian density. In the case of Gaussian item-responses and Gaussian latent variables one can marginalize the log-likelihood in closed form because the log likelihood is itself an additive function of two Gaussian densities. In contrast, with Bernoulli item-

responses and Gaussian latent variables, the log likelihood is of indeterminate form meaning that the likelihood cannot be marginalized in closed form. Consequently, marginal maximization requires integration of the log likelihood over the Gaussian latent variables. In order to understand the concept of numerical approximations to marginal maximum likelihood it is useful to understand the procedure given by Bock and Aitkin (1981), which gave an early and useful means of approximating the integral in equation 1.14.

Following Bartholomew and Knott (1999) the conditional response function may be expressed generally as $\pi_i(\theta)$, thus permitting the derivation of the estimator under either the probit or logit response functions. The EM-type algorithm given by Bock and Aitkin (1981) requires Gauss-Hermite quadrature approximations to the marginal and conditional likelihood functions, given respectively as

$$f(\mathbf{x}_h) = \sum_{q=1}^{r} f(\mathbf{x}_h|\theta_q)w(\theta_q) \quad (h = 1, 2, \ldots, n) \tag{1.15}$$

and

$$f(\mathbf{x}_h|\theta_q) = \prod_{i=1}^{P} [\pi_i(\theta_q)]^{x_{ih}} [1 - \pi_i(\theta_q)]^{1-x_{ih}}, \tag{1.16}$$

where $\mathbf{x}_h$ is the observed response pattern, of which there may be n, and θ_q is the q_{th} Gauss-Hermite quadrature node, derived as the q_{th} root of the Legendre polynomial, with corresponding quadrature weight $w(\theta_q)$. Estimation requires maximization of $\ell = \sum_{h=1}^{n} f(\mathbf{x}_h)$. Expressing the response model as $\alpha_{i0} + \alpha_{i1}\theta$ permits a simplified definition of the gradient function:

$$\frac{\partial \ell}{\partial \alpha_{il}} = \sum_{q=1}^{r} \frac{\partial \pi_i(\theta_q)}{\partial \alpha_{il}} \frac{[r_{iq} - N_q \pi_i(\theta_q)]}{\pi_i(\theta_q)[1 - \pi_i(\theta_q)]} \tag{1.17}$$

where r_{iq} and N_q are the expected response pattern and corresponding expected frequency of that response pattern respectively. Given provisional values of α_{il}, r_{iq} and N_q may be solved for, then holding r_{iq} and N_q fixed, values of α_{il} may be obtained by maximum likelihood, where maximization is based on probit regression in the case of the ogive model and logistic regression in the case of the logistic model. Conditional upon $l = 0, 1$,

this process is repeated until convergence is obtained for estimates of α_{il}. The asymptotic covariance matrix need not be computed for estimates to be obtained, thus circumventing the limitations of the approach taken by Bock and Lieberman (1970). Of course, the lack of this matrix also prevents computation of standard errors for the item parameters, α_{il}, which was one of the advantages of the Bock and Lieberman (1970) method.

The full information approach given by Bock and Aitkin (1981) was based on evaluation of the integral via summation over Gauss-Hermite quadrature nodes where integration nodes were distributed across the range of the distribution of θ, assuming $\theta \sim N(0, 1)$, at fixed intervals (generally user specified intervals). This approach is commonly referred to as fixed quadrature approximations to an integral. While this approach works well when the posterior distribution is normally distributed and when models have few dimensions of integration, the models considered in this dissertation require an alternative procedure that can accommodate the high-dimensional nature of the models of interest. Two alternative approximations to the integral exist for just such circumstances, they are adaptive quadrature and Monte-Carlo integration.

As can be seen in equation 1.15, the integral approximation is based on evaluating quadrature points distributed at fixed intervals along the assumed range of θ. Given that under most applications we assume that $\theta \sim N(0, 1)$, quadrature nodes are generally arrayed at fixed intervals between ± 3. An added complexity unique to the models that I consider here is that the latent ability is now vector-valued for each respondent and is no longer scalar. Because the integrand evaluated in equation 1.15 is proportional to the posterior density (Skrondal & Rabe-Hesketh, 2004), when the posterior is asymmetric or departs substantially from Gaussian form, fixed quadrature may provide a poor solution. Rather than basing integration on the assumption that the mode of the integrand lies within the domain corresponding to a standard Gaussian density, we can estimate the moments of the integrand and relocate the quadrature points according to the estimated moments of the integrand. Estimation of the moments of the integrand requires iterative

updating, and therefore at each iteration the quadrature nodes adapt to new locations given the current estimates of the integrand moments. This is the essence of adaptive quadrature. To be more precise, following Naylor and Smith (1982) the expectation of the integrand may be expressed as:

$$\varepsilon[f(\boldsymbol{\theta}_q)] = \int_{-\infty}^{\infty} f(\boldsymbol{\theta})p(\boldsymbol{\theta})\partial\boldsymbol{\theta}. \tag{1.18}$$

Starting with $\boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{I})$, and defining $f(\boldsymbol{\theta})$ as a transformation which gives a non-standard Gaussian distribution gives a means of defining a function which can be updated to determine the parameters of the integrand. For example, following the work of Schilling and Bock (2005), $f(\boldsymbol{\theta}) = \mathbf{T}\boldsymbol{\theta} + \boldsymbol{\mu}$, where $\mathbf{T}$ corresponds to the Cholesky decomposition of the covariance matrix of the posterior density and $\boldsymbol{\mu}$ the posterior mean, estimates of which help guide the location of the quadrature nodes used in approximating the integrand. Many implementations of adaptive quadrature estimate the mode ($\tilde{\boldsymbol{\mu}}$) and information matrix at the mode ($I^{-1}(\tilde{\boldsymbol{\mu}})$) rather than the mean and covariance matrix because they are easier to estimate. Scaling by $f(\boldsymbol{\theta})$ gives the adaptive quadrature analog of equation 1.15:

$$f(\boldsymbol{\theta}_q) = |\mathbf{T}| \sum_{i_d=1}^{q} W_{i_d} \ldots \sum_{i_1=1}^{q} W_{i_1} f(\mathbf{T}\boldsymbol{\theta}_{i_1 \ldots i_d} + \tilde{\boldsymbol{\mu}}), \tag{1.19}$$

where $\boldsymbol{\theta}_{i_q}$ is a quadrature point and W_{i_q} is the corresponding weight.

Monte-Carlo integration differs from adaptive quadrature in how the moments of the integrand are estimated. Whereas in adaptive quadrature moments are estimated by evaluating draws of empirical Bayes estimates of $\boldsymbol{\theta}$ at each iteration, in Monte-Carlo integration the moments are estimated by simulating $s = 1 \ldots t$ random draws of size n from $p(\boldsymbol{\theta})$. As a result, $f(\boldsymbol{\theta}_h)$ is identical to that given in equation 1.19, only weights are not given because rather than evaluating nodes and weights, simulated draws from each dimension of the prior substitute for the empirical Bayes estimates of $\boldsymbol{\theta}$. Consequently,

the Monte-Carlo analog of equation 1.15 is given as:

$$f(\boldsymbol{\theta}_q) = |\mathbf{T}|\frac{1}{n}\sum_{s=1}^{t} f(\tilde{\boldsymbol{\theta}}_s), \tag{1.20}$$

where $\tilde{\boldsymbol{\theta}}_s$ is the vector-valued draw of simulated values of $\boldsymbol{\theta}$ simulated from a multivariate Gaussian density $\tilde{\boldsymbol{\theta}}_s \sim N(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

In sum, each of these approaches to estimation have strengths and limitations. Weighted least squares with mean and variance adjustment (WLSMV) is widely recognized as being capable of fitting large models to few items (Joreskog & Moustaki, 2001), in addition certain parameterizations permit easier specification of error covariances/correlations. Consequently, I would expect this to be a stable estimator when few items are repeatedly sampled and a highly parameterized model is indicated. Full information estimation based on adaptive quadrature approximations to the integral work well for large item sets when models are relatively parsimonious. Thus, the fitting of the model for correlated true scores should be easily fit in this framework even with item sets that would prove prohibitively large for WLSMV (Skrondal & Rabe-Hesketh, 2004). However, the accuracy of the solution depends on the number of quadrature points employed per dimension, and as dimensions increase, quadrature points must decrease to offset the computing burden. Thus, even though a complex model may be fit using adaptive quadrature, the solution may not be optimal due to increased error in approximating the integral with fewer pieces of information per dimension (Schilling & Bock, 2005). Lastly, for highly parameterized models of many dimensions irrespective of item set size, I would expect a simulated integral, like that employed in Monte-Carlo integration, to more readily accommodate optimization than quadrature based approaches. Therefore, I would expect Monte-Carlo integration to perform well when fitting a model for either or both the correlated true scores and errors, though the increased computing time associated with simulating the integral would only justify employing this estimation routine for the model with correlated true scores *and* errors.

1.3 Hypotheses

I propose five key hypotheses relating to model complexity, estimators, difficulty range width, and sample size, as well as the interaction between estimators across sample size.

1.3.1 Model Complexity Hypothesis

Two general classes of models are considered in this dissertation: A model for correlated true scores only, denoted model 1, and A model for both correlated true scores *and* errors, denoted model 2. Comparing these two models in terms of computing time and rates of estimation failure (non-convergence and boundary solutions) will demonstrate advantages for Model 1. Models with more complex error correlation structures will be harder to fit and exhibit greater bias and inflated root mean squared error (RMSE) compared to models with no error correlations. Though the bias and RMSE may not be compared directly across model types, the difference in absolute magnitude of the estimation error will be discussed across stratified analyses.

1.3.2 Estimator Hypothesis

Compared to WLSMV estimation, FIML estimators, either quadrature or Monte-Carlo based, will differ little in point estimate accuracy. However, observed differences will favor FIML estimation. In particular, FIML estimation will be less biased and have lower RMSE than limited information estimation. Within quadrature based estimation, consistent with the work of Schilling and Bock (2005), accuracy of point estimates will increase as the number of quadrature points per dimension increase. I expect the relative ranking of the estimators, in terms of bias and RMSE, to favor FIML based on 7 QP more than 3 QP, which will be favored over WLSMV. The difference between the limited information estimator and FIML estimators in terms of computing time will demonstrate advantages of limited information estimation over full information. Limited information estimation will differ from full information estimation in the rate of estimation problems

such that WLSMV will have more cases of non-convergence or improper solutions than will FIML.

1.3.3 Difficulty Range Hypothesis

Consistent with the ideas put forth by Schilling and Bock (2005), when few items cover a wide range of θ, as when the difficulty parameters are spread widely across θ, limited information estimators and quadrature based estimators employing few quadrature points will fail or exhibit high degrees of bias and RMSE.

1.3.4 Sample Size Hypotheses

Estimates based on larger sample sizes will show less bias and less dispersion (as measured by RMSE). In addition, rates of non-convergence (NCV) will decrease as sample size increases.

1.3.5 Sample Size By Estimator Hypothesis

NCV rates will decrease as sample size increases for all estimators. Because NCV rates for WLSMV are a function of degenerate contingency tables, which are themselves a function of sample size, the decrease in NCV rates will be greatest for WLSMV. In addition, because sparse, though not degenerate, contingency tables can pose estimation problems for WLSMV, compared to other estimators, WLSMV will have higher rates of bias and RMSE at lower sample sizes. Consequently, increasing sample size will decrease the difference in bias between the WLSMV and FIML estimators, thus the difference between estimators will be greatest at small sample sizes and smallest at large sample sizes. There is no evidence to suggest that a similar effect would be observed for FIML estimators.

Testing of these hypotheses will permit me to examine two useful and theoretically important longitudinal IRT models. Because there are no existing simulation studies of this or any other longitudinal IRT model, these hypotheses will permit the explication of the effects of difficulty range, sample size, and estimator. These are three of the most

basic and important considerations in any IRT application. Therefore, within existing computational constraints, this study provides a thorough preliminary examination of the performance of the proposed models under the most fundamental and important aspects considered in the specification of any IRT model.

CHAPTER 2

Method

In order to empirically test the proposed hypotheses I employed a simulation study. All estimation was based on the Mplus system (Muthén & Muthén, 2007). Due to computational considerations I made heavy use of fractional experimental design rather than a full factorial design. Outcomes of interest in this study included non-convergence rates, rates of boundary solutions, bias, and root mean squared error (RMSE) of estimates. I simulated 300 replications per cell of the design.

2.0.6 Model Complexity Manipulation

Two longitudinal models were examined, model 1 and 2. For each model, nine items were simulated over four repeated measures. Model 1 was the proposed second-order growth curve with no correlated errors, having 4 dimensions of integration, one for each time-point. Model 2 extended model 1 to include constant error correlation of .3 for 4 of the 9 items. Because for FIML estimators the error correlations were parameterized via testlet factors, model 2 had a total of 8 dimensions of integration. Item and structural parameters for models 1 and 2 are given in Table 2.1. Item and structural parameters were held constant across models. The only difference between model 1 and model 2 was that the constant error correlation of .3 was simulated for the 4 items in model 2.

2.0.7 Estimator Manipulation

The estimator component of the design had three levels: WLSMV, quadrature, and Monte-Carlo based estimation, and was fractionally nested within model as a result of

computing limitations. Consequently, within model 1 WLSMV and quadrature based estimation were contrasted. The dimensions of integration associated with model 2 likely exceeded or were at the limit of the available computing resources afforded by quadrature based estimation, thus, only WLSMV and Monte-Carlo based estimation were contrasted within this model. Monte-Carlo EM estimation was based on $N = 500$ simulated draws from the posterior per dimension of integration. Within model 1, quadrature-based solutions were contrasted across quadrature point conditions. The quadrature point effect had two levels: the minimum number greater than one and maximum feasible number of odd quadrature points per dimension, which were 3 and 7 quadrature points respectively.

For both full and limited information estimation, Mplus parameterizes the response model with a slope and intercept ($\boldsymbol{\varphi}_k = c_k + a_k\boldsymbol{\theta}$), where $c_k = a_k \times b_k$. Because the generating model is based on the slope and threshold parameterization, $\boldsymbol{\varphi}_k = a_k(\boldsymbol{\theta} - b_k)$, in order to compare accuracy of item parameter estimates to generating parameters, intercept estimates were re-parameterized to thresholds as $b_k = \frac{c_k}{a_k}$. Even after this conversion, an additional rescaling was required for WLSMV estimates obtained from the theta parameterization, denoted with a ϑ subscript, to translate them from the probit to logit metric:

$$\begin{aligned} a_{fiml} &= a_\vartheta \\ b_{fiml} &= \left(\frac{c_\vartheta}{a_\vartheta}\right) 1.7 \\ \sigma^2_{fiml} &= (\sigma^2_\vartheta)\, 1.7^2 \\ Cov_{fiml} &= (Cov_\vartheta)\, 1.7^2 \\ \mu_{fiml} &= (\mu_\vartheta)\, 1.7. \end{aligned} \tag{2.1}$$

where 1.7 is a scaling factor for the difference between the standardized Gaussian and standardized Logistic distribution functions.

2.0.8 Difficulty Range Manipulation

Width of the difficulty, or threshold, parameter range was manipulated as follows: holding the distribution parameters for $\boldsymbol{\theta}$ constant, when the difficulty parameters covered a narrow range of θ, test information plateaus were approximately bounded between $\pm$ 1 standard deviations of θ at each time point. When the difficulty parameters covered a wide range, test information plateaus were approximately bounded between $\pm$ 2 standard deviations of theta at each time point. The number of items were held fixed at 9 items per assessment. Each model examined had a total of 4 repeated measures, resulting in a total item set of $9 \times 4 = 36$ items. As stated, item parameters within item were held constant over time, with two parameters per item, this simulation examined a total of 16 item parameters per model. There were 16, and not 18, because one of the items had item parameters constrained and not estimated for purposes of identification. Slope and threshold parameters were modeled after those presented by Curran et al. (2008) so as to be representative of the diversity and magnitude of item parameters encountered in psychopathology research. Structural parameters for the higher and lower order factors were held constant across the two response pattern conditions.

Table 2.1: Population Values for Item and Structural Parameters

	Generating values	
P	Set 1	Set 2
a_1	0.46	0.46
a_2	0.69	0.69
a_4	0.92	1.92
a_5	1.15	1.20
a_6	1.37	1.80

Continued on next page

Table 2.1 – continued from previous page

Generating values		
P	Set 1	Set 2
a_7	1.68	1.68
a_8	1.76	1.76
a_9	0.30	0.30
b_1	2.30	3.30
b_2	-0.50	-1.00
b_4	3.00	4.00
b_5	1.50	2.50
b_6	1.00	1.20
b_7	-0.30	-1.50
b_8	2.00	2.80
b_9	-1.00	-2.00
μ_α	1.39	1.39
μ_β	0.50	0.50
τ_α	0.67	0.67
$\tau_{\alpha\beta}$	0.05	0.05
τ_β	0.05	0.05
ψ_1	0.67	0.67
ψ_2	0.81	0.81
ψ_3	1.05	1.05
ψ_4	1.39	1.39

Table 2.1 presents the generating parameters employed in parameter set 1 and 2. Note that for both parameter sets the only difference is the item parameters, with the structural model parameters being fixed across sets. To characterize the difficulty range effect, ICC plots for the two sets are plotted for each time-point. For ease of interpretation, the ICCs are based on standardized parameters, and not the raw parameters listed in Table 2.1. The raw parameters were converted to the standardized metric using the following procedure: Given

$$\varphi_{kt} = a_{kt}(\theta_t - b_{kt}) \mid \theta_t \sim N(\mu_t, \sigma_t^2), \tag{2.2}$$

I can obtain a standardized expression for φ_{kt} under $\theta_t^\star \sim N(0,1)$ by a Gaussian change of variable function. Let

$$\theta_t^\star = \frac{\theta_t - \mu_t}{\sigma_t} \;\rightarrow\; \theta_t^\star \sim N(0,1), \tag{2.3}$$

then I can express

$$\varphi_{kt}^\star = a_{kt}^\star(\theta_t^\star - b_{kt}^\star) \mid \theta_t^\star \sim N(0,1), \tag{2.4}$$

where

$$a_{kt}^\star = a_{kt} * \sigma_t \text{ and } b_{kt}^\star = \frac{b_{kt} - \mu_t}{\sigma_t}. \tag{2.5}$$

As can be seen in the ICC plots for both sets, the generating procedure succeeded in producing data which are representative of longitudinal studies of substance abuse or internalizing behavior, in which the phenomenon of interest is rarely observed early on and items are consequently difficult to endorse and not strongly discriminating, while later in observation the phenomenon is more common and items appear less difficult but more highly discriminating. Though both parameter sets were reflective of this process, they differed in the range of $\boldsymbol{\theta}$ measured by the scale. The first parameter set was representative of a scale which measured a much more narrow range of the latent construct, while the second parameter set measured a much broader range of the latent construct. Because the number of observations and items were held constant across sets,

expanding the range of the latent construct being measured resulted in a sparseness of the response patterns under set 2.

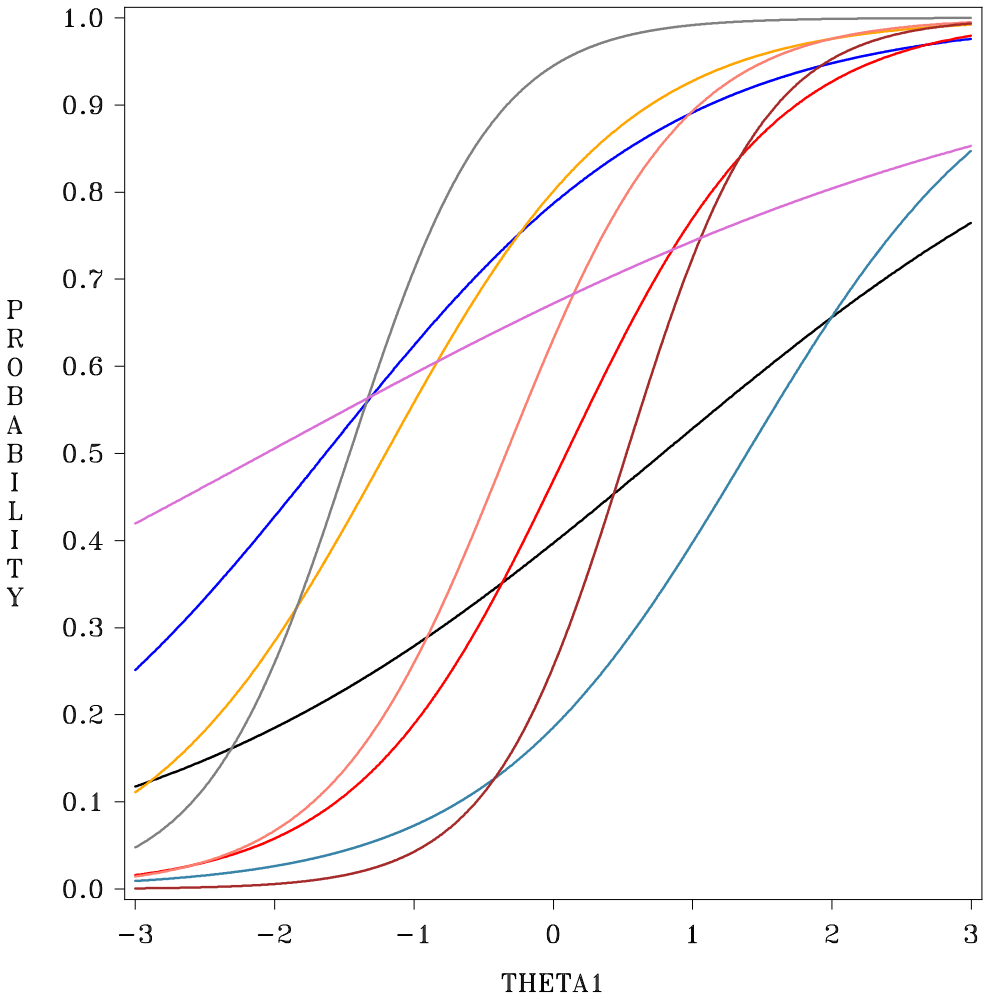

Figure 2.1: ICCs for Parameter Set 1 at Time 1

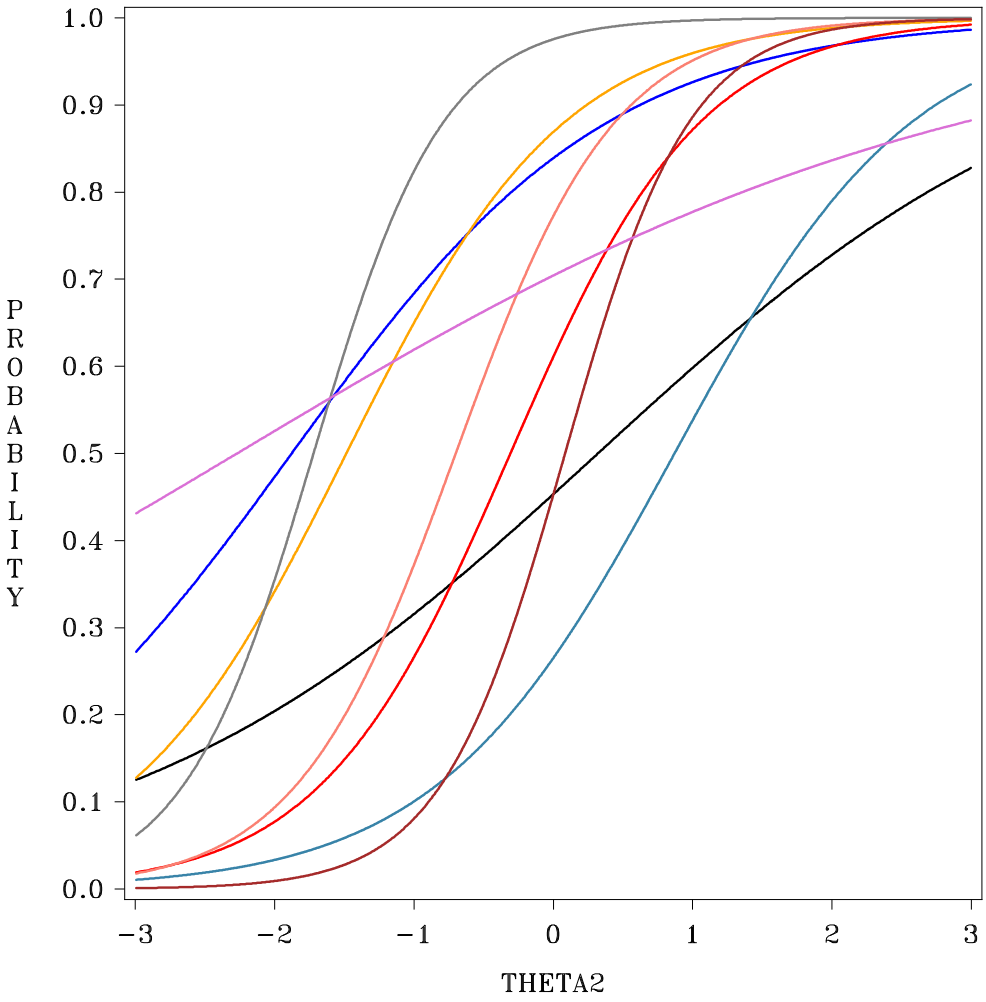

Figure 2.2: ICCs for Parameter Set 1 at Time 2

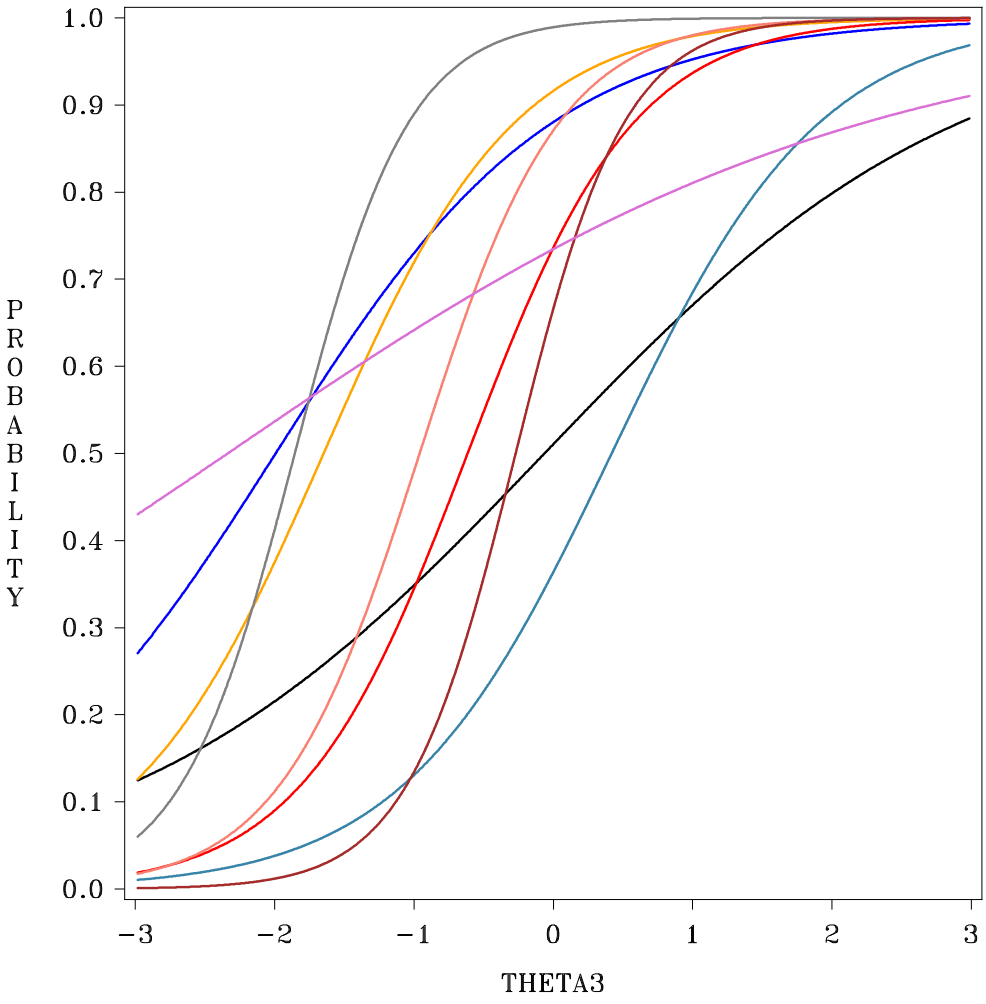

Figure 2.3: ICCs for Parameter Set 1 at Time 3

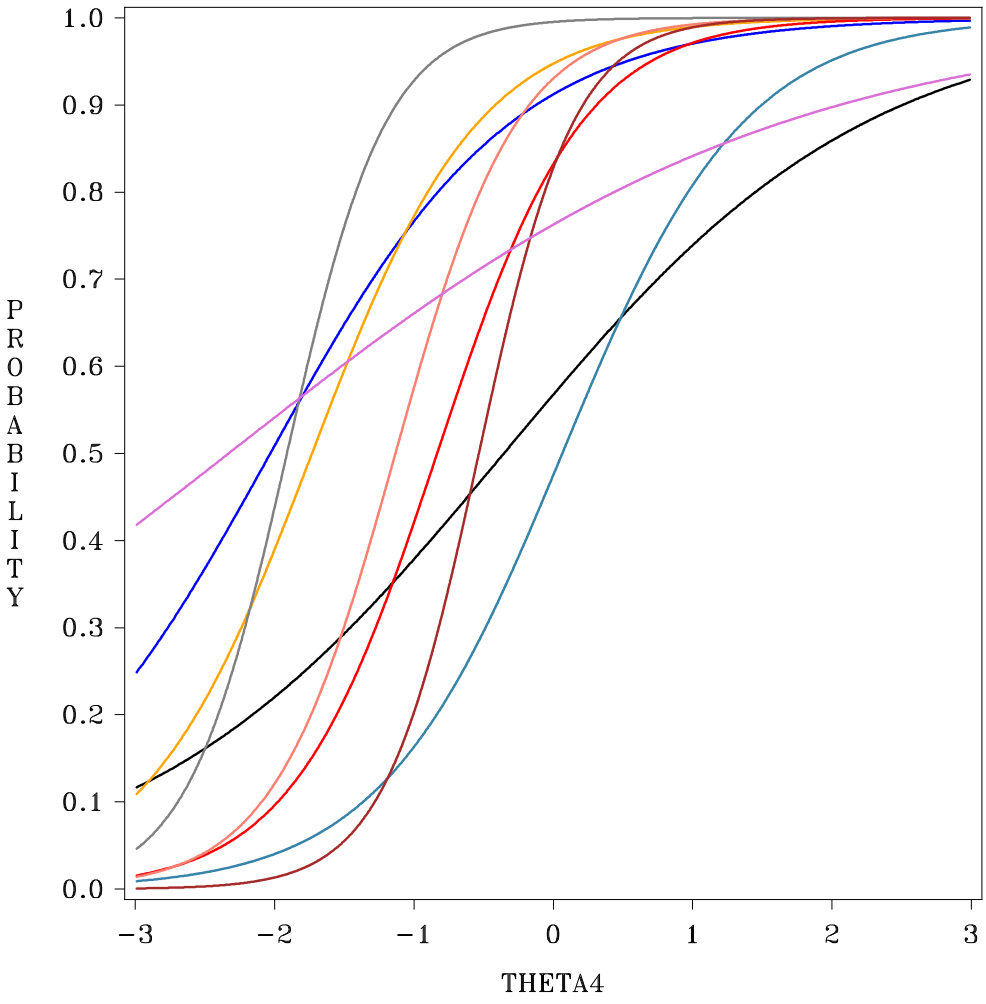

Figure 2.4: ICCs for Parameter Set 1 at Time 4

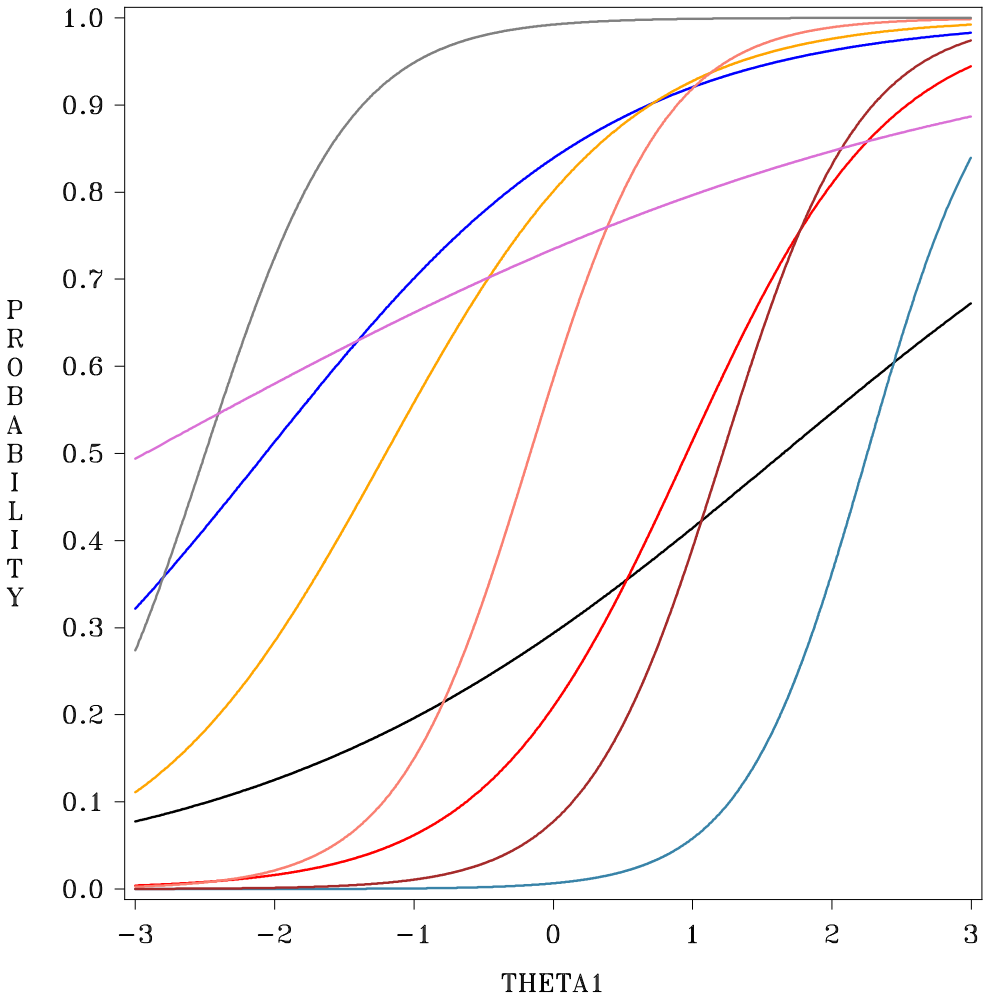

Figure 2.5: ICCs for Parameter Set 2 at Time 1

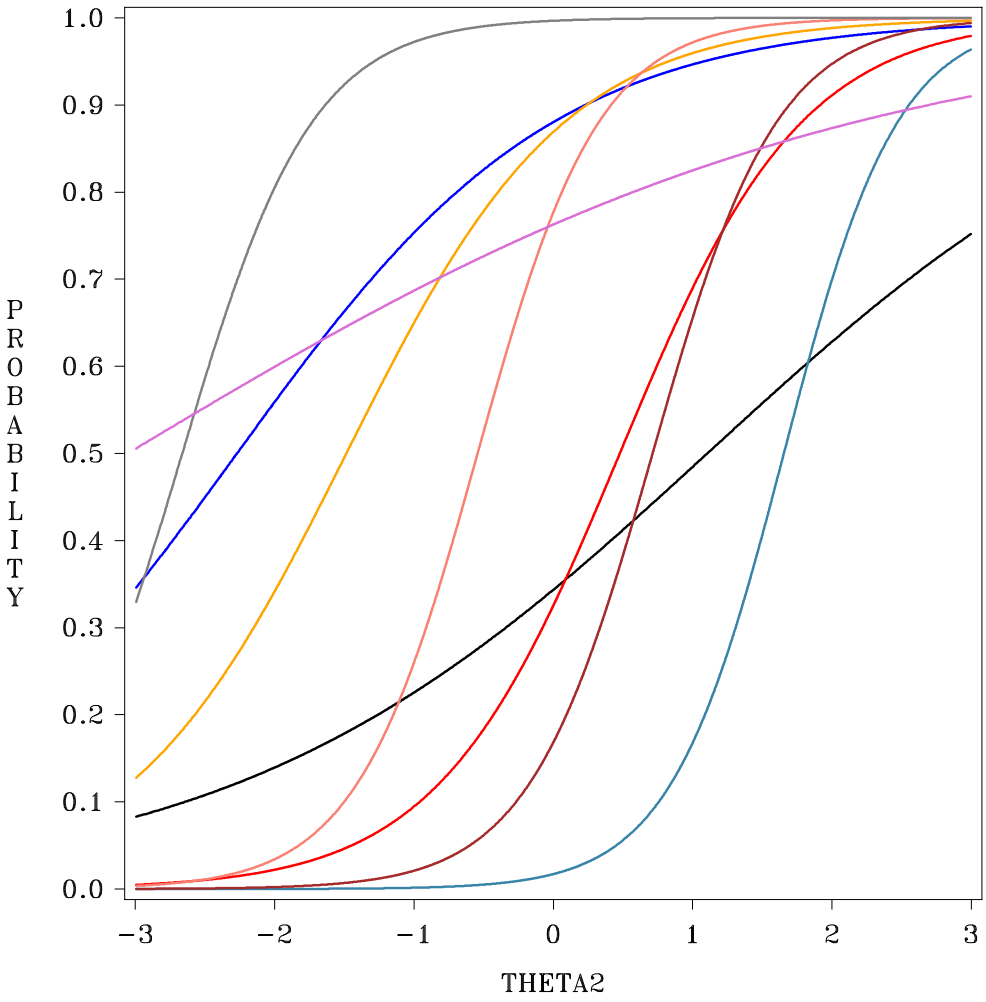

Figure 2.6: ICCs for Parameter Set 2 at Time 2

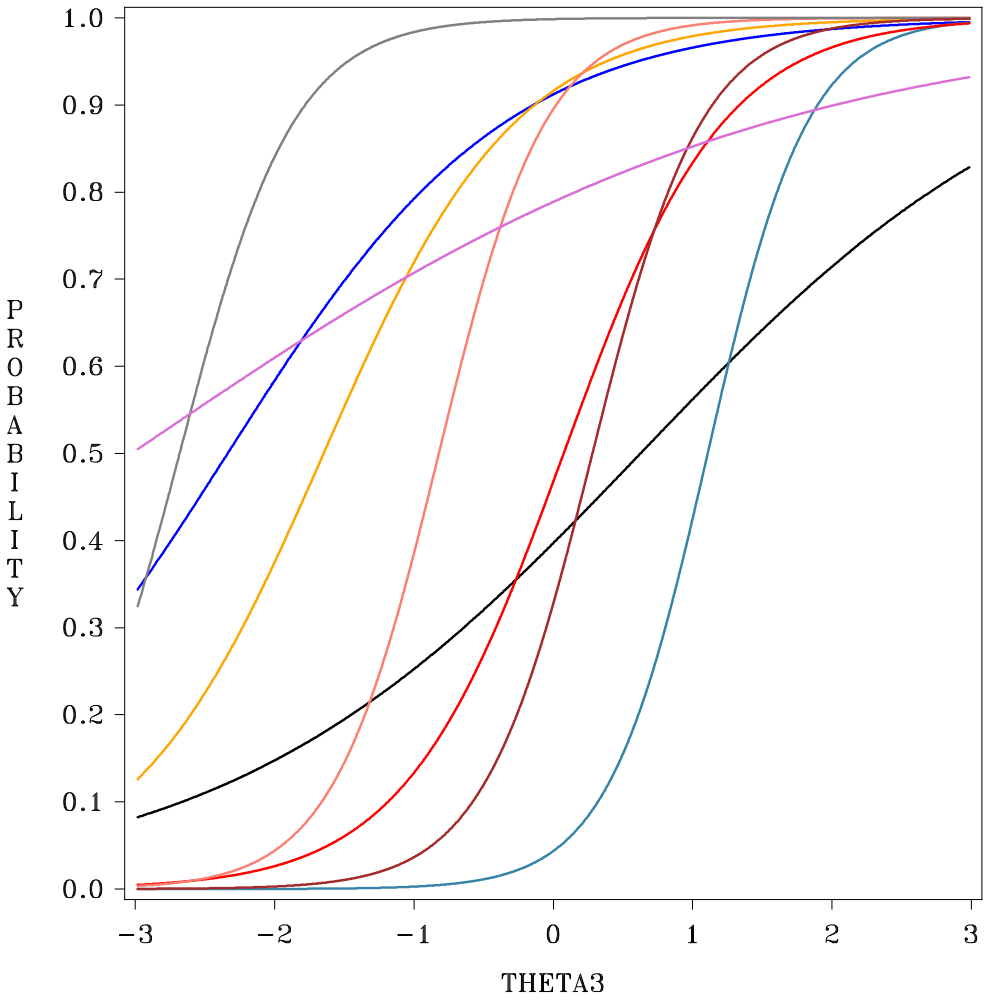

Figure 2.7: ICCs for Parameter Set 2 at Time 3

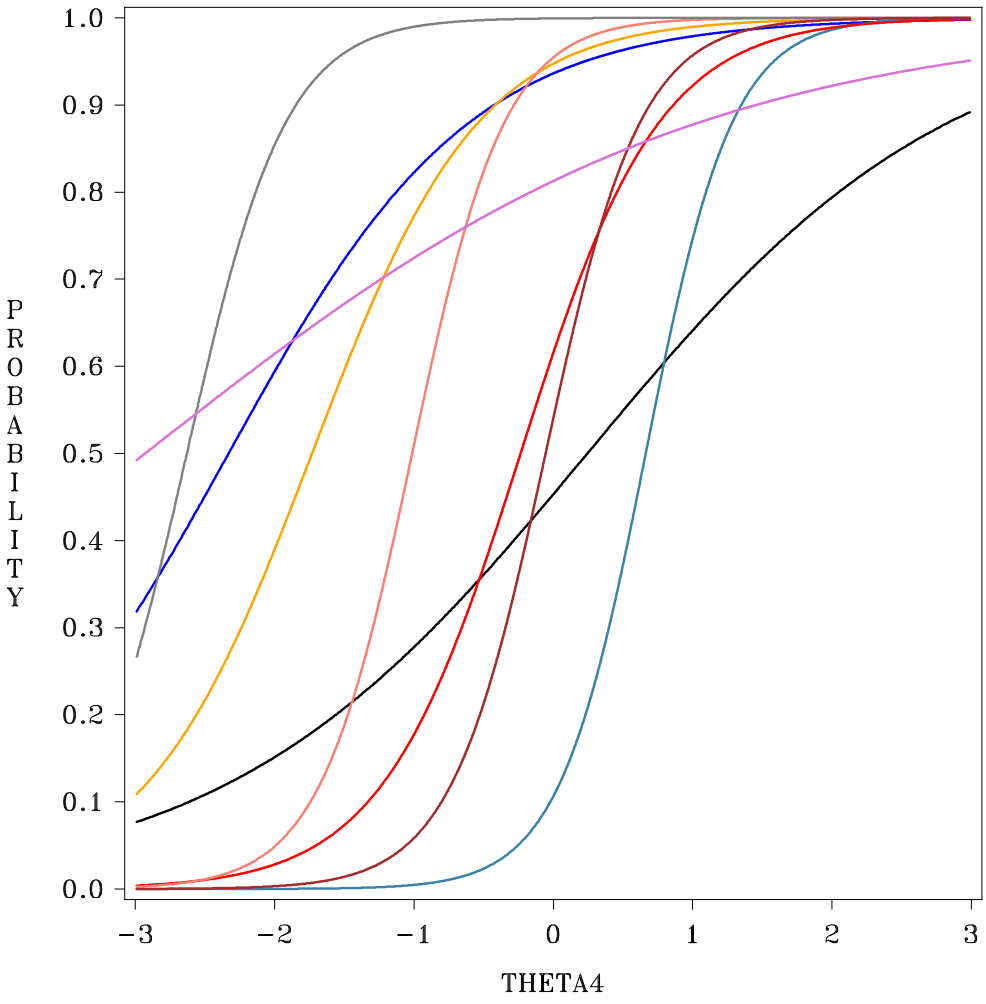

Figure 2.8: ICCs for Parameter Set 2 at Time 4

2.0.9 Implied Moments

The generating parameters given in Table 2.1 coupled with the coding of $\boldsymbol{\lambda}$ can be used to calculate the implied moments of $\boldsymbol{\theta}$. As stated, the implied mean vector may be calculated as $\boldsymbol{\mu_\theta} = \boldsymbol{\lambda\mu_\eta}$ and the implied covariance matrix may be calculated as $\boldsymbol{\Sigma_\theta} = \boldsymbol{\lambda\tau_\eta\lambda}' + \boldsymbol{\psi}$.

This resulted in

$$\boldsymbol{\mu_\theta} = \begin{bmatrix} 1.392 & 1.892 & 2.392 & 2.892 \end{bmatrix} \tag{2.6}$$

and

$$\boldsymbol{\Sigma_\theta} = \begin{bmatrix} 1.34 & - & - & - \\ 0.7157575 & 1.6230301 & - & - \\ 0.761515 & 0.9072725 & 2.1060601 & - \\ 0.8072725 & 1.0030301 & 1.1987876 & 2.7890902 \end{bmatrix}, \tag{2.7}$$

with corresponding correlation matrix

$$\mathbf{R_\theta} = \begin{bmatrix} 1 & - & - & - \\ 0.4853446 & 1 & - & - \\ 0.4533052 & 0.490726 & 1 & - \\ 0.4175769 & 0.4714324 & 0.4946244 & 1 \end{bmatrix}. \tag{2.8}$$

Because the structural parameters were held fixed across parameter sets, the implied moments given here are the same for both parameter sets.

2.0.10 Sample Size Manipulation

All cells of the simulation were examined under two fixed sample sizes: $N = 750$ and $N = 3000$. These sample sizes were chosen in order to minimize optimization problems for WLSMV estimation under the wide difficulty range condition and because they were representative of the large samples common in IRT applications. The lower sample size of $N = 750$ was selected so as to be representative of what we considered to be a minimally sufficient sample size for the estimation of the models of interest.

2.0.11 Non-Convergence

For every cell of the design, non-converged solutions were identified. Within a model and parameter set combination, replications were fixed across estimator within sample size cells. Thus, 300 replications were generated when $N = 750$ and each of these replications was used in each estimator. Every unique replication which did not converge across estimators was identified and within a sample size non-converged replications were deleted and resampling of replications was conducted until a total of 300 converged solutions were obtained in each sample size.

2.0.12 Improper Solutions

Following the work of Chen, Bollen, Paxton, Curran, and Kirby (2001) I differentiated the importance of improper solutions (non-positive definite (NPD) covariance matrices and negative variance estimates) based on the magnitude of the departure of the parameter from its support. Thus, variance parameters for which estimates were only slightly negative, or covariance matrices with eigenvalues that were trivially negative or near zero were deemed to result from sampling variation, particularly when they were associated with generating values which were themselves close to the boundary of parametric support. In addition, even when point estimates appear proper, the covariance matrix may be degenerate, having a negative eigenvalue. Consequently, all variance estimates were screened for degeneracy and replications having trivially degenerate solutions were identified and described. These replications were not removed from analysis as this could have induced a selection bias in the outcomes of interest. Rather, analyses were conducted with results from all 300 converged replications. However, in order to assess the potential impact of the degenerate solutions, sensitivity analyses, in which models for bias were re-run omitting replications with improper solutions, were conducted to assess any impact of degeneracy on conclusions.

CHAPTER 3

Results

To more clearly distill the experimental manipulation, and the fractional design employed, effects of interest related to sample size, estimator, model, and difficulty range width are outlined in Table 3.1. This table also reflects the order with which results are presented. First I will present results for Model 1 for the first and then second item parameter sets, focussing on differences across sample size and estimator cells of the design. This is then followed by the same presentation for Model 2. Convergence and NPD solutions are discussed first. Bias and meta-model results are then discussed in detail. Contrasts of interest in the meta-models include the sample size main effect, the estimator main effects, and all sample size by estimator interactions. As detailed in the results section, because patterns observed for RMSE did not differ from those for bias, RMSE is not examined in detail, and results are relegated to appendices.

Table 3.1: Experimental Design

		$N = 750$ & $N = 3000$		
		Estimator		
Model	Set	Limited Information	Full Information	
1	1	$WLSMV$	$QP = 3$	$QP = 7$
1	2	$WLSMV$	$QP = 3$	$QP = 7$
2	1	$WLSMV$	$MCEM$	–
2	2	$WLSMV$	$MCEM$	–

3.1 Model 1

3.1.1 Convergence

Non-converged replications were encountered only when estimation was based on full information with three quadrature points per dimension. Of the 300 replications simulated per cell, 40 failed to converge when $N = 750$, and 17 failed to converge when $N = 3000$ under item parameter set 1. Under item parameter set 2, when $N = 750$ a total of 27 replications failed to converge, and when $N = 3000$ a total of 5 replications failed to converge. The converged solutions corresponding to these replications were deleted from the 7 quadrature point and WLSMV solutions. An additional 100 replications were generated in order to replace the failed or deleted replications. This produced solutions corresponding to 300 identical replications across all analyzed cells.

3.1.2 NPD Solutions

Under item parameter set 1, when $N = 750$ a total of 36 unique replications had degenerate variance estimates. All estimates were associated with τ_η. A total of 8 replications were common to all estimators, 7 were common to FIML based on 3 QP

and WLSMV, 4 were common to FIML based on 7 QP and WLSMV. A total of 16 were unique to WLSMV, and 1 was unique to FIML based on 3 QP. When $N = 3000$ a total of 2 unique replications had degenerate variance estimates. One degenerate replication occurred under WLSMV, and the other was observed for FIML based on 3 QP.

Under item parameter set 2, when $N = 750$ a total of 34 unique replications had improper solutions for variance estimates. A total of 3 replications were common to all estimators, 4 replications were shared between FIML based on 3 QP and WLSMV, 5 replications were common to FIML based on 7 QP and WLSMV, one replication was shared between 3 and 7 QP only. A single replication was unique to 3 QP, likewise for 7 QP, while a total of 19 replications were unique to WLSMV. When $N = 3000$ no replications had degenerate solutions under either FIML estimator, and only 2 replications produced improper solutions under WLSMV.

Across both parameter sets, the majority of degenerate solutions were associated with estimates of τ_β, whose generating value, .05, was near the boundary of the parametric support. Consequently, sampling variations likely account for a large number of the degenerate solutions.

3.1.3 First Parameter Set Bias

Examination of the results presented in Table 3.2 reveals that, aside from the covariance between the random intercept and slope, all estimators exhibit minimal degrees of bias. Column heading P contains the parameters, heading θ contains the population values, heading $\hat{\theta}(SD)$ contains the estimate averaged over the 300 replications and corresponding standard deviation, $B_{\hat{\theta}}$ contains the bias of the estimate, and $RB_{\hat{\theta}}$ contains the relative bias of the estimate obtained by dividing bias by the population value. Values of $\hat{\theta}(SD)$, $B_{\hat{\theta}}$, and $RB_{\hat{\theta}}$ are given for each estimator. Tables are organized by blocking on estimator, with the 3 QP FIML estimator, denoted $QP = 3$, followed by the 7 QP FIML estimator, denoted $QP = 7$, with WLSMV last.

Notable trends include the elevated bias of WLSMV relative to FIML, the equality of

3 and 7 quadrature points, and the preponderance of negative bias exhibited by WLSMV estimates of $\hat{b}_i$. Nearly identical results are observed in Table 3.3, suggesting the absence of a sample size effect, even though sample size was quadrupled under this condition. It is important to note that though there appear to be some differences in degree of bias, nearly all of the bias observed is so low as to be of little consequence. In point of fact, the highest relative bias observed corresponds to $\tau_{\alpha\beta}$, which ranges from 12.3% to 22.1% bias, however, examination of the raw bias and average point estimate reveals that there is very little discrepancy between the population parameter and the estimate. This is the case for every parameter: degrees of bias were acceptably low based on the ±10% relative bias criterion. Throughout, estimates for which relative bias met or exceeded ±10% have their values of and $RB_{\hat{\theta}}$ placed in bold text.

Table 3.2: Item and Structural Parameter Bias for Model 1, Set 1, N=750

N=750										
		$QP = 3$			$QP = 7$			$WLSMV$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.46(0.05)	0.003	0.006	0.46(0.04)	-0.003	-0.006	0.50(0.05)	0.037	0.081
a_2	0.69	0.70(0.07)	0.005	0.007	0.69(0.07)	-0.002	-0.003	0.69(0.07)	0.002	0.002
a_4	0.92	0.94(0.08)	0.016	0.018	0.93(0.08)	0.005	0.006	0.96(0.10)	0.043	0.046
a_5	1.15	1.16(0.09)	0.009	0.008	1.15(0.09)	0.002	0.002	1.20(0.11)	0.054	0.047
a_6	1.37	1.38(0.12)	0.012	0.009	1.37(0.12)	0.003	0.002	1.41(0.13)	0.043	0.031
a_7	1.68	1.72(0.17)	0.041	0.024	1.69(0.16)	0.012	0.007	1.66(0.21)	-0.020	-0.012
a_8	1.76	1.75(0.13)	-0.011	-0.006	1.76(0.14)	-0.004	-0.002	1.78(0.17)	0.021	0.012
a_9	0.30	0.30(0.04)	-0.001	-0.004	0.30(0.04)	-0.005	-0.015	0.31(0.05)	0.014	0.046
b_1	2.30	2.31(0.12)	0.013	0.006	2.31(0.12)	0.013	0.006	2.26(0.12)	-0.035	-0.015
b_2	-0.50	-0.50(0.19)	-0.001	0.002	-0.53(0.20)	-0.029	0.058	-0.55(0.22)	-0.049	0.098
b_4	3.00	3.00(0.14)	-0.003	-0.001	3.00(0.14)	0.002	0.001	2.94(0.14)	-0.065	-0.022
b_5	1.50	1.51(0.07)	0.006	0.004	1.50(0.07)	0.000	0.000	1.48(0.07)	-0.016	-0.011
b_6	1.00	1.01(0.07)	0.007	0.007	1.00(0.07)	-0.003	-0.003	1.00(0.08)	-0.003	-0.003
b_7	-0.30	-0.28(0.13)	0.021	-0.069	-0.31(0.13)	-0.008	0.028	-0.28(0.15)	0.019	-0.063
b_8	2.00	2.00(0.08)	0.005	0.002	2.00(0.08)	0.002	0.001	1.97(0.07)	-0.033	-0.017
b_9	-1.00	-1.05(0.41)	-0.055	0.055	-1.09(0.42)	-0.093	0.093	-1.09(0.45)	-0.085	0.085
μ_α	1.39	1.40(0.07)	0.005	0.004	1.39(0.07)	0.000	0.000	1.38(0.07)	-0.010	-0.007
μ_β	0.50	0.50(0.04)	0.002	0.004	0.50(0.04)	0.004	0.009	0.48(0.04)	-0.019	-0.038
τ_α	0.67	0.69(0.13)	0.017	0.025	0.69(0.13)	0.021	0.032	0.66(0.14)	-0.012	-0.019
$\boldsymbol{\tau_{\alpha\beta}}$	0.05	0.04(0.04)	-0.007	**−0.161**	0.04(0.04)	-0.006	**−0.123**	0.04(0.05)	-0.010	**−0.214**
τ_β	0.05	0.05(0.03)	0.003	0.060	0.05(0.03)	0.004	0.074	0.05(0.03)	-0.001	-0.025
ψ_1	0.67	0.64(0.15)	-0.031	-0.046	0.67(0.15)	0.002	0.002	0.63(0.15)	-0.041	-0.062
ψ_2	0.81	0.79(0.14)	-0.024	-0.029	0.81(0.14)	0.002	0.003	0.76(0.15)	-0.049	-0.061
ψ_3	1.05	1.05(0.19)	-0.006	-0.005	1.08(0.19)	0.024	0.022	0.99(0.19)	-0.061	-0.058
ψ_4	1.39	1.38(0.28)	-0.013	-0.010	1.42(0.28)	0.021	0.015	1.31(0.28)	-0.086	-0.061

Table 3.3: Item and Structural Parameter Bias for Model 1, Set 1, N=3000

N=3000										
		$QP = 3$			$QP = 7$			$WLSMV$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.46(0.02)	0.005	0.010	0.46(0.02)	-0.001	-0.002	0.50(0.03)	0.041	0.089
a_2	0.69	0.70(0.03)	0.006	0.008	0.69(0.03)	-0.002	-0.002	0.69(0.03)	0.003	0.004
a_4	0.92	0.93(0.04)	0.010	0.011	0.92(0.04)	0.000	0.000	0.95(0.04)	0.034	0.036
a_5	1.15	1.16(0.05)	0.007	0.006	1.15(0.05)	0.000	0.000	1.20(0.05)	0.053	0.046
a_6	1.37	1.38(0.06)	0.010	0.007	1.37(0.06)	0.001	0.000	1.41(0.07)	0.042	0.031
a_7	1.68	1.71(0.09)	0.027	0.016	1.68(0.09)	-0.002	-0.001	1.60(0.09)	-0.085	-0.050
a_8	1.76	1.76(0.07)	-0.004	-0.002	1.76(0.08)	0.003	0.002	1.79(0.09)	0.028	0.016
a_9	0.30	0.30(0.02)	0.002	0.006	0.30(0.02)	-0.002	-0.005	0.32(0.02)	0.017	0.058
b_1	2.30	2.30(0.07)	0.003	0.001	2.30(0.07)	0.003	0.001	2.25(0.06)	-0.051	-0.022
b_2	-0.50	-0.48(0.09)	0.019	-0.037	-0.51(0.09)	-0.009	0.018	-0.53(0.10)	-0.026	0.052
b_4	3.00	3.00(0.08)	-0.005	-0.002	3.00(0.08)	0.001	0.000	2.93(0.07)	-0.072	-0.024
b_5	1.50	1.50(0.04)	0.002	0.002	1.50(0.04)	-0.003	-0.002	1.48(0.03)	-0.022	-0.015
b_6	1.00	1.01(0.03)	0.009	0.009	1.00(0.03)	0.000	0.000	1.00(0.03)	0.000	0.000
b_7	-0.30	-0.27(0.07)	0.026	-0.085	-0.30(0.07)	-0.004	0.012	-0.30(0.07)	0.004	-0.014
b_8	2.00	2.00(0.04)	0.002	0.001	2.00(0.04)	0.000	0.000	1.96(0.04)	-0.039	-0.020
b_9	-1.00	-0.99(0.19)	0.008	-0.008	-1.03(0.19)	-0.029	0.029	-1.01(0.21)	-0.012	0.012
μ_α	1.39	1.40(0.04)	0.004	0.003	1.39(0.04)	-0.001	-0.001	1.38(0.03)	-0.012	-0.009
μ_β	0.50	0.50(0.02)	-0.002	-0.004	0.50(0.02)	0.001	0.001	0.48(0.02)	-0.025	-0.049
τ_α	0.67	0.67(0.07)	-0.001	-0.002	0.67(0.07)	0.002	0.004	0.64(0.07)	-0.033	-0.049
$\tau_{\alpha\beta}$	0.05	0.04(0.02)	-0.003	-0.071	0.04(0.02)	-0.001	-0.024	0.04(0.02)	-0.010	$\mathbf{-0.221}$
τ_β	0.05	0.05(0.01)	0.000	0.009	0.05(0.02)	0.001	0.018	0.05(0.02)	-0.004	-0.075
ψ_1	0.67	0.64(0.07)	-0.028	-0.042	0.67(0.08)	0.005	0.007	0.63(0.07)	-0.044	-0.066
ψ_2	0.81	0.79(0.07)	-0.022	-0.027	0.82(0.08)	0.005	0.006	0.75(0.08)	-0.058	-0.072
ψ_3	1.05	1.03(0.10)	-0.020	-0.019	1.06(0.10)	0.010	0.009	0.97(0.10)	-0.081	-0.077
ψ_4	1.39	1.36(0.13)	-0.033	-0.023	1.40(0.14)	0.004	0.003	1.26(0.13)	-0.132	-0.094

3.1.4 First Parameter Set Meta Model Results

Tables 3.2 and 3.3 present the descriptive statistics for model 1, set 1. I next employed a series of meta models to formally test my proposed hypotheses in regards to raw bias. The model of interest characterized main effects for sample size (750 Vs. 3000) and estimator (3 vs. 7 QP, $3QP$ Vs. WLSMV, $7QP$ Vs. WLSMV, and all FIML Vs. WLSMV), as well as all two-way interactions. The model was fit separately for each parameter presented in Tables 3.2 and 3.3. In the case of item parameters and the residual variances contained in $\boldsymbol{\psi}$, bias was aggregated over parameters within a

parameter class. Thus rather than fit a model to the bias of each individual a parameter, a parameter bias was aggregated and the average a bias was modeled. The decision to aggregate was based on the desire for parsimony and the fact that examination of the results for individual parameters within these classes indicated little deviation from the trend observed in aggregate. The meta model tables for the item parameters and residual variance include a block of columns indicating any discrepancies observed in the individual-parameter models. Asterisks indicate correspondence to the aggregate results, while dashes indicate divergence from the aggregate results. Complete results for the parameter-specific models for every cell of the entire simulation design are presented in a PDF posted at

`www.unc.edu\~curran\serrano.pdf`

.

Readers interested in augmenting their understanding of the direction of effects are referred to Table 5.7, which contains the cell means from which all contrasts parameterized in the meta models may be computed. This table, and the cell means tables for all analyses are contained in Appendix 3. Cell means for each parameter are given, though only aggregate cell means are given for the a, b, and ψ parameters.

Results presented in Table 3.4 for aggregated $\hat{a}$ raw bias reveal an absence of sample size effects and interactions, but the hypothesized estimator effects were observed. The direction of effects was consistent with those hypothesized: Estimates based on 7 QP were significantly less biased than 3 QP, though bias was only lower by .008 units ($\beta(95\%CI) = -0.008(-0.015, -0.002), t = -2.5, p \leq .05$). FIML estimates obtained under 3 QP were .012 units less biased than WLSMV ($\beta(95\%CI) = -0.012(-0.018, -0.005), t = -3.6, p \leq .001$). Likewise, FIML estimates based on 7 QP exhibited bias .02 units lower than WLSMV estimates ($\beta(95\%CI) = -0.020(-0.026, -0.014), t = -6.1, p \leq .0001$). Lastly, the average of 3 and 7 QP, comprising the aggregate FIML effect, was less biased

than WLSMV ($\beta(95\%CI) = -0.016(-0.021, -0.010), t = -5.6, p \leq .0001$). This model accounted for 2% of the variance in bias.

There were few individual items which diverged from the aggregate findings. Notable exceptions included a_2 and a_7. In the case of the former, a lack of estimator main effects was observed, while for the latter, the sample-size main effect was observed, along with all 2-way interactions. In addition, 50% of the parameters did not exhibit a difference between FIML based on 3 and 7 QP.

Table 3.4: Meta Model Results: $\hat{a}$ Raw Bias for Model 1, Set 1

Raw $\hat{a}$ Bias										
Results when $\hat{a}_i$ Aggregated			Correspondence to Aggregate by $\hat{a}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	a_1	a_2	a_4	a_5	a_6	a_7	a_8	a_9
$1: N3k - N750$	$-0.004(-0.009, 0.002)$	$-1.3, NS$	*	*	-	*	*	-	*	*
$2: 3QP - 7QP$	$-0.008(-0.015, -0.002)$	$-2.5, p \leq .05$	*	*	*	-	-	*	-	-
$3: 3QP - WLSMV$	$-0.012(-0.018, -0.005)$	$-3.6, p \leq .001$	*	-	*	*	*	*	*	*
$4: 7QP - WLSMV$	$-0.020(-0.026, -0.014)$	$-6.1, p \leq .0001$	*	-	*	*	*	*	*	*
$5: FIML - WLSMV$	$-0.016(-0.021, -0.010)$	$-5.6, p \leq .0001$	*	-	*	*	*	*	*	*
$6: 1 \times 2$	$0.000(-0.013, 0.013)$	$-0.0, NS$	*	*	*	*	*	*	*	*
$7: 1 \times 3$	$-0.006(-0.019, 0.007)$	$-0.9, NS$	*	*	*	*	*	-	*	*
$8: 1 \times 4$	$-0.006(-0.019, 0.007)$	$-0.9, NS$	*	*	*	*	*	-	*	*
$9: 1 \times 5$	$-0.006(-0.017, 0.005)$	$-1.0, NS$	*	*	*	*	*	-	*	*

With the exception of the significant sample size effect, all results for the aggregated $\hat{b}$ bias matched those observed for the aggregated $\hat{a}$. However, closer inspection of the point estimates contained in Table 3.5 suggests an opposite direction of effect for the estimator contrasts, indicating WLSMV was less biased for $\hat{b}$ compared to both FIML estimators. Examination of the estimator cell means helps understand this counterintuitive finding. Aggregate bias for $\hat{b}$ based on WLSMV was negative, as was that based on 7 QP, though it was positive for estimates based on 3 QP. Given these cell means, the estimator contrasts may be re-constructed using WLSMV as the reference. Though the direction of effects for contrasts in Table 3.5 suggest WLSMV is less biased, it is clear from the cell means that WLSMV is more biased, and that the di-

rection of effect is obscured by the larger negative bias of WLSMV during the calculation of differences in cell means. The same issue was observed for the sample size effect. On average, bias was negative; however bias was larger when $N = 750$ than when $N = 3000$. Thus, bias was significantly larger when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = 0.008(0.002, 0.014), t = 2.7, p \leq .01$). Bias was significantly lower when 3 QP were employed compared to 7 QP($\beta(95\%CI) = -0.013(-0.020, -0.006), t = -3.6, p <= .001$). Estimates based on 3 QP were significantly less biased than WLSMV estimates ($\beta(95\%) = 0.034(0.027, 0.041), t = 9.3, p <= .0001$), as was the case when estimates were based on 7 QP($\beta(95\%CI) = 0.021(0.013, 0.028), t = 5.6, p <= .0001$). In addition, the aggregate FIML bias was less than that of WLSMV ($\beta(95\%CI) = 0.027(0.021, 0.033)t = 8.6, p <= .0001$). This model accounted for 5% of the variance in bias.

As with the estimates of $\hat{a}$, results from the parameter-specific models reveal some discrepancies relative to the aggregate results. Specifically, there was an absence of the sample size effect in all but three parameters (b_1, b_2, and b_9). Moreover, the difference between bias for 3 and 7 QP estimates was absent for all but b_2, b_6, and b_7. One parameter in particular, b_9 diverges substantially, not exhibiting any of the estimator effects.

Table 3.5: Meta Model Results: $\hat{b}$ Raw Bias for Model 1, Set 1

Raw $\hat{b}$ Bias										
Results when $\hat{b}_i$ Aggregated			Correspondence to Aggregate by $\hat{b}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	b_1	b_2	b_4	b_5	b_6	b_7	b_8	b_9
$1 : N3k - N750$	$0.008(0.002, 0.014)$	$2.7, p \leq .01$	*	*	-	-	-	-	-	*
$2 : 3QP - 7QP$	$-0.013(-0.020, -0.006)$	$-3.6, p \leq .001$	-	*	-	-	*	*	-	-
$3 : 3QP - WLSMV$	$0.034(0.027, 0.041)$	$9.3, p \leq .0001$	*	*	*	*	*	-	*	-
$4 : 7QP - WLSMV$	$0.021(0.013, 0.028)$	$5.6, p \leq .0001$	*	*	*	*	-	*	*	-
$5 : FIML - WLSMV$	$0.027(0.021, 0.033)$	$8.6, p \leq .0001$	*	*	*	*	-	-	*	-
$6 : 1 \times 2$	$0.000(-0.015, 0.014)$	$0.0, NS$	*	*	*	*	*	*	*	*
$7 : 1 \times 3$	$-0.003(-0.017, 0.012)$	$-0.4, NS$	*	*	*	*	*	*	*	*
$8 : 1 \times 4$	$-0.003(-0.017, 0.011)$	$-0.4, NS$	*	*	*	*	*	*	*	*
$9 : 1 \times 5$	$-0.003(-0.015, 0.010)$	$-0.4, NS$	*	*	*	*	*	*	*	*

Design effects observed for fixed effect bias, presented in Table 3.6, are nearly identical across μ_α and μ_β, the primary difference being the existence of a sample-size effect for μ_β only. Otherwise, results confirmed that the negative bias of WLSMV exceeded that of both FIML estimators, both individually, and in aggregate. When considering μ_α, FIML based on 3 QP was, on average, less biased than WLSMV by .02 units($\beta(95\%CI) = 0.02(0.01, 0.02), t = 4.7, p <= .0001$). Likewise, FIML based on 7 QP was less biased than WLSMV by .01 units ($\beta(95\%CI) = 0.01(0.00, 0.02), t = 3.2, p <= .01$). In aggregate, FIML was less biased than WLSMV ($\beta(95\%CI) = 0.01(0.01, 0.02), t = 4.5, p <= .0001$). This model accounted for 1.3% of the variance in bias. In contrast, bias for μ_β was significantly, though trivially, lower when $N = 750$ versus $N = 3000$ ($\beta(95\%CI) = -0.004(-0.01, -0.001), t = -3.1, p <= .01$). Estimates obtained from 3 QP were significantly less biased than WLSMV estimates ($\beta(95\%CI) = 0.02(0.02, 0.03), t = 12.1, p <= .0001$), as were 7 QP estimates ($\beta(95\%CI) = 0.02(0.02, 0.03), t = 13.4, p <= .0001$). Consistent with these results, the aggregate FIML bias was significantly lower than that observed for WLSMV ($\beta(95\%CI) = 0.02(0.02, 0.03), t = 14.8, p <= .0001$). This model accounted for 11.3% of the variance in bias.

Table 3.6: Meta Model Results: Fixed Effect Raw Bias for Model 1, Set 1

Raw Bias				
	$\hat{\mu}_\alpha$		$\hat{\mu}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1 : N3k - N750$	$-0.00(-0.01, 0.00)$	$-0.5, NS$	$-0.00(-0.01, -0.00)$	$-3.1, p \leq .01$
$2 : 3QP - 7QP$	$-0.01(-0.01, 0.00)$	$-1.5, NS$	$0.00(0.00, 0.01)$	$1.3, NS$
$3 : 3QP - WLSMV$	$0.02(0.01, 0.02)$	$4.7, p \leq .0001$	$0.02(0.02, 0.03)$	$12.1, p \leq .0001$
$4 : 7QP - WLSMV$	$0.01(0.00, 0.02)$	$3.2, p \leq .01$	$0.02(0.02, 0.03)$	$13.4, p \leq .0001$
$5 : FIML - WLSMV$	$0.01(0.01, 0.02)$	$4.5, p \leq .0001$	$0.02(0.02, 0.03)$	$14.8, p \leq .0001$
$6 : 1 \times 2$	$0.00(-0.01, 0.01)$	$0.0, NS$	$0.00(-0.01, 0.01)$	$0.0, NS$
$7 : 1 \times 3$	$-0.00(-0.01, 0.01)$	$-0.3, NS$	$-0.00(-0.01, 0.01)$	$-0.6, NS$
$8 : 1 \times 4$	$-0.00(-0.01, 0.01)$	$-0.3, NS$	$-0.00(-0.01, 0.00)$	$-0.6, NS$
$9 : 1 \times 5$	$-0.00(-0.01, 0.01)$	$-0.3, NS$	$-0.00(-0.01, 0.00)$	$-0.7, NS$

Results for the model of random effect variance bias are presented in Table 3.7. In the case of τ_α, estimates were less biased when $N = 750$ than when $N = 3000$ ($\beta(95\%CI) = -0.02(-0.03, -0.01), t = -3.8, p \leq .001$). Estimates based on 3 QP were less biased than WLSMV ($\beta(95\%CI) = 0.03(0.02, 0.04), t = 4.9, p <= .0001$). The 7 QP estimates were less biased than WLSMV estimates $\beta(95\%CI) = 0.03(0.02, 0.05), t = 5.5, p <= .0001$). The aggregate FIML bias was significantly lower than the WLSMV bias $\beta(95\%CI) = 0.03(0.02, 0.04), t = 6.0, p <= .0001$). This model accounted for 3% of the variance in bias.

In contrast to τ_α, τ_β estimates were less biased when $N = 3000$ than when $N = 750$ ($\beta(95\%CI) = -0.002(-0.004, -0.0004), t = -2.4, p <= .05$). FIML estimates based on 3 QP were less biased than WLSMV estimates $\beta(95\%CI) = 0.004(0.001, 0.01), t = 3.1, p <= .01$), as were estimates based on 7 QP $\beta(95\%CI) = 0.004(0.002, 0.01), t = 3.5, p <= .001$). In addition, the average FIML estimate was less biased than WLSMV $\beta(95\%CI) = 0.004(0.002, 0.01), t = 3.8, p <= .001$). This model accounted for 1% of the variance in bias.

Table 3.7: Meta Model Results: Random Effect Variance Estimate Raw Bias for Model 1, Set 1

Raw Bias				
	$\hat{\tau}_\alpha$		$\hat{\tau}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.02(-0.03, -0.01)$	$-3.8, p \leq .001$	$-0.00(-0.00, -0.00)$	$-2.4, p \leq .05$
$2: 3QP - 7QP$	$0.00(-0.01, 0.02)$	$0.6, NS$	$0.00(0.00, 0.00)$	$0.4, NS$
$3: 3QP - WLSMV$	$0.03(0.02, 0.04)$	$4.9, p \leq .0001$	$0.00(0.00, 0.01)$	$3.1, p \leq .01$
$4: 7QP - WLSMV$	$0.03(0.02, 0.05)$	$5.5, p \leq .0001$	$0.00(0.00, 0.01)$	$3.5, p \leq .001$
$5: FIML - WLSMV$	$0.03(0.02, 0.04)$	$6.0, p \leq .0001$	$0.00(0.00, 0.01)$	$3.8, p \leq .001$
$6: 1 \times 2$	$0.00(-0.02, 0.03)$	$0.0, NS$	$0.00(-0.01, 0.01)$	$0.1, NS$
$7: 1 \times 3$	$0.00(-0.03, 0.02)$	$-0.2, NS$	$0.00(-0.01, 0.01)$	$0.0, NS$
$8: 1 \times 4$	$0.00(-0.03, 0.02)$	$-0.1, NS$	$0.00(0.00, 0.01)$	$0.1, NS$
$9: 1 \times 5$	$0.00(-0.02, 0.02)$	$-0.2, NS$	$0.00(0.00, 0.00)$	$0.1, NS$

As can be seen in Table 3.8, the random effect covariance $\tau_{\alpha\beta}$, did not exhibit a sample-size effect. Estimates of $\tau_{\alpha\beta}$ were significantly less biased for 3 QP versus WLSMV ($\beta(95\%CI) = 0.004(0.0004, 0.01), t = 2.2, p <= .05$). likewise, estimates based on 7 QP were less biased than WLSMV ($\beta(95\%CI) = 0.01(0.002, 0.01), t = 3.1, p <= .01$). FIML estimates, in aggregate, were less biased than WLSMV ($\beta(95\%CI) = 0.01(0.002, 0.01), t = 3.1, p <= .01$). This model accounted for 1% of the variance in bias.

Table 3.8: Meta Model Results: Random Effect Covariance Estimate Raw Bias for Model 1, Set 1

Raw Bias		
	$\hat{\tau}_{\alpha\beta}$	
Contrast	$\beta(95\%CI)$	$t, pval$
$1 : N3k - N750$	$0.00(0.00, 0.01)$	$1.6, NS$
$2 : 3QP - 7QP$	$0.00(0.00, 0.01)$	$0.9, NS$
$3 : 3QP - WLSMV$	$0.00(0.00, 0.01)$	$2.2, p \leq .05$
$4 : 7QP - WLSMV$	$0.01(0.00, 0.01)$	$3.1, p \leq .01$
$5 : FIML - WLSMV$	$0.01(0.00, 0.01)$	$3.1, p \leq .01$
$6 : 1 \times 2$	$0.00(-0.01, 0.01)$	$-0.1, NS$
$7 : 1 \times 3$	$0.00(-0.01, 0.00)$	$-1.1, NS$
$8 : 1 \times 4$	$0.00(-0.01, 0.00)$	$-1.1, NS$
$9 : 1 \times 5$	$0.00(-0.01, 0.00)$	$-1.3, NS$

Aggregate bias for the residual variance estimates, presented in Table 3.9, was significantly lower when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = -0.01(-0.02, -0.00), t = -2.0, p <= .05$). In addition, aggregate bias was significantly

lower for 7 QP estimates compared to 3 QP ($\beta(95\%CI) = 0.03(0.02, 0.04), t = 4.7, p <= .0001$). Both 3 QP estimates ($\beta(95\%CI) = 0.05(0.03, 0.06), t = 7.1, p <= .0001$) and 7 QP estimates ($\beta(95\%CI) = 0.08(0.07, 0.09), t = 11.9, p <= .0001$) significantly differed from WLSMV estimates. In addition, the aggregate FIML bias was significantly less than WLSMV bias ($\beta(95\%CI) = 0.06(0.05, 0.07), t = 11.0, p <= .0001$). This model accounted for 8% of the variance in bias. Few individual parameters deviated from the aggregate trend, though two did not exhibit the sample size effect, and ψ_1 did not show any difference in bias between 3 QP and WLSMV.

Table 3.9: Meta Model Results: $\hat{\psi}$ Raw Bias for Model 1, Set 1

Raw $\hat{\psi}$ Bias						
Results when $\hat{\psi}_i$ Aggregated			Correspondence to Aggregate by $\hat{\psi}_i$			
Contrast	$\beta(95\%CI)$	$t, pval$	ψ_1	ψ_2	ψ_3	ψ_4
$1 : N3k - N750$	$-0.01(-0.02, -0.00)$	$-2.0, p \leq .05$	-	-	*	*
$2 : 3QP - 7QP$	$0.03(0.02, 0.04)$	$4.7, p \leq .0001$	*	*	*	*
$3 : 3QP - WLSMV$	$0.05(0.03, 0.06)$	$7.1, p \leq .0001$	-	*	*	*
$4 : 7QP - WLSMV$	$0.08(0.07, 0.09)$	$11.9, p \leq .0001$	*	*	*	*
$5 : FIML - WLSMV$	$0.06(0.05, 0.07)$	$11.0, p \leq .0001$	*	*	*	*
$6 : 1 \times 2$	$-0.00(-0.03, 0.02)$	$-0.1, NS$	*	*	*	*
$7 : 1 \times 3$	$-0.01(-0.04, 0.01)$	$-0.9, NS$	*	*	*	*
$8 : 1 \times 4$	$-0.01(-0.04, 0.01)$	$-1.0, NS$	*	*	*	*
$9 : 1 \times 5$	$-0.01(-0.04, 0.01)$	$-1.1, NS$	*	*	*	*

3.1.5 Second Parameter Set Bias

As can be seen in Table 3.10 with $N = 750$, the degree of bias for model 1 under parameter set 2 was higher than that observed for model 1 under parameter set 1. Though bias was higher in relative terms, bias remained acceptably low in absolute terms, with only a few parameters, notably b_7 under WLSMV estimation, exhibiting large degrees of bias. With $N = 3000$ WLSMV estimates of b_7 remained profoundly biased, though as can be seen in Table 3.11, for most other parameters all estimators performed with minimal bias. Again, though bias remained low in absolute terms, WLSMV did appear more biased than either FIML estimator.

Table 3.10: Item and Structural Parameter Bias for Model 1, Set 2, N=750

N=750										
		$QP = 3$			$QP = 7$			$WLSMV$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.48(0.05)	0.018	0.039	0.46(0.05)	0.001	0.002	0.50(0.07)	0.037	0.081
a_2	0.69	0.72(0.07)	0.033	0.048	0.70(0.07)	0.007	0.011	0.68(0.09)	-0.007	-0.010
a_4	1.92	2.01(0.18)	0.085	0.044	1.95(0.20)	0.031	0.016	1.65(0.25)	-0.268	**−0.140**
a_5	1.20	1.24(0.10)	0.042	0.035	1.20(0.10)	0.001	0.001	1.26(0.14)	0.057	0.047
a_6	1.80	1.90(0.18)	0.099	0.055	1.82(0.18)	0.023	0.013	1.90(0.26)	0.104	0.058
a_7	1.68	1.79(0.24)	0.106	0.063	1.72(0.26)	0.043	0.026	1.00(0.15)	-0.685	**−0.408**
a_8	1.76	1.83(0.16)	0.070	0.040	1.77(0.15)	0.013	0.007	1.76(0.21)	-0.004	-0.002
a_9	0.30	0.31(0.04)	0.010	0.034	0.30(0.04)	-0.001	-0.003	0.30(0.05)	0.003	0.010
b_1	3.30	3.24(0.18)	-0.064	-0.019	3.32(0.18)	0.016	0.005	3.26(0.23)	-0.041	-0.013
b_2	-1.00	-0.93(0.23)	0.071	-0.071	-1.00(0.23)	0.000	0.000	-1.06(0.31)	-0.065	0.065
b_4	4.00	3.90(0.18)	-0.098	-0.025	4.00(0.19)	-0.003	-0.001	4.03(0.26)	0.027	0.007
b_5	2.50	2.45(0.10)	-0.047	-0.019	2.50(0.11)	0.004	0.002	2.47(0.13)	-0.035	-0.014
b_6	1.20	1.19(0.06)	-0.007	-0.006	1.20(0.07)	0.000	0.000	1.20(0.07)	0.000	0.000
b_7	-1.50	-1.42(0.24)	0.076	-0.050	-1.52(0.27)	-0.015	0.010	-2.45(0.49)	-0.947	**0.631**
b_8	2.80	2.74(0.12)	-0.061	-0.022	2.80(0.12)	-0.001	0.000	2.76(0.15)	-0.039	-0.014
b_9	-2.00	-1.98(0.50)	0.024	-0.012	-2.08(0.52)	-0.084	0.042	-2.22(0.66)	-0.224	**0.112**
μ_α	1.39	1.37(0.07)	-0.020	-0.014	1.39(0.07)	-0.004	-0.003	1.38(0.07)	-0.015	-0.011
μ_β	0.50	0.49(0.04)	-0.013	-0.025	0.50(0.04)	0.003	0.006	0.49(0.05)	-0.008	-0.015
τ_α	0.67	0.64(0.13)	-0.027	-0.040	0.69(0.14)	0.018	0.027	0.68(0.17)	0.014	0.020
$\tau_{\alpha\beta}$	0.05	0.04(0.04)	-0.010	**−0.210**	0.04(0.04)	-0.008	**−0.185**	0.04(0.05)	-0.007	**−0.159**
τ_β	0.05	0.05(0.03)	0.002	0.044	0.06(0.03)	0.006	**0.124**	0.06(0.04)	0.005	**0.108**
ψ_1	0.67	0.61(0.14)	-0.056	-0.084	0.66(0.15)	-0.014	-0.020	0.60(0.19)	-0.066	-0.099
ψ_2	0.81	0.75(0.13)	-0.059	-0.073	0.81(0.14)	0.002	0.002	0.75(0.18)	-0.057	-0.070
ψ_3	1.05	0.98(0.17)	-0.071	-0.067	1.06(0.18)	0.011	0.010	1.00(0.24)	-0.054	-0.052
ψ_4	1.39	1.28(0.25)	-0.112	-0.080	1.39(0.26)	-0.009	-0.007	1.33(0.34)	-0.069	-0.050

Table 3.11: Item and Structural Parameter Bias for Model 1, Set 2, N=3000

N=3000										
		$QP = 3$			$QP = 7$			$WLSMV$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.48(0.02)	0.018	0.039	0.46(0.02)	0.000	0.000	0.50(0.03)	0.044	0.095
a_2	0.69	0.72(0.04)	0.029	0.042	0.69(0.04)	0.002	0.002	0.68(0.04)	-0.013	-0.019
a_4	1.92	1.99(0.09)	0.071	0.037	1.93(0.10)	0.005	0.003	1.64(0.11)	-0.278	**−0.145**
a_5	1.20	1.24(0.05)	0.044	0.036	1.20(0.05)	-0.001	-0.001	1.27(0.06)	0.069	0.058
a_6	1.80	1.88(0.10)	0.082	0.046	1.80(0.09)	0.001	0.000	1.82(0.10)	0.023	0.013
a_7	1.68	1.76(0.12)	0.084	0.050	1.69(0.12)	0.014	0.008	1.19(0.11)	-0.485	**−0.289**
a_8	1.76	1.83(0.08)	0.066	0.037	1.76(0.08)	0.005	0.003	1.86(0.11)	0.104	0.059
a_9	0.30	0.31(0.02)	0.011	0.037	0.30(0.02)	-0.001	-0.003	0.31(0.02)	0.013	0.044
b_1	3.30	3.22(0.10)	-0.081	-0.024	3.31(0.10)	0.006	0.002	3.19(0.10)	-0.111	-0.034
b_2	-1.00	-0.92(0.12)	0.076	-0.076	-1.00(0.12)	0.000	0.000	-1.06(0.14)	-0.057	0.057
b_4	4.00	3.89(0.10)	-0.106	-0.027	4.00(0.10)	-0.001	0.000	3.94(0.12)	-0.060	-0.015
b_5	2.50	2.45(0.05)	-0.054	-0.022	2.50(0.06)	0.001	0.001	2.42(0.05)	-0.079	-0.032
b_6	1.20	1.19(0.03)	-0.008	-0.007	1.20(0.03)	-0.001	-0.001	1.19(0.03)	-0.014	-0.012

Continued on next page

Table 3.11 – continued from previous page

		N=3000								
		$QP = 3$			$QP = 7$			$WLSMV$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
b_7	-1.50	-1.41(0.13)	0.095	-0.063	-1.50(0.14)	-0.002	0.001	-1.92(0.24)	-0.424	**0.283**
b_8	2.80	2.73(0.06)	-0.066	-0.024	2.80(0.07)	0.000	0.000	2.70(0.07)	-0.104	-0.037
b_9	-2.00	-1.92(0.26)	0.082	-0.041	-2.03(0.27)	-0.033	0.016	-2.03(0.29)	-0.029	0.015
μ_α	1.39	1.37(0.04)	-0.018	-0.013	1.39(0.04)	-0.001	-0.001	1.37(0.03)	-0.024	-0.017
μ_β	0.50	0.48(0.02)	-0.017	-0.033	0.50(0.02)	0.000	0.001	0.48(0.02)	-0.024	-0.049
τ_α	0.67	0.62(0.07)	-0.045	-0.068	0.67(0.08)	0.002	0.004	0.62(0.08)	-0.047	-0.070
$\tau_{\alpha\beta}$	0.05	0.04(0.02)	-0.004	-0.077	0.04(0.02)	-0.002	-0.042	0.04(0.03)	-0.009	**−0.200**
τ_β	0.05	0.05(0.01)	-0.003	-0.067	0.05(0.02)	0.001	0.013	0.05(0.02)	-0.001	-0.019
ψ_1	0.67	0.63(0.07)	-0.043	-0.065	0.67(0.08)	0.004	0.006	0.56(0.09)	-0.107	**−0.159**
ψ_2	0.81	0.75(0.07)	-0.063	-0.077	0.81(0.07)	0.003	0.004	0.69(0.08)	-0.122	**−0.150**
ψ_3	1.05	0.97(0.09)	-0.080	-0.076	1.06(0.10)	0.008	0.008	0.93(0.10)	-0.125	**−0.119**
ψ_4	1.39	1.28(0.12)	-0.116	-0.083	1.39(0.13)	-0.004	-0.003	1.23(0.13)	-0.166	**−0.119**

3.1.6 Second Parameter Set Meta Model Results

Meta model results for aggregate $\hat{a}_i$ bias under item parameter set 2 differ substantially from those observed for set 1. In particular, many sample size by estimator interactions were observed, and the magnitude of the point estimates were larger, reflecting greater degrees and wider variation of bias in set 2. Bias for estimates obtained under 7 QP was significantly smaller than that associated with 3 QP estimates ($\beta(95\%CI) = -0.05(-0.05, -0.04), t = -10.7, p \leq .0001$). Both 3 QP ($\beta(95\%CI) = 0.14(0.13, 0.14), t = 31.8, p \leq .0001$) and 7 QP ($\beta(95\%CI) = 0.09(0.08, 0.10), t = 21.1, p \leq .0001$) were significantly less biased compared to WLSMV. In addition, the average FIML bias was significantly lower than the bias observed for WLSMV estimates ($\beta(95\%CI) = 0.11(0.10, 0.12), t = 30.5, p \leq .0001$). However, main effects involving the contrast with WLSMV were not possible to interpret given the interaction effects observed in this model. The difference in bias between estimates based on 3 QP and WLSMV was significantly greater when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = 0.04(0.02, 0.05), t = 4.4, p \leq .0001$). This was also the case for the difference in bias between 7 QP and WLSMV ($\beta(95\%CI) = 0.04(0.03, 0.06), t = 4.9, p \leq .0001$), and the aggregate FIML bias versus WLSMV ($\beta(95\%CI) = 0.04(0.03, 0.05), t = 5.4, p \leq .0001$).

Thus, as sample size increased, the discrepancy in bias between estimators diminished. Reflecting the greater variation in bias observed under parameter set 2, this model accounted for a larger proportion of the observed variation in bias.

As with Set 1, there were some individual $\hat{a}_i$ for which bias trends did not correspond to the aggregate trend. Notably, 50% of the eight estimated $\hat{a}_i$ did not have any of the sample size by estimator interactions, and nearly all of the $\hat{a}_i$ exhibited a main effect for sample size not observed in aggregate.

Table 3.12: Meta Model Results: $\hat{a}$ Raw Bias for Model 1, Set 2

Raw $\hat{a}$ Bias										
Results when $\hat{a}_i$ Aggregated			Correspondence to Aggregate by $\hat{a}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	a_1	a_2	a_4	a_5	a_6	a_7	a_8	a_9
$1: N3k - N750$	$0.00(0.00, 0.01)$	$1.1, NS$	*	*	-	*	-	-	-	-
$2: 3QP - 7QP$	$-0.05(-0.05, -0.04)$	$-10.7, p \leq .0001$	*	*	*	*	*	*	*	*
$3: 3QP - WLSMV$	$0.14(0.13, 0.14)$	$31.8, p \leq .0001$	*	*	*	*	*	*	*	-
$4: 7QP - WLSMV$	$0.09(0.08, 0.10)$	$21.1, p \leq .0001$	*	*	*	*	*	*	*	*
$5: FIML - WLSMV$	$0.11(0.10, 0.12)$	$30.5, p \leq .0001$	*	*	*	*	-	*	-	-
$6: 1 \times 2$	$0.01(-0.01, 0.02)$	$0.6, NS$	*	*	*	*	*	*	*	*
$7: 1 \times 3$	$0.04(0.02, 0.05)$	$4.4, p \leq .0001$	-	-	-	-	*	*	*	*
$8: 1 \times 4$	$0.04(0.03, 0.06)$	$4.9, p \leq .0001$	-	-	-	-	*	*	*	*
$9: 1 \times 5$	$0.04(0.03, 0.05)$	$5.4, p \leq .0001$	-	-	-	-	*	*	*	*

The magnitude of bias for estimates of $\hat{b}_i$ differed significantly as a function of sample size and estimator main effects, however, these main effects were subsumed by the existence of multiple significant interactions. The difference in bias between estimates based on 3 QP and WLSMV was significantly greater when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = 0.05(0.03, 0.07), t = 5.6, p \leq .0001$). This was also the case for the difference in bias between 7 QP and WLSMV ($\beta(95\%CI) = 0.05(0.03, 0.07), t = 5.5, p \leq .0001$), and the aggregate FIML bias versus WLSMV ($\beta(95\%CI) = 0.05(0.03, 0.06), t = 6.4, p \leq .0001$). Consistent with the results observed for $\hat{a}_i$ as sample size increased, the discrepancy in bias for $\hat{b}_i$ between estimators diminished. This model accounted for approximately 40% of the variation in aggregate $\hat{b}_i$

bias.

Some individual parameters did not follow the trends observed in aggregate. A notable discrepancy was the existence of the main effect for 3 versus 7 QP for all parameters individually, which was not observed in aggregate. In addition, a majority of parameters did not have a significant main effect for the difference between 3 QP and WLSMV.

Table 3.13: Meta Model Results: $\hat{b}$ Raw Bias for Model 1, Set 2

Raw $\hat{b}$ Bias										
Results when $\hat{b}_i$ Aggregated			Correspondence to Aggregate by $\hat{b}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	b_1	b_2	b_4	b_5	b_6	b_7	b_8	b_9
$1: N3k - N750$	$0.02(0.02, 0.03)$	$6.2, p \leq .0001$	*	-	*	*	*	*	*	*
$2: 3QP - 7QP$	$0.00(-0.01, 0.01)$	$0.8, NS$	-	-	-	-	-	-	-	-
$3: 3QP - WLSMV$	$0.13(0.12, 0.14)$	$28.4, p \leq .0001$	-	*	*	-	-	*	-	*
$4: 7QP - WLSMV$	$0.13(0.12, 0.14)$	$29.2, p \leq .0001$	*	*	-	*	*	*	*	*
$5: FIML - WLSMV$	$0.13(0.12, 0.14)$	$33.3, p \leq .0001$	*	*	*	*	-	*	*	*
$6: 1 \times 2$	$-0.00(-0.02, 0.02)$	$-0.1, NS$	*	*	*	*	*	*	*	*
$7: 1 \times 3$	$0.05(0.03, 0.07)$	$5.6, p \leq .0001$	*	-	*	*	*	*	*	*
$8: 1 \times 4$	$0.05(0.03, 0.07)$	$5.5, p \leq .0001$	*	-	*	*	*	*	*	*
$9: 1 \times 5$	$0.05(0.03, 0.06)$	$6.4, p \leq .0001$	*	-	*	*	*	*	*	*

When considering bias in estimates of the growth curve fixed effects, substantial discrepancies were observed. In the case of μ_α, only estimator main effects were observed, with 7 QP being less biased than 3 QP ($\beta(95\%CI) = 0.02(0.01, 0.02), t = 5.2, p \leq .0001$), and 7 QP being less biased than WLSMV ($\beta(95\%CI) = 0.02(0.01, 0.02), t = 5.3, p \leq .0001$), as well as the aggregate FIML bias being smaller than WLSMV ($\beta(95\%CI) = 0.01(0.00, 0.01), t = 3.2, p \leq .01$). This model accounted for 2.2% of the variation in bias.

However, for μ_β, main effects *and* interactions were observed. Estimator differences varied as a function of sample size. For example, the difference in bias between 3 QP and WLSMV was significantly smaller when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = -0.01(-0.02, -0.01), t = -3.3, p \leq .001$). The same trend was observed for 7 QP versus WLSMV when $N = 750$ versus $N = 3000$ ($\beta(95\%CI) = -0.01(-0.02, -0.01), t = -3.7, p \leq .001$). Consequently, in contrast

to the bias observed for $\hat{a}$ and $\hat{b}$, the estimator difference for 3 and 7 QP versus WLSMV *increased* as a function of sample size. This was also true for the aggregate FIML bias compared to WLSMV, which diverged in bias as sample size increased ($\beta(95\%CI) = -0.01(-0.02, -0.01), t = -4.1, p \leq .0001$). This model accounted for 7.5% of the variation in bias.

Table 3.14: Meta Model Results: Fixed Effect Raw Bias for Model 1, Set 2

Raw Bias				
	$\hat{\mu}_\alpha$		$\hat{\mu}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.00(-0.01, 0.00)$	$-0.4, NS$	$-0.01(-0.01, -0.00)$	$-5.0, p \leq .0001$
$2: 3QP - 7QP$	$0.02(0.01, 0.02)$	$5.2, p \leq .0001$	$0.02(0.01, 0.02)$	$8.5, p \leq .0001$
$3: 3QP - WLSMV$	$0.00(-0.01, 0.01)$	$0.2, NS$	$0.00(0.00, 0.01)$	$0.7, NS$
$4: 7QP - WLSMV$	$0.02(0.01, 0.02)$	$5.3, p \leq .0001$	$0.02(0.01, 0.02)$	$9.2, p \leq .0001$
$5: FIML - WLSMV$	$0.01(0.00, 0.01)$	$3.2, p \leq .01$	$0.01(0.01, 0.01)$	$5.7, p \leq .0001$
$6: 1 \times 2$	$-0.00(-0.01, 0.01)$	$-0.3, NS$	$-0.00(-0.01, 0.01)$	$-0.4, NS$
$7: 1 \times 3$	$-0.01(-0.02, 0.00)$	$-1.6, NS$	$-0.01(-0.02, -0.01)$	$-3.3, p \leq .001$
$8: 1 \times 4$	$-0.01(-0.02, 0.00)$	$-1.8, NS$	$-0.01(-0.02, -0.01)$	$-3.7, p \leq .001$
$9: 1 \times 5$	$-0.01(-0.02, 0.00)$	$-2.0, NS$	$-0.01(-0.02, -0.01)$	$-4.1, p \leq .0001$

In the case of the random intercept variance, τ_α, bias differed as a function of sample size and estimator main effects, though these effects were subsumed by interactions. Though in the case of the FIML estimator contrast, main effects and not interactions were observed, with FIML based on 7 QP being less biased than FIML based on 3 QP ($\beta(95\%CI) = 0.05(0.03, 0.06), t = 6.9, p \leq .0001$). The difference in bias between FIML based on 3 QP and WLSMV was significantly greater when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = -0.04(-0.07, -0.02), t = -3.2, p \leq .01$), indicating a convergence in estimator behavior as sample size increased. In contrast, the opposite was true for FIML based on 7 QP versus WLSMV, as sample size increased the difference in bias between these estimators increased ($\beta(95\%CI) = -0.05(-0.07, -0.02), t = -3.4, p \leq .001$), and the same was true for the difference in bias between the aggregate FIML and WLSMV ($\beta(95\%CI) = -0.04(-0.07, -0.02), t = -3.8, p \leq .001$). This model accounted

for 5% of the variance in bias.

The variance for the slope random effect, τ_β, only exhibited sample size and estimator differences. These estimates were much less biased when $N = 3000$ compared to $N = 750$ ($\beta(95\%CI) = -0.01(-0.01, -0.00), t = -5.2, p \leq .0001$). FIML estimates based on 3 QP tended to be less biased than those based on 7 QP ($\beta(95\%CI) = 0.004(0.001, 0.01), t = 2.9, p \leq .01$), and WLSMV estimation as well ($\beta(95\%CI) = -0.003(-0.01, -0.0001), t = -2.0, p \leq .05$). Approximately 2% of the variance in bias was accounted for by this model.

Table 3.15: Meta Model Results: Random Effect Variance Estimate Raw Bias for Model 1, Set 2

Raw Bias				
	$\hat{\tau}_\alpha$		$\hat{\tau}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.03(-0.04, -0.02)$	$-5.8, p \leq .0001$	$-0.01(-0.01, -0.00)$	$-5.2, p \leq .0001$
$2: 3QP - 7QP$	$0.05(0.03, 0.06)$	$6.9, p \leq .0001$	$0.00(0.00, 0.01)$	$2.9, p \leq .01$
$3: 3QP - WLSMV$	$-0.02(-0.03, -0.01)$	$-2.9, p \leq .01$	$-0.00(-0.01, -0.00)$	$-2.0, p \leq .05$
$4: 7QP - WLSMV$	$0.03(0.01, 0.04)$	$4.0, p \leq .0001$	$0.00(0.00, 0.00)$	$0.9, NS$
$5: FIML - WLSMV$	$0.00(-0.01, 0.01)$	$0.6, NS$	$-0.00(0.00, 0.00)$	$-0.7, NS$
$6: 1 \times 2$	$-0.00(-0.03, 0.02)$	$-0.2, NS$	$0.00(-0.01, 0.01)$	$0.0, NS$
$7: 1 \times 3$	$-0.04(-0.07, -0.02)$	$-3.2, p \leq .01$	$-0.00(-0.01, 0.00)$	$-0.3, NS$
$8: 1 \times 4$	$-0.05(-0.07, -0.02)$	$-3.4, p \leq .001$	$-0.00(-0.01, 0.00)$	$-0.3, NS$
$9: 1 \times 5$	$-0.04(-0.07, -0.02)$	$-3.8, p \leq .001$	$-0.00(-0.01, 0.00)$	$-0.3, NS$

In the case of the covariance, $\tau_{\alpha\beta}$, estimates were less biased when $N = 3000$ ($\beta(95\%CI) = 0.004(0.0001, 0.01), t = 2.0, p \leq .05$), though this effect differed as a function of sample size, with the difference in bias of FIML and WLSMV being smaller when $N = 750$ versus when $N = 3000$ ($\beta(95\%CI) = -0.01(-0.02, -0.001), t = -2.2, p \leq .05$). This model accounted for approximately 1% of the variance in bias.

Table 3.16: Meta Model Results: Random Effect Covariance Estimate Raw Bias for Model 1, Set 2

Raw Bias		
	$\hat{\tau}_{\alpha\beta}$	
Contrast	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$0.00(0.00, 0.01)$	$2.0, p \leq .05$
$2: 3QP - 7QP$	$0.00(0.00, 0.01)$	$0.6, NS$
$3: 3QP - WLSMV$	$0.00(0.00, 0.01)$	$0.8, NS$
$4: 7QP - WLSMV$	$0.00(0.00, 0.01)$	$1.4, NS$
$5: FIML - WLSMV$	$0.00(0.00, 0.01)$	$1.2, NS$
$6: 1 \times 2$	$0.00(-0.01, 0.01)$	$-0.1, NS$
$7: 1 \times 3$	$-0.01(-0.02, 0.00)$	$-1.8, NS$
$8: 1 \times 4$	$-0.01(-0.02, 0.00)$	$-1.9, NS$
$9: 1 \times 5$	$-0.01(-0.02, -0.00)$	$-2.2, p \leq .05$

Bias for the residual variance exhibited main effects for sample size and estimator, though only the FIML contrast was not subsumed by interactions, with FIML estimation based on 3 QP significantly more biased than FIML estimation based on 7 QP ($\beta(95\%CI) = 0.08(0.06, 0.09), t = 10.8, p \leq .0001$). The difference in bias between 3 QP and WLSMV was significantly greater when $N = 750$ versus $N = 3000$ ($\beta(95\%CI) = -0.07(-0.09, -0.04), t = -4.8, p \leq .0001$), as was the case for FIML based on 7 QP versus WLSMV ($\beta(95\%CI) = -0.07(-0.10, -0.05), t = -5.3, p \leq .0001$), and the aggregate FIML bias versus WLSMV ($\beta(95\%CI) = -0.07(-0.09, -0.05), t = -5.8, p \leq .0001$). This model accounted for 12.5% of the variation in bias for the residual variance. As was the case in the first item parameter set, there was virtually no divergence from the aggre-

gate trend among the individual parameters, though a notable exception is the WLSMV versus 3 QP effect, which was not observed for 50% of the parameters.

Table 3.17: Meta Model Results: $\hat{\psi}$ Raw Bias for Model 1, Set 2

Raw $\hat{\psi}$ Bias						
Results when $\hat{\psi}_i$ Aggregated			Correspondence to Aggregate by $\hat{\psi}_i$			
Contrast	$\beta(95\%CI)$	$t, pval$	ψ_1	ψ_2	ψ_3	ψ_4
$1: N3k - N750$	$-0.02(-0.03, -0.01)$	$-3.7, p \leq .001$	-	*	*	*
$2: 3QP - 7QP$	$0.08(0.06, 0.09)$	$10.8, p \leq .0001$	*	*	*	*
$3: 3QP - WLSMV$	$0.02(0.01, 0.03)$	$3.0, p \leq .01$	*	*	-	-
$4: 7QP - WLSMV$	$0.10(0.08, 0.11)$	$13.8, p \leq .0001$	*	*	*	*
$5: FIML - WLSMV$	$0.06(0.05, 0.07)$	$9.7, p \leq .0001$	*	*	*	*
$6: 1 \times 2$	$-0.01(-0.03, 0.02)$	$-0.5, NS$	*	*	*	*
$7: 1 \times 3$	$-0.07(-0.09, -0.04)$	$-4.8, p \leq .0001$	*	*	*	*
$8: 1 \times 4$	$-0.07(-0.10, -0.05)$	$-5.3, p \leq .0001$	*	*	*	*
$9: 1 \times 5$	$-0.07(-0.09, -0.05)$	$-5.8, p \leq .0001$	*	*	*	*

3.1.7 RMSE

RMSE results for both item parameter set 1 and 2 in model 1 and model 2 did not differ in any substantial manner from the results observed for bias. The magnitude of RMSE was relatively low, estimators maintain the same relative ranking in terms of RMSE as bias, and the sample size effect is trivial. Consequently, in the interest of devoting space to only the most relevant information, RMSE results are relegated to a series of appendices.

3.1.8 Summary of Findings

Under Model 1, non-converged (NCV) solutions were only observed for the FIML estimator based on 3 QP. NCV solutions were more common when $N = 750$ compared to $N = 3000$. Though I hypothesized a higher rate of NCV solutions under parameter set 2, when $N = 750$ more NCV solutions were observed under parameter set 1 than were observed for parameter set 2, while the opposite was true when $N = 3000$. Non-positive definite (NPD) solutions were restricted to the random effect covariance matrix,

$\hat{\boldsymbol{\tau}}_{\boldsymbol{\eta}}$, particularly, $\hat{\tau}_\beta$. NPD solutions under the first parameter set varied in frequency as a function of estimator, with the majority being observed for WLSMV, followed by FIML based on 3 QP, and the least observed for FIML based on 7 QP. Under the second parameter set the same pattern was observed, though FIML based on 7 QP produced one more NPD solution than 3 QP. For both parameter sets, virtually no NPD solutions were observed when $N = 3000$.

Under model 1 bias for all estimators was acceptably low, with only estimates of $\hat{\tau}_{\alpha\beta}$ exceeding 10% relative bias for all estimators when $N = 750$, and only WLSMV estimates of $\hat{\tau}_{\alpha\beta}$ when $N = 3000$. Though estimates of $\hat{\tau}_{\alpha\beta}$ had the highest relative bias value, higher raw bias values were observed for other parameters for all estimators. Bias values for all estimators were higher under the second parameter set, though WLSMV estimates were the most biased. The highest raw and relative bias values were observed for WLSMV estimates of the slope and threshold parameters for item 7. Meta model results for raw bias within the first parameter set produced mostly subtle estimator differences with FIML estimators less biased than WLSMV. For some parameters, specifically aggregate $\hat{b}$, $\hat{\mu}_\beta$, $\hat{\tau}_\alpha$, $\hat{\tau}_\beta$, and aggregate $\hat{\psi}$, sample size main effects were observed. In addition, for some parameters differences in bias were observed between FIML estimators, with 7 QP less biased than 3 QP for aggregate $\hat{a}$ and $\hat{\psi}$, but the opposite true for $\hat{b}$. In contrast, under the second parameter set, though main effects were observed, the majority of parameters exhibited significant sample size by estimator interactions. For item parameters and $\hat{\psi}$, differences between FIML and WLSMV bias diminished as sample size increased, while the opposite was true for all estimators of $\hat{\mu}_\beta$, and in the case of $\hat{\tau}_\beta$ the difference between 3 QP and WLSMV decreased with sample size, though it increased with sample size for the difference between 7 QP and WLSMV as well as the aggregate FIML bias.

3.2 Model 2

Recall that within the cells of the experimental design, full information estimation of Model 2 was infeasible with adaptive Gauss-Hermite based quadrature approximations

to the posterior distribution due to the increased dimensions of integration associated with the testlet factors for error correlations. As such, the full information estimator comparator to WLSMV was Monte-Carlo integration, implemented in the Monte-Carlo EM algorithm implemented in Mplus. Throughout, MCEM is the FIML comparator, and should be viewed as an analog of the adaptive Gauss-Hermite integration employed in Model 1, as such, expressed hypotheses predict better performance for MCEM than WLSMV.

3.2.1 Convergence

As with Model 1, WLSMV never failed to converge. Under parameter set 1, when $N = 750$ a total of 32 replications failed to converge under MCEM, while only one failed to converge when $N = 3000$. When replications were generated under the second parameter set many more non-converged (NCV) replications were observed. With $N = 750$ a total of 75 replications failed to converge, more than double the rate observed for parameter set 1, and with $N = 3000$ 25 replications failed to converge.

An additional 100 replications were generated to replace these failed simulates. Replications which failed to converge under MCEM were deleted from the WLSMV solutions and replaced with estimates from the converged replications obtained from the additional 100 replications. Consequently, analysis was based on 300 identical replications across estimators.

3.2.2 NPD Solutions

Under MCEM estimation, with $N = 750$, a total of 26 replications had non-positive definite (NPD) random effect covariance matrices, and each NPD solution was associated with covariance matrices whose point estimates were within the parametric support having very slightly negative eigenvalues, the largest being $1.29834E - 9$. With $N = 3000$, only 3 replications had NPD estimates of $\boldsymbol{\tau_\eta}$, and again, each estimate was characterized by apparently non-degenerate point estimates but slightly negative eigenvalues, the

largest being $-1.5637E-10$.

When estimation was based on WLSMV degeneracy was observed with greater frequency. When $N = 750$ a total of 43 replications had NPD estimates of $\boldsymbol{\tau_\eta}$, though most NPD cases were associated with seemingly proper point estimates, multiple negative variance estimates were observed for τ_β. In addition, the negative eigenvalues were much larger, for example, the smallest eigenvalue was -0.000302, while the largest was -0.079986. Consequently, the NPD solutions under WLSMV appear to be more substantial departures from the parametric support. With $N = 3000$, 5 replications had NPD estimates of the random effect covariance matrix. Point estimates appeared proper, but negative eigenvalues were again relatively large, with a range of $-.000563818$ to $-.001532073$.

The second parameter set was associated with a much higher rate of degenerate solutions for both estimators. With $N = 750$, 27 replications had degenerate solutions, all estimates were within the parametric support, and as with the first parameter set, negative eigenvalues were very small, having a range of $-8.2765E-12$ to $-7.266E-10$. Given the same number sample size, WLSMV had nearly triple the number of degenerate solutions, with a total of 61 NPD estimates of $\boldsymbol{\tau_\eta}$. Roughly half (31) of these cases were associated with negative point estimates for τ_β. As with model 1, the NPD solutions were more profoundly degenerate, with the range of observed eigenvalues being -0.001730 to -0.098104. With $N = 3000$, only 6 replications had NPD solutions for $\boldsymbol{\tau_\eta}$ under MCEM, each of which had positive point estimates for τ_β. The observed range of the improper eigenvalues was $-2.6545E-11$ to $-5.0596E-11$, indicating trivial departure from the parametric support. WLSMV estimation produced only 2 NPD solutions, each of which had seemingly proper point estimates of the elements of $\boldsymbol{\tau_\eta}$, though again, eigenvalues associated with the degenerate solutions were much larger than was observed for MCEM, having a range from $-.001072593$ to $-.007005037$.

3.2.3 First Parameter Set Bias

As can be seen in Table 3.18, at $N = 750$, the degree of bias for the correlated error model was higher than that observed for Model 1. Even given the uniformly high degree of bias, estimators could be differentiated in degree of bias. Specifically, MCEM was substantially more biased than WLSMV for nearly every parameter. As with Model 1, WLSMV demonstrated a pronounced negative bias for threshold estimates. Though MCEM exhibited particularly high rates of bias, with virtually every parameter exceeding 10% relative bias, MCEM estimates of the fixed effects and random effect covariance components were relatively unbiased in the raw metric, though relative bias indicated greater problems. Note that while WLSMV estimates the error correlations directly, MCEM estimates the root of the correlation as λ, and under the testlet parameterization in MCEM, $\rho = \lambda^2$, consequently, the bias values for $\hat{\rho}$ under MCEM are in relation to the generating parameter for λ, which was $\sqrt{\rho} = .5477$.

Table 3.18: Item and Structural Parameter Bias for Model 2, Set 1, N=750

N=750							
		WLSMV			*MCEM*		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.49(0.06)	0.03	0.066	**0.593(0.12)**	**0.133**	**0.289**
a_2	0.69	0.693(0.07)	0.003	0.005	**0.864(0.17)**	**0.174**	**0.252**
a_4	0.92	0.951(0.1)	0.031	0.034	**1.212(0.24)**	**0.292**	**0.317**
a_5	1.15	1.252(0.11)	0.102	0.089	**1.46(0.27)**	**0.31**	**0.269**
a_6	1.37	1.472(0.13)	0.102	0.074	**1.725(0.33)**	**0.355**	**0.259**
a_7	1.68	1.715(0.2)	0.035	0.021	**2.121(0.41)**	**0.441**	**0.263**
a_8	1.76	1.867(0.19)	0.107	0.061	**2.204(0.42)**	**0.444**	**0.252**

Continued on next page

Table 3.18 – continued from previous page

N=750							
		WLSMV			*MCEM*		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_9	0.3	0.327(0.05)	0.027	0.089	**0.378(0.08)**	**0.078**	**0.259**
b_1	2.3	2.164(0.12)	-0.136	-0.059	2.187(0.18)	-0.113	-0.049
b_2	-0.5	-0.512(0.21)	-0.012	0.023	**−0.134(0.32)**	**0.366**	**−0.732**
b_4	3	2.815(0.14)	-0.185	-0.062	2.731(0.25)	-0.269	-0.09
b_5	1.5	1.419(0.07)	-0.081	-0.054	1.528(0.11)	0.028	0.019
b_6	1	0.953(0.07)	-0.047	-0.047	**1.123(0.12)**	**0.123**	**0.123**
b_7	-0.3	**−0.155(0.13)**	**0.145**	**−0.482**	**0.18(0.24)**	**0.48**	**−1.6**
b_8	2	1.883(0.07)	-0.117	-0.059	1.931(0.13)	-0.069	-0.034
b_9	-1	-1.059(0.45)	-0.059	0.059	**−0.561(0.46)**	**0.439**	**−0.439**
μ_α	1.39	1.302(0.07)	-0.09	-0.065	1.423(0.11)	0.031	0.023
μ_β	0.5	0.475(0.04)	-0.025	-0.05	**0.413(0.07)**	**−0.087**	**−0.175**
τ_α	0.67	**0.594(0.13)**	**−0.076**	**−0.113**	**0.473(0.18)**	**−0.197**	**−0.295**
$\tau_{\alpha\beta}$	0.05	**0.037(0.05)**	**−0.009**	**−0.186**	**0.027(0.03)**	**−0.019**	**−0.418**
τ_β	0.05	**0.042(0.03)**	**−0.008**	**−0.153**	**0.028(0.02)**	**−0.022**	**−0.431**
ψ_1	0.67	0.618(0.15)	-0.052	-0.078	**0.485(0.19)**	**−0.185**	**−0.276**
ψ_2	0.81	**0.675(0.13)**	**−0.136**	**−0.168**	**0.452(0.17)**	**−0.36**	**−0.443**
ψ_3	1.05	**0.9(0.18)**	**−0.153**	**−0.145**	**0.634(0.24)**	**−0.419**	**−0.398**
ψ_4	1.39	**1.217(0.27)**	**−0.177**	**−0.127**	**0.872(0.35)**	**−0.523**	**−0.375**
ρ_1	0.3	0.28(0.08)	-0.02	-0.066	**0.347(0.14)**	**−0.201**	**−0.366**
ρ_2	0.3	**0.246(0.11)**	**−0.054**	**−0.18**	**0.294(0.22)**	**−0.253**	**−0.462**
ρ_3	0.3	**0.268(0.13)**	**−0.032**	**−0.108**	**0.336(0.22)**	**−0.212**	**−0.387**
ρ_4	0.3	0.272(0.11)	-0.028	-0.093	**0.346(0.19)**	**−0.202**	**−0.368**

As can be seen in Table 3.19 quadrupling the sample size to $N = 3000$ had little impact on the degree of observed bias of the estimators. Again, the degree of bias for the correlated error model was higher than that observed for the model without correlated errors. As was the case with $N = 750$, WLSMV was substantially less biased than MCEM. With $N = 3000$, WLSMV threshold estimates remained negatively biased, though not uniformly. As with $N = 750$, MCEM estimates only demonstrated acceptable levels of bias for fixed effect and random effect covariance component estimates.

Table 3.19: Item and Structural Parameter Bias for Model 2, Set 1, N=3000

N=3000							
		WLSMV			*MCEM*		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.491(0.03)	0.031	0.067	**0.596(0.07)**	**0.136**	**0.295**
a_2	0.69	0.691(0.04)	0.001	0.001	**0.863(0.1)**	**0.173**	**0.251**
a_4	0.92	0.945(0.05)	0.025	0.027	**1.208(0.13)**	**0.288**	**0.313**
a_5	1.15	1.252(0.06)	0.102	0.088	**1.465(0.16)**	**0.315**	**0.274**
a_6	1.37	1.467(0.07)	0.097	0.071	**1.73(0.18)**	**0.36**	**0.263**
a_7	1.68	1.671(0.1)	-0.009	-0.005	**2.123(0.24)**	**0.443**	**0.264**
a_8	1.76	1.858(0.09)	0.098	0.056	**2.203(0.24)**	**0.443**	**0.251**
a_9	0.3	0.33(0.02)	0.03	0.099	**0.383(0.04)**	**0.083**	**0.277**
b_1	2.3	2.151(0.06)	-0.149	-0.065	2.159(0.09)	-0.141	-0.061
b_2	-0.5	-0.499(0.11)	0.001	-0.003	**−0.093(0.19)**	**0.407**	**−0.814**
b_4	3	2.803(0.07)	-0.197	-0.066	**2.691(0.13)**	**−0.309**	**−0.103**
b_5	1.5	1.411(0.03)	-0.089	-0.059	1.514(0.06)	0.014	0.009
b_6	1	0.95(0.03)	-0.05	-0.05	**1.122(0.07)**	**0.122**	**0.122**
b_7	-0.3	**−0.163(0.07)**	**0.137**	**−0.457**	**0.205(0.15)**	**0.505**	**−1.684**

Continued on next page

Table 3.19 – continued from previous page

N=3000							
		$WLSMV$			$MCEM$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
b_8	2	1.875(0.04)	-0.125	-0.063	1.911(0.07)	-0.089	-0.044
b_9	-1	-0.987(0.21)	0.013	-0.013	**−0.472(0.24)**	**0.528**	**−0.528**
μ_α	1.39	1.299(0.03)	-0.093	-0.067	1.416(0.06)	0.024	0.017
μ_β	0.5	0.469(0.02)	-0.031	-0.061	**0.401(0.04)**	**−0.099**	**−0.198**
τ_α	0.67	**0.585(0.07)**	**−0.085**	**−0.127**	**0.438(0.1)**	**−0.232**	**−0.346**
$\tau_{\alpha\beta}$	0.05	**0.034(0.02)**	**−0.012**	**−0.266**	**0.026(0.02)**	**−0.02**	**−0.433**
τ_β	0.05	**0.042(0.01)**	**−0.008**	**−0.164**	**0.026(0.01)**	**−0.024**	**−0.481**
ψ_1	0.67	0.607(0.07)	-0.063	-0.095	**0.457(0.1)**	**−0.213**	**−0.318**
ψ_2	0.81	**0.675(0.07)**	**−0.136**	**−0.168**	**0.429(0.09)**	**−0.383**	**−0.472**
ψ_3	1.05	**0.88(0.09)**	**−0.173**	**−0.164**	**0.583(0.13)**	**−0.47**	**−0.446**
ψ_4	1.39	**1.169(0.13)**	**−0.226**	**−0.162**	**0.799(0.18)**	**−0.595**	**−0.427**
ρ_1	0.3	0.287(0.04)	-0.013	-0.044	**0.358(0.07)**	**−0.19**	**−0.346**
ρ_2	0.3	**0.259(0.06)**	**−0.041**	**−0.136**	**0.344(0.1)**	**−0.204**	**−0.372**
ρ_3	0.3	**0.262(0.07)**	**−0.038**	**−0.126**	**0.325(0.12)**	**−0.222**	**−0.406**
ρ_4	0.3	0.274(0.05)	-0.026	-0.085	**0.354(0.09)**	**−0.194**	**−0.353**

3.2.4 First Parameter Set Meta Model Results

As with model 1, parameter sets were aggregated. Thus models were fit to the average bias for $\hat{a}, \hat{b}, \hat{\psi}$,and$\hat{\rho}$, though in order to characterize any discrepancy between the aggregate and the individual parameters idiosyncracies are noted for each individual parameter. Results which diverge from the aggregate are denoted with a dash, while results which conform to that observed for the aggregate are denoted with an asterisk. For the vast majority of parameters modeled, only estimator main effects were observed.

Table 3.20: Meta Model Results: $\hat{a}$ Raw Bias for Model 2, Set 1

Raw $\hat{a}$ Bias										
Results when $\hat{a}_i$ Aggregated			Correspondence to Aggregate by $\hat{a}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	a_1	a_2	a_4	a_5	a_6	a_7	a_8	a_9
$1: N3k - N750$	$0(-0.02, 0.01)$	$-0.36, NS$	*	*	*	*	*	*	*	*
$2: MCEM - WLSMV$	$0.23(0.21, 0.25)$	$26.58, p \leq .0001$	*	*	*	*	*	*	*	*
$3: 1 \times 2$	$-0.01(-0.04, 0.02)$	$-0.56, NS$	*	*	*	*	*	*	*	*

As can be seen from the cell means, and the model results presented in Table 3.20, bias did not vary as a function of sample size, though MCEM was significantly more biased than WLSMV ($\beta(95\%CI) = 0.23(0.21, 0.25),\ t = 26.58, p \leq .0001$). In addition, the distance between the difference of MCEM and WLSMV did not vary as a function of sample size. All of the individual $\hat{a}$ conformed to the findings observed for the aggregate. This model accounted for 39% of the variance in bias.

Table 3.21: Meta Model Results: $\hat{b}$ Raw Bias for Model 2, Set 1

Raw $\hat{b}$ Bias										
Results when $\hat{b}_i$ Aggregated			Correspondence to Aggregate by $\hat{b}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	b_1	b_2	b_4	b_5	b_6	b_7	b_8	b_9
$1: N3k - N750$	$0.01(0, 0.02)$	$1.09, NS$	-	-	-	-	*	*	-	-
$2: MCEM - WLSMV$	$0.19(0.18, 0.2)$	$37.81, p \leq .0001$	*	*	*	*	*	*	*	*
$3: 1 \times 2$	$0(-0.02, 0.02)$	$-0.22, NS$	*	*	*	*	*	*	*	*

Results for $\hat{b}$, presented in Table 3.21, match those of $\hat{a}$. There was no difference as a function of sample size, though MCEM was significantly more biased than WLSMV ($\beta(95\%CI) = 0.19(0.18, 0.2),\ t = 37.81, p \leq .0001$). As can be seen from the cell means, the aggregate bias for $\hat{b}$ exhibited the same negative bias observed for model 1. In addition, the distance between the difference of MCEM and WLSMV did not vary as a function of sample size. All of the individual $\hat{b}$ conformed to the findings observed for the aggregate estimator and interaction effects, though the majority of the individual $\hat{b}$ exhibited a sample size effect not observed in aggregate. This sample size effect was likely due to the stronger negative bias for $\hat{b}_{WLSMV}$ when sample size was $N = 750$. This model accounted for 56% of the variance in bias.

Table 3.22: Meta Model Results: Fixed Effect Raw Bias for Model 2, Set 1

Raw Bias				
	$\hat{\mu}_\alpha$		$\hat{\mu}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.01(-0.01, 0)$	$-1.24, NS$	$-0.01(-0.01, -0.00)$	$-3.17, p \leq .01$
$2: MCEM - WLSMV$	$0.12(0.11, 0.13)$	$27.55, p \leq .0001$	$-0.07(-0.07, -0.06)$	$-24.12, p \leq .0001$
$3: 1 \times 2$	$0(-0.01, 0.02)$	$0.52, NS$	$0.01(0, 0.02)$	$1.12, NS$

Bias for fixed effect estimates, $\hat{\mu}_\alpha$ and $\hat{\mu}_\beta$, presented in Table 3.22, match the pattern observed for item parameters though direction of effect varied as a function of parameter. There was no difference as a function of sample size for $\hat{\mu}_\alpha$, though WLSMV was significantly more biased than MCEM ($\beta(95\%CI) = 0.12(0.11, 0.13),\ t = 27.55, p \leq .0001$). In the case of $\hat{\mu}_\beta$, estimates were significantly (though trivially) more negatively biased with $N = 3000$ compared to $N = 750$ ($\beta(95\%CI) = -0.01(-0.01, -0.003),\ t = -3.17, p \leq .01$), and estimates were significantly more negatively biased under MCEM than WLSMV ($\beta(95\%CI) = -0.07(-0.07, -0.06),\ t = -24.12, p \leq .0001$). This model accounted for 40% of the variance in bias for $\hat{\mu}_\alpha$ $R^2 = 0.40$, and 35% of the bias in $\hat{\mu}_\beta$.

Table 3.23: Meta Model Results: Random Effect Variance Estimate Raw Bias for Model 2, Set 1

Raw Bias				
	$\hat{\tau}_\alpha$		$\hat{\tau}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.02(-0.04, -0.01)$	$-3, p \leq .01$	$0(0, 0)$	$-1.28, NS$
$2: MCEM - WLSMV$	$-0.13(-0.15, -0.12)$	$-18.32, p \leq .0001$	$-0.01(-0.02, -0.01)$	$-12.48, p \leq .0001$
$3: 1 \times 2$	$0.03(0, 0.05)$	$1.73, NS$	$0(0, 0.01)$	$0.81, NS$

In the case of the random effect variance bias, given in Table 3.23, both sample size and estimator main effects were observed for $\hat{\tau}_\alpha$ while only estimator main effects were observed for $\hat{\tau}_\beta$. Estimates of $\hat{\tau}_\alpha$ were, significantly less biased under $N = 750$ compared to $N = 3000$ ($\beta(95\%CI) = -0.02(-0.04, -0.01)$, $t = -3, p \leq .01$), and estimates were significantly more biased for MCEM than for WLSMV ($\beta(95\%CI) = -0.13(-0.15, -0.12)$, $t = -18.32, p \leq .0001$). In the case of $\hat{\tau}_\beta$, MCEM estimates were significantly more negatively biased than those of WLSMV ($\beta(95\%CI) = -0.01(-0.02, -0.01)$, $t = -12.48, p \leq .0001$). This model accounted for 24% of the variance in bias for $\hat{\tau}_\alpha$, and 12% of the variance in bias for $\hat{\tau}_\beta$.

Table 3.24: Meta Model Results: Random Effect Covariance Estimate Raw Bias for Model 2, Set 1

Raw Bias		
	$\hat{\tau}_{\alpha\beta}$	
Contrast	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$0(-0.01, 0)$	$-1.12, NS$
$2: MCEM - WLSMV$	$-0.01(-0.01, -0.01)$	$-4.71, p \leq .0001$
$3: 1 \times 2$	$0(-0.01, 0)$	$-0.76, NS$

Table 3.24 reveals that the random effect covariance bias, differed trivially as a function of estimator, with $\hat{\tau}_{\alpha\beta}$ estimates slightly more negatively biased under MCEM then WLSMV ($\beta(95\%CI) = -0.01(-0.01, -0.01),\ t = -4.71, p \leq .0001$). This model accounted for only 2% of the bias in $\hat{\tau}_{\alpha\beta}$.

Table 3.25: Meta Model Results: $\hat{\psi}$ Raw Bias for Model 2, Set 1

Raw $\hat{\psi}$ Bias						
Results when $\hat{\psi}_i$ Aggregated			Correspondence to Aggregate by $\hat{\psi}_i$			
Contrast	$\beta(95\%CI)$	$t, pval$	ψ_1	ψ_2	ψ_3	ψ_4
$1: N3k - N750$	$-0.03(-0.05, -0.02)$	$-3.77, p \leq .001$	*	-	*	*
$2: MCEM - WLSMV$	$-0.25(-0.27, -0.24)$	$-30.15, p \leq .0001$	*	*	*	*
$3: 1 \times 2$	$0.02(-0.01, 0.06)$	$1.42, NS$	*	*	*	*

Table 3.25 demonstrates that the aggregate ψ bias was significantly less negatively biased with $N = 750$ compared to $N = 3000$, though this difference was small ($\beta(95\%CI) = -0.03(-0.05, -0.02),\ t = -3.77, p \leq .001$). In addition bias for MCEM estimates was more than double that observed for WLSMV ($\beta(95\%CI) = -0.25(-0.27, -0.24),\ t = -30.15, p \leq .0001$). Results for all individual parameters conformed to the aggregate results, except that bias for ψ_2 did not vary as a function of sample size. This model accounted for 45% of the variance in bias for ψ.

Table 3.26: Meta Model Results: $\hat{\rho}$ Raw Bias for Model 2, Set 1

Raw $\hat{\rho}$ Bias						
Results when $\hat{\rho}_i$ Aggregated			Correspondence to Aggregate by $\hat{\rho}_i$			
Contrast	$\beta(95\%CI)$	$t, pval$	ρ_1	ρ_2	ρ_3	ρ_4
$1: N3k - N750$	$-0.01(-0.01, -0.00)$	$-2.76, p \leq .01$	-	-	*	*
$2: MCEM - WLSMV$	$-0.13(-0.14, -0.12)$	$-46.55, p \leq .0001$	*	*	*	*
$3: 1 \times 2$	$0.02(0.01, 0.04)$	$4.29, p \leq .0001$	-	*	*	*

Aggregate bias for estimates of the error correlation, ρ, presented in Table 3.26, reveal significant main effects and interactions. Bias was significantly less negatively biased with $N = 750$ versus$N = 3000$, though this difference was small ($\beta(95\%CI) = -0.01(-0.01, -0.002),\ t = -2.76, p \leq .01$). In addition, MCEM estimates were significantly more biased than WLSMV estimates ($\beta(95\%CI) = -0.13(-0.14, -0.12),\ t = -46.55, p \leq .0001$). The difference between estimator bias was greater with $N = 3000$ compared to $N = 750$, though only slightly $\beta(95\%CI) = 0.02(0.01, 0.04),\ t = 4.29, p \leq .0001$). A majority of individual parameters did not exhibit a sample size effect, and ρ_1 did not demonstrate a significant estimator by sample size interaction. This model accounted for 66% of the variance in bias for ρ.

3.2.5 Second Parameter Set Bias

As can be seen in Table 3.27 with $N = 750$, the degree of bias for the correlated error model was higher than that observed for Model 1. MCEM was uniformly more biased than WLSMV for nearly every parameter, and the magnitude of observed bias was quite large. WLSMV demonstrated a preponderance of negative bias for threshold estimates. MCEM estimates were least biased for fixed effects. As noted under model 2 for the first parameter set, WLSMV estimates the error correlations directly, while MCEM estimates the root of the correlation as λ, and under the testlet parameterization in MCEM, consequently, MCEM bias values are in relation to the generating parameter for λ, which was $\sqrt{\rho} = .5477$.

Table 3.27: Item and Structural Parameter Bias for Model 2, Set 2, N=750

		N=750					
		WLSMV			*MCEM*		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.49(0.07)	0.035	0.075	0.65(0.13)	0.195	**0.424**
a_2	0.69	0.68(0.09)	-0.009	-0.013	0.91(0.18)	0.22	**0.319**
a_4	1.92	1.67(0.24)	-0.255	**−0.133**	2.71(0.55)	0.789	**0.411**
a_5	1.2	1.31(0.14)	0.105	0.088	1.64(0.32)	0.441	**0.368**
a_6	1.8	1.97(0.27)	0.168	0.093	2.31(0.46)	0.511	**0.284**

Continued on next page

Table 3.27 – continued from previous page

N=750							
		WLSMV			*MCEM*		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_7	1.68	1.24(0.17)	-0.441	**−0.263**	2.43(0.52)	0.751	**0.447**
a_8	1.76	1.83(0.22)	0.067	0.038	2.41(0.47)	0.646	**0.367**
a_9	0.3	0.32(0.05)	0.017	0.056	0.4(0.09)	0.104	**0.346**
b_1	3.3	3.1(0.22)	-0.199	-0.06	2.81(0.3)	-0.489	**−0.148**
b_2	-1	-1.03(0.32)	-0.027	0.027	-0.51(0.37)	0.485	**−0.485**
b_4	4	3.86(0.24)	-0.143	-0.036	3.33(0.39)	-0.675	**−0.169**
b_5	2.5	2.37(0.12)	-0.132	-0.053	2.23(0.2)	-0.272	**−0.109**
b_6	1.2	1.14(0.07)	-0.056	-0.046	1.24(0.11)	0.039	0.032
b_7	-1.5	-0.91(0.26)	0.589	**−0.392**	-0.13(0.28)	1.374	**−0.916**
b_8	2.8	2.65(0.14)	-0.148	-0.053	2.45(0.23)	-0.35	**−0.125**
b_9	-2	-2.13(0.63)	-0.127	0.064	-1.27(0.59)	0.733	**−0.366**
μ_α	1.39	1.26(0.07)	-0.128	-0.092	1.34(0.11)	-0.056	-0.04
μ_β	0.5	0.51(0.05)	0.006	0.012	0.41(0.07)	-0.089	**−0.179**
τ_α	0.67	0.63(0.16)	-0.041	-0.061	0.43(0.16)	-0.241	**−0.36**
$\tau_{\alpha\beta}$	0.05	0.04(0.05)	-0.003	-0.061	0.02(0.03)	-0.029	**−0.645**
τ_β	0.05	0.04(0.03)	-0.014	**−0.283**	0.02(0.02)	-0.031	**−0.621**
ψ_1	0.67	0.77(0.2)	0.099	**0.148**	0.54(0.2)	-0.129	**−0.192**
ψ_2	0.81	0.61(0.16)	-0.199	**−0.245**	0.34(0.15)	-0.471	**−0.58**
ψ_3	1.05	0.86(0.21)	-0.197	**−0.187**	0.5(0.2)	-0.553	**−0.526**
ψ_4	1.39	1.19(0.29)	-0.207	**−0.148**	0.72(0.28)	-0.672	**−0.482**
ρ_1	0.3	0.27(0.09)	-0.028	-0.093	0.36(0.14)	-0.192	**−0.351**
ρ_2	0.3	0.23(0.12)	-0.073	**−0.242**	0.36(0.28)	-0.186	**−0.34**
ρ_3	0.3	0.25(0.15)	-0.055	**−0.183**	0.33(0.22)	-0.215	**−0.392**
ρ_4	0.3	0.31(0.3)	0.008	0.028	0.21(0.46)	-0.335	**−0.611**

As can be seen in Table 3.28 quadrupling the the number of observations to $N = 3000$ did not reduce bias in a substantial manner, and for some estimates, notably among MCEM point estimates, bias increased. Again, the degree of bias for the correlated error model was higher than that observed for the model without correlated errors. As was the case with $N = 750$, WLSMV was substantially less biased than MCEM. With $N = 3000$, WLSMV threshold estimates remained negatively biased, though not uniformly. As with $N = 750$, MCEM estimates only demonstrated acceptable levels of raw bias for fixed effects and some random effect covariance parameters, though the relative bias values reflected more substantial estimation error. In some cases, MCEM was grossly biased, for example, the estimate of b_7.

Table 3.28: Item and Structural Parameter Bias for Model 2, Set 2, N=3000

N=3000							
		$WLSMV$			$MCEM$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.5(0.03)	0.04	0.087	0.69(0.1)	0.235	**0.51**
a_2	0.69	0.68(0.04)	-0.009	-0.013	0.96(0.13)	0.269	**0.39**
a_4	1.92	1.69(0.1)	-0.232	**−0.121**	2.85(0.38)	0.927	**0.483**
a_5	1.2	1.33(0.07)	0.129	**0.107**	1.75(0.24)	0.549	**0.457**
a_6	1.8	1.89(0.1)	0.091	0.051	2.44(0.34)	0.636	**0.353**
a_7	1.68	1.47(0.11)	-0.213	**−0.127**	2.57(0.38)	0.889	**0.529**
a_8	1.76	1.92(0.1)	0.157	0.089	2.55(0.36)	0.788	**0.448**
a_9	0.3	0.33(0.02)	0.028	0.092	0.43(0.07)	0.134	**0.448**
b_1	3.3	3.04(0.09)	-0.261	-0.079	2.69(0.19)	-0.611	**−0.185**
b_2	-1	-0.99(0.14)	0.007	-0.007	-0.39(0.24)	0.615	**−0.615**
b_4	4	3.76(0.11)	-0.237	-0.059	3.16(0.24)	-0.837	**−0.209**
b_5	2.5	2.32(0.05)	-0.18	-0.072	2.15(0.12)	-0.352	**−0.141**
b_6	1.2	1.13(0.03)	-0.069	-0.057	1.24(0.05)	0.035	0.029
b_7	-1.5	-0.65(0.1)	0.846	**−0.564**	-0.02(0.19)	1.48	**−0.987**
b_8	2.8	2.59(0.06)	-0.209	-0.075	2.36(0.14)	-0.443	**−0.158**
b_9	-2	-1.94(0.27)	0.061	-0.031	-1.03(0.37)	0.972	**−0.486**
μ_α	1.39	1.25(0.03)	-0.138	-0.099	1.32(0.05)	-0.067	-0.048
μ_β	0.5	0.49(0.02)	-0.009	-0.017	0.38(0.05)	-0.119	**−0.239**
τ_α	0.67	0.58(0.07)	-0.092	**−0.137**	0.36(0.1)	-0.313	**−0.467**
$\tau_{\alpha\beta}$	0.05	0.03(0.02)	-0.012	**−0.266**	0.01(0.01)	-0.031	**−0.683**
τ_β	0.05	0.04(0.01)	-0.014	**−0.278**	0.01(0.01)	-0.035	**−0.703**
ψ_1	0.67	0.69(0.08)	0.018	0.027	0.46(0.12)	-0.208	**−0.311**
ψ_2	0.81	0.57(0.06)	-0.245	**−0.301**	0.29(0.09)	-0.522	**−0.643**
ψ_3	1.05	0.79(0.08)	-0.266	**−0.253**	0.42(0.12)	-0.638	**−0.606**
ψ_4	1.39	1.09(0.12)	-0.3	**−0.215**	0.61(0.17)	-0.785	**−0.563**
ρ_1	0.3	0.28(0.04)	-0.02	-0.066	0.36(0.07)	-0.188	**−0.343**
ρ_2	0.3	0.25(0.06)	-0.053	**−0.177**	0.4(0.14)	-0.149	**−0.271**
ρ_3	0.3	0.26(0.07)	-0.036	**−0.121**	0.31(0.1)	-0.234	**−0.428**
ρ_4	0.3	0.35(0.12)	0.046	**0.154**	0.2(0.31)	-0.348	**−0.635**

3.2.6 Second Parameter Set Meta Model Results

As with model 1, parameter sets were aggregated. Thus models were fit to the average bias for $\hat{a}, \hat{b}, \hat{\psi}$, and $\hat{\rho}$, though in order to characterize any discrepancy between the aggregate and the individual parameters idiosyncracies are noted for each individual parameter. Results which diverge from the aggregate are denoted with a dash, while results which conform to that observed for the aggregate are denoted with an asterisk. For the vast majority of parameters modeled, only estimator main effects were observed.

Table 3.29: Meta Model Results: $\hat{a}$ Raw Bias for Model 2, Set 2

Raw $\hat{a}$ Bias										
Results when $\hat{a}_i$ Aggregated			Correspondence to Aggregate by $\hat{a}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	a_1	a_2	a_4	a_5	a_6	a_7	a_8	a_9
$1: N3k - N750$	$0.07(0.04, 0.09)$	$5.55, p \leq .0001$	*	*	*	*	-	*	*	*
$2: MCEM - WLSMV$	$0.53(0.5, 0.55)$	$43.5, p \leq .0001$	*	*	*	*	*	*	*	*
$3: 1 \times 2$	$-0.06(-0.11, -0.01)$	$-2.42, p \leq .05$	*	*	*	*	*	*	-	*

As can be seen from the cell means, and the model results presented in Table 3.29, bias was significantly lower when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = 0.07(0.04, 0.09),\ t = 5.55, p \leq .0001$). MCEM estimates were significantly more biased than WLSMV ($\beta(95\%CI) = 0.53(0.5, 0.55),\ t = 43.5, p \leq .0001$). In addition, the difference between MCEM and WLSMV differed as a function of sample size, with the difference in bias between estimators increasing with sample size ($\beta(95\%CI) = -0.06(-0.11, -0.01),\ t = -2.42, p \leq .05$). All of the individual $\hat{a}$ conformed to the findings observed for the aggregate except a_6, whose bias did not vary as a function of sample size, and a_8, whose bias did not have an estimator by sample size interaction. This model accounted for 63% of the variance in bias.

Table 3.30: Meta Model Results: $\hat{b}$ Raw Bias for Model 2, Set 2

Raw $\hat{b}$ Bias										
Results when $\hat{b}_i$ Aggregated			Correspondence to Aggregate by $\hat{b}_i$							
Contrast	$\beta(95\%CI)$	$t, pval$	b_1	b_2	b_4	b_5	b_6	b_7	b_8	b_9
$1: N3k - N750$	$0.01(0, 0.02)$	$2.71, p \leq .01$	*	*	*	*	*	*	*	*
$2: MCEM - WLSMV$	$0.12(0.11, 0.13)$	$25.03, p \leq .0001$	*	*	*	*	*	*	*	*
$3: 1 \times 2$	$0.02(0, 0.04)$	$2.38, p \leq .05$	*	*	*	*	-	*	-	-

As can be seen from the cell means, and the model results presented in Table 3.30, bias was significantly, though trivially lower when $N = 750$ compared to when $N = 3000$ ($\beta(95\%CI) = 0.01(0.003, 0.02),\ t = 2.71, p \leq .01$). MCEM estimates were significantly more biased than WLSMV ($\beta(95\%CI) = 0.12(0.11, 0.13),\ t = 25.03, p \leq .0001$). In addition, the difference between MCEM and WLSMV differed as a function of sample size, with the difference in bias between estimators increasing with sample size ($\beta(95\%CI) = 0.02(0.004, 0.04),\ t = 2.38, p \leq .05$). All of the individual $\hat{a}$ conformed to the findings observed for the aggregate except b_6, b_8, andb_9 whose bias did not exhibit an estimator by sample size interaction. This model accounted for 36% of the variance in bias.

Table 3.31: Meta Model Results: Fixed Effect Raw Bias for Model 2, Set 2

Raw Bias				
	$\hat{\mu}_\alpha$		$\hat{\mu}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.01(-0.02, -0.00)$	$-2.41, p \leq .05$	$-0.02(-0.03, -0.02)$	$-7.42, p \leq .0001$
$2: MCEM - WLSMV$	$0.07(0.06, 0.08)$	$16.23, p \leq .0001$	$-0.1(-0.11, -0.1)$	$-34.39, p \leq .0001$
$3: 1 \times 2$	$0(-0.02, 0.02)$	$0.17, NS$	$0.02(0.00, 0.03)$	$2.61, p \leq .01$

Bias for fixed effect estimates, $\hat{\mu}_\alpha$ and $\hat{\mu}_\beta$, are presented in Table 3.31. Estimates of $\hat{\mu}_\alpha$ were significantly, though trivially, less biased under $N = 750$ ($\beta(95\%CI) = -0.01(-0.02, -0.002),\ t = -2.41, p \leq .05$). MCEM estimates of $\hat{\mu}_\alpha$ were significantly

less biased than WLSMV estimates ($\beta(95\%CI) = 0.07(0.06, 0.08),\ t = 16.23, p \leq .0001$). In the case of $\hat{\mu}_\beta$, estimates were significantly, though trivially, less biased under $N = 750$ ($\beta(95\%CI) = -0.02(-0.03, -0.02),\ t = -7.42, p \leq .0001$). MCEM estimates were profoundly more biased than WLSMV ($\beta(95\%CI) = -0.1(-0.11, -0.1),\ t = -34.39, p \leq .0001$). Furthermore, the degree to which MCEM estimates were more biased than WLSMV estimates significantly increased as a function of sample size ($\beta(95\%CI) = 0.02(0.003, 0.03),\ t = 2.61, p \leq .01$). This model accounted for 20% of the variance in bias for $\hat{\mu}_\alpha$, and 52% of the bias in $\hat{\mu}_\beta$.

Table 3.32: Meta Model Results: Random Effect Variance Estimate Raw Bias for Model 2, Set 2

Raw Bias				
	$\hat{\tau}_\alpha$		$\hat{\tau}_\beta$	
Contrast	$\beta(95\%CI)$	$t, pval$	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.06(-0.08, -0.05)$	$-7.88, p \leq .0001$	$0(0, 0)$	$-1.65, NS$
$2: MCEM - WLSMV$	$-0.21(-0.23, -0.2)$	$-27.29, p \leq .0001$	$-0.02(-0.02, -0.02)$	$-16.34, p \leq .0001$
$3: 1 \times 2$	$0.02(-0.01, 0.05)$	$1.35, NS$	$0(0, 0.01)$	$1.86, NS$

Consistent with the results observed for model 2 under parameter set 2, in the case of the random effect variance bias, given in Table 3.32, both sample size and estimator main effects were observed for $\hat{\tau}_\alpha$ while only estimator main effects were observed for $\hat{\tau}_\beta$. Estimates of $\hat{\tau}_\alpha$ were, significantly less biased under $N = 750$ compared to $N = 3000$ ($\beta(95\%CI) = -0.06(-0.08, -0.05),\ t = -7.88, p \leq .0001$), and estimates were profoundly more biased for MCEM than for WLSMV ($\beta(95\%CI) = -0.21(-0.23, -0.2),\ t = -27.29, p \leq .0001$). In the case of $\hat{\tau}_\beta$, MCEM estimates were significantly, though only slightly, more negatively biased than those of WLSMV ($\beta(95\%CI) = -0.02(-0.02, -0.02),\ t = -16.34, p \leq .0001$). This model accounted for 41% of the variance in bias for $\hat{\tau}_\alpha$, and 20% of the variance in bias for $\hat{\tau}_\beta$.

Table 3.33: Meta Model Results: Random Effect Covariance Estimate Raw Bias for Model 2, Set 2

Raw Bias		
	$\hat{\tau}_{\alpha\beta}$	
Contrast	$\beta(95\%CI)$	$t, pval$
$1: N3k - N750$	$-0.01(-0.01, -0.00)$	$-2.87, p \leq .01$
$2: MCEM - WLSMV$	$-0.02(-0.03, -0.02)$	$-11.79, p \leq .0001$
$3: 1 \times 2$	$-0.01(-0.02, -0.00)$	$-1.97, p \leq .05$

Table 3.33 reveals that the random effect covariance bias, differed trivially as a function of all effects, with $\hat{\tau}_{\alpha\beta}$ estimates slightly more negatively biased under under $N = 3000$ ($\beta(95\%CI) = -0.01(-0.01, -0.002),\ t = -2.87, p \leq .01$) and more negatively biased under MCEM than WLSMV ($\beta(95\%CI) = -0.02(-0.03, -0.02),\ t = -11.79, p \leq .0001$). Moreover, the difference in bias as a function of estimator increased with sample size, though only slightly ($\beta(95\%CI) = -0.01(-0.02, -0.00001),\ t = -1.97, p \leq .05$). This model accounted for only 12% of the bias in $\hat{\tau}_{\alpha\beta}$.

Table 3.34: Meta Model Results: $\hat{\psi}$ Raw Bias for Model 2, Set 2

Raw $\hat{\psi}$ Bias						
Results when $\hat{\psi}_i$ Aggregated			Correspondence to Aggregate by $\hat{\psi}_i$			
Contrast	$\beta(95\%CI)$	$t, pval$	ψ_1	ψ_2	ψ_3	ψ_4
$1: N3k - N750$	$-0.08(-0.09, -0.06)$	$-9.03, p \leq .0001$	*	*	*	*
$2: MCEM - WLSMV$	$-0.34(-0.35, -0.32)$	$-39.28, p \leq .0001$	*	*	*	*
$3: 1 \times 2$	$0.01(-0.02, 0.04)$	$0.58, NS$	*	*	*	*

Results from the meta model for the aggregate ψ bias, presented in Table 3.34, reveal that estimates were significantly more negatively biased with $N = 3000$ compared

to $N = 750$ ($\beta(95\%CI) = -0.08(-0.09, -0.06)$, $t = -9.03, p \leq .0001$). In addition bias for MCEM estimates was almost triple that observed for WLSMV ($\beta(95\%CI) = -0.34(-0.35, -0.32)$, $t = -39.28, p \leq .0001$). Results for all individual parameters conformed to the aggregate results. This model accounted for 59% of the variance in bias for ψ.

Table 3.35: Meta Model Results: $\hat{\rho}$ Raw Bias for Model 2, Set 2

Raw $\hat{\rho}$ Bias						
Results when $\hat{\rho}_i$ Aggregated			Correspondence to Aggregate by $\hat{\rho}_i$			
Contrast	$\beta(95\%CI)$	$t, pval$	ρ_1	ρ_2	ρ_3	ρ_4
$1 : N3k - N750$	$-0.02(-0.03, -0.01)$	$-3.68, p \leq .001$	-	-	*	*
$2 : MCEM - WLSMV$	$-0.11(-0.12, -0.1)$	$-23.74, p \leq .0001$	*	*	*	*
$3 : 1 \times 2$	$0.08(0.06, 0.09)$	$8.31, p \leq .0001$	*	*	*	*

Aggregate bias for estimates of the error correlation, ρ, presented in Table 3.35, reveal significant main effects and interactions. Though bias was significantly less negative with $N = 750$, this difference was trivial ($\beta(95\%CI) = -0.02(-0.03, -0.01)$, $t = -3.68, p \leq .001$). MCEM estimates were substantially more biased than WLSMV estimates ($\beta(95\%CI) = -0.11(-0.12, -0.1)$, $t = -23.74, p \leq .0001$). The difference between estimator bias increased with sample size ($\beta(95\%CI) = 0.08(0.06, 0.09)$, $t = 8.31, p \leq .0001$). Though most parameters conformed to the aggregate results, ρ_1 and ρ_2 did not exhibit a sample size effect. This model accounted for 36% of the variance in bias for ρ.

3.2.7 Summary of Findings

Under Model 2, NCV solutions only occurred under MCEM estimation, with an identical sample size effect as observed under model 1, though NCV rates were higher for MCEM than any estimator under model 1. NPD solution rates under the first parameter set mirrored the effects observed for model 1, with NPD solutions restricted to $\hat{\boldsymbol{\tau}}_{\boldsymbol{\eta}}$, more

common when $N = 750$, with the majority associated with WLSMV estimates. Under the second parameter set the same pattern was observed, though NPD rates under WLSMV were the highest observed in the entire simulation.

Though MCEM was hypothesized to outperform WLSMV in terms of bias and RMSE under model 2, this was virtually never the case. Of the 29 parameters estimated in model 2 under parameter set 1 when $N = 750$, only 5 had relative bias vales *below* 10%, and in the case of $\hat{b}_7$, MCEM exhibited 160% relative bias. In contrast, though WLSMV was a far cry from exhibiting unbiased estimates, only 9 out of 29 parameters had relative bias values exceeding 10%. Both estimators had highest relative bias for estimates of thresholds, random effect variances, and variances of the time-specific $\boldsymbol{\theta}$. In addition, MCEM displayed marked relative bias in estimates of the error correlations. Out of the 29 parameters estimated in this model, only three displayed less bias under MCEM than under WLSMV, these parameters were the thresholds for $item_1$ and $item_5$ and $\hat{\mu}_\alpha$. Under the second parameter set, estimates exhibited much higher degrees of bias, with MCEM performing even more poorly than under the first parameter set, and WLSMV performing only slightly worse than under the first parameter set. When $N = 750$ only 2 of 29 MCEM estimates had relative bias below 10% while only 10 of 29 WLSMV estimates *exceeded* 10%. WLSMV was more biased than MCEM in only two parameters ($\hat{b}_6$ and $\hat{\mu}_\alpha$). Whereas high bias for WLSMV was restricted to estimates of thresholds, random effect variance estimates, variances of the time-specific $\boldsymbol{\theta}$, and error correlations, MCEM displayed uniformly high bias across all parameters in both the raw and relative bias metrics.

Quadrupling the sample size to $N = 3000$ did not help MCEM estimates, with only 4 out of 29 parameters having relative bias below 10%, while in the case of WLSMV only 9 out of 29 parameters had relative bias values *exceeding* 10%. In addition, there were only four parameters for which MCEM was less biased than WLSMV, these parameters were $\hat{b}_1,\ \hat{b}_5,\ \hat{b}_8,$ and $\hat{\mu}_\alpha$. As was the case when $N = 750$, the highest relative bias values

for both estimators were associated with estimates of item thresholds, random effect variances and covariances, variances of the time-specific $\boldsymbol{\theta}$, and the error correlations. Under the second parameter set, no real change was observed relative to $N = 750$, with both MCEM and WLSMV performing even more poorly than under the first parameter set. When $N = 3000$ only 2 of 29 MCEM estimates had relative bias below 10% while 13 of 29 WLSMV estimates *exceeded* 10%. WLSMV was more biased than MCEM for the same two parameters when $N = 3000$ as under $N = 750$ ($\hat{b}_6$ and $\hat{\mu}_\alpha$). Whereas high bias for WLSMV was restricted to estimates of thresholds, random effect variance estimates, and variances of the time-specific $\boldsymbol{\theta}$, MCEM displayed uniformly high bias across all parameters in both the raw and relative bias metrics.

Compared to model 1, Meta model results for raw bias under model 2 within the first parameter set produced more pronounced estimator differences with WLSMV uniformly less biased than MCEM. In addition to estimator effects, estimates of $\hat{\mu}_\beta$, $\hat{\tau}_\alpha$, aggregate $\hat{\psi}$, and $\hat{\rho}$, exhibited modest sample size main effects, suggesting that bias increased with sample size. The difference in bias between WLSMV and MCEM increased with sample size for estimates of $\hat{\rho}$. Under the second parameter set, estimator main effects indicated that WLSMV remained less biased than MCEM. Sample size effects were the same as under the first parameter set, indicating that bias increased with sample size for model 2. In addition, several interactions were observed under the second parameter set indicating that the difference in bias between WLSMV and MCEM for aggregate $\hat{a}$, $\hat{b}$, and $\hat{\rho}$, along with $\hat{\mu}_\beta$ and $\hat{\tau}_{\alpha\beta}$ increased as sample size increased.

3.3 Additional Analyses

3.3.1 Model 1 with MCEM

To investigate the poor behavior of MCEM in model 2, MCEM was used to estimate model 1 for the first parameter set when $N = 750$. Results are contrasted with WLSMV and FIML based on 7 QP in Table 3.36. As can be seen, MCEM was the most biased estimator of model parameters both in relative and absolute terms. Contrasting estimators on relative bias, we see that the only case in which relative bias exceeded 10% for FIML and WLSMV was for $\hat{\tau}_{\alpha\beta}$. In contrast, only 6 out of 25 parameters for MCEM had relative bias values *below* 10%. In the case of $\hat{b}_7$, whose generating value was $-.3$, MCEM exhibited 74% relative bias, and the raw bias associated with this parameter, .22, was also substantial. We can therefore conclude that the implementation of MCEM in Mplus is a poor one, especially given the optimal performance of this estimator for similar models demonstrated in Schilling and Bock (2005).

Table 3.36: MCEM Item and Structural Parameter Bias for Model 1, Set 1, N=750

N=750										
		$QP = 7$			$WLSMV$			$MCEM$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
a_1	0.46	0.46(0.04)	-0.003	-0.006	0.50(0.05)	0.037	0.081	0.540(0.09)	0.080	**0.173**
a_2	0.69	0.69(0.07)	-0.002	-0.003	0.69(0.07)	0.002	0.002	0.810(0.14)	0.120	**0.174**
a_4	0.92	0.93(0.08)	0.005	0.006	0.96(0.10)	0.043	0.046	1.086(0.17)	0.166	**0.181**
a_5	1.15	1.15(0.09)	0.002	0.002	1.20(0.11)	0.054	0.047	1.354(0.20)	0.204	**0.177**
a_6	1.37	1.37(0.12)	0.003	0.002	1.41(0.13)	0.043	0.031	1.613(0.24)	0.243	**0.177**
a_7	1.68	1.69(0.16)	0.012	0.007	1.66(0.21)	-0.020	-0.012	1.964(0.31)	0.284	**0.169**
a_8	1.76	1.76(0.14)	-0.004	-0.002	1.78(0.17)	0.021	0.012	2.047(0.28)	0.287	**0.163**
a_9	0.30	0.30(0.04)	-0.005	-0.015	0.31(0.05)	0.014	0.046	0.347(0.06)	0.047	**0.158**
b_1	2.30	2.31(0.12)	0.013	0.006	2.26(0.12)	-0.035	-0.015	2.179(0.14)	-0.121	-0.053
b_2	-0.50	-0.53(0.20)	-0.029	0.058	-0.55(0.22)	-0.049	0.098	-0.266(0.26)	0.234	**−0.468**
b_4	3.00	3.00(0.14)	0.002	0.001	2.94(0.14)	-0.065	-0.022	2.775(0.18)	-0.225	-0.075
b_5	1.50	1.50(0.07)	0.000	0.000	1.48(0.07)	-0.016	-0.011	1.482(0.08)	-0.018	-0.012
b_6	1.00	1.00(0.07)	-0.003	-0.003	1.00(0.08)	-0.003	-0.003	1.050(0.09)	0.050	0.050
b_7	-0.30	-0.31(0.13)	-0.008	0.028	-0.28(0.15)	0.019	-0.063	-0.078(0.20)	0.222	**−0.741**
b_8	2.00	2.00(0.08)	0.002	0.001	1.97(0.07)	-0.033	-0.017	1.914(0.09)	-0.086	-0.043
b_9	-1.00	-1.09(0.42)	-0.093	0.093	-1.09(0.45)	-0.085	0.085	-0.762(0.45)	0.238	**−0.238**
μ_α	1.39	1.39(0.07)	0.000	0.000	1.38(0.07)	-0.010	-0.007	1.381(0.08)	-0.011	-0.008
μ_β	0.50	0.50(0.04)	0.004	0.009	0.48(0.04)	-0.019	-0.038	0.438(0.06)	-0.062	**−0.125**
τ_α	0.67	0.69(0.13)	0.021	0.032	0.66(0.14)	-0.012	-0.019	0.519(0.14)	-0.151	**−0.225**

Continued on next page

Table 3.36 – continued from previous page

N=750										
		$QP = 7$			$WLSMV$			$MCEM$		
P	θ	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$	$\hat{\theta}(SD)$	$B_{\hat{\theta}}$	$RB_{\hat{\theta}}$
$\tau_{\alpha\beta}$	0.05	0.04(0.04)	-0.006	**−0.123**	0.04(0.05)	-0.010	**−0.214**	0.038(0.02)	-0.008	**−0.170**
τ_{β}	0.05	0.05(0.03)	0.004	0.074	0.05(0.03)	-0.001	-0.025	0.034(0.04)	-0.016	**−0.321**
ψ_1	0.67	0.67(0.15)	0.002	0.002	0.63(0.15)	-0.041	-0.062	0.473(0.15)	-0.197	**−0.293**
ψ_2	0.81	0.81(0.14)	0.002	0.003	0.76(0.15)	-0.049	-0.061	0.581(0.15)	-0.230	**−0.284**
ψ_3	1.05	1.08(0.19)	0.024	0.022	0.99(0.19)	-0.061	-0.058	0.775(0.21)	-0.278	**−0.264**
ψ_4	1.39	1.42(0.28)	0.021	0.015	1.31(0.28)	-0.086	-0.061	1.036(0.29)	-0.358	**−0.257**

3.3.2 NPD Sensitivity Analyses

For model 1 under the first parameter set a total of 38 replications had NPD solutions for $\boldsymbol{\tau}_\eta$. Sensitivity analyses for the first parameter set were conducted by removing all replications with NPD solutions for each estimator, resulting in an unbalanced design, as rates of NPD replications differed across estimator. A total of 16 replications were deleted from the 3 QP estimator replications, 12 from the 7 QP estimator replications, and 35 from WLSMV. After deletion meta-models were re-run for each parameter group and results were compared to the complete-case data. For every parameter except $\tau_{\alpha\beta}$ and the aggregated ψ, the direction of effect and patterns of significant findings were identical after deletion of replications with NPD solutions. In the case of $\tau_{\alpha\beta}$, after deleting the replications with NPD solutions, the formerly non-significant sample size effect became significant ($p = .0232$) though all other contrasts remained identical. In the case of the aggregated ψ, the formerly significant sample size effect became non-significant ($p = .094$) after deletion, though all other contrasts remained identical.

A total of 34 replications had NPD solutions for $\boldsymbol{\tau}_\eta$ under the second parameter set. Sensitivity analyses conducted after deletion of 9 replications with NPD solutions in the 3 QP estimator, and the 10 replications in the 7 QP estimator, and the 31 replications in WLSMV with NPD solutions revealed a similar pattern of findings as those observed in the first parameter set. For all parameters except $\tau_{\alpha\beta}$ and τ_β, results did not change in direction or significance after case-deletion. In the case of τ_β, the formerly significant 3

QP versus WLSMV contrast became non-significant ($p = 0.0766$), and in the case of $\tau_{\alpha\beta}$ the significant sample size main effect became non-significant ($p = 0.0986$) and the non-significant sample size by 3 QP versus WLSMV ($p = 0.0332$) and sample size by 7 QP versus WLSMV ($p = 0.0340$) interactions became significant. Thus, in the case of model 1, sensitivity analyses revealed that case-deletion only affected the results observed for parameters which could have been impacted by NPD solutions: namely, variances and covariances. It is not clear whether this is due to selection bias truncating the distribution of observed bias, or whether the NPD solutions themselves produced misleading estimates which obscured the true effects for bias in these parameters. Nonetheless, it is important to note that virtually all other contrasts were unaffected, and where change was observed, it was only for a small number of contrasts. Out of the total 144 contrasts estimated across all parameters and item parameter sets, only 6, or .042%, contrasts deviated from the observed trends once replications with NPD solutions were deleted.

For model 2, set 1, a total of 29 replications had NPD solutions under MCEM estimation and 48 replications had NPD solutions under WLSMV estimation. Sensitivity analyses revealed no change for any parameter in terms of sign or significance after deleting replications with NPD solutions under the first parameter set. In the case of the second parameter set, 63 WLSMV replications had NPD solutions and 33 MCEM replications had NPD solutions. Only the formerly significant sample size by estimator interaction for $\tau_{\alpha\beta}$ became non-significant ($p = 0.0541$) after deletion of replications with NPD solutions. For all other parameters deletion of replications produced no change under the second parameter set. Consequently, out of the 54 contrasts estimated across all parameters and item parameter sets for model 2, only 1, or .02%, showed differences under sensitivity analyses. As with model 1, differences observed for model 2 under sensitivity analyses were restricted to parameters which would be expected to be impacted, namely a covariance of a matrix identified as being degenerate upon which selection for deletion was based. We can therefore conclude that the inclusion of the NPD solutions was

associated with trivial, if any, impact on conclusions.

CHAPTER 4

Discussion

There were five key issues that arose in this dissertation: convergence, sample size effects, estimator effects, WLSMV bias and contingency table sparseness, and marginal likelihood approximations. I next expand on these themes in order to more adequately address the nature of the observed phenomena. Each of these issues ran counter to hypotheses in one way or another, and I attempt to elucidate likely mechanisms which lead to the departure of observed phenomena from expectation.

4.1 Convergence

Though hypotheses related to convergence posited that WLSMV would fail to converge more than any other estimator due to degenerate contingency tables, in fact WLSMV never failed to converge. Instead, and rather unexpectedly, NCV solutions were only observed for FIML based on 3 QP. The lack of NCV solutions for WLSMV may be explained through the fact that the algorithm implemented in Mplus augments degenerate contingency tables by inserting trivially small numbers into zero cells in order to prevent convergence failure. No such explanation exists for why FIML based on 3 QP would fail to converge, though. However, the fact that replications failed to converge under 3 QP but not under 7 QP points to an answer. If the only difference between converging and not is quadrature points, then the likely explanation is that replications which fail to converge do so because the number of quadrature points used to approximate the integrals required to obtain and maximize the marginal log-likelihood are insufficient

and approximation fails. If approximating the integrals fails, then the marginal log-likelihood can not be defined. And if the marginal log-likelihood can not be defined then it can not be maximized. Thus bad approximations to the integral render the marginal log-likelihood ill or undefined, which prevents optimization of the target function, leading to a failure to converge. To verify this hypothesis, replications which failed to converge were run individually outside of the simulation automation program and error messages were examined. In virtually every case convergence failure was caused by the log likelihood decreasing in the final iteration. Increasing the number of iterations did not fix the problem. Examination of the iterations revealed that the likelihood evaluations in these cases alternated between positive and negative values at each iteration. This behavior is characteristic of likelihood surfaces with poorly defined maxima, a phenomenon which is itself consistent with the conjecture that the cause of convergence failure was poor approximations of the marginal likelihood.

4.2 Sample Size Effect

The sample size manipulation was implemented to test the hypothesis that bias would diminish markedly by quadrupling sample size from 750 to 3000. While meta-model results indicated some significant sample size differences, differences in absolute terms were very small for bias. However, it is worth noting that, in general, RMSE exhibited a 50% reduction, as expected, when sample size was quadrupled from 750 to 3000. In the case of model 1, even though bias decreased as a function of sample size, bias was acceptably low even at $N = 750$, and thus the estimators, particularly FIML, were operating as we would expect asymptotically even at the smallest observed sample size. Under model 2, in most cases where significant sample size effects were observed, differences nearly uniformly indicated that bias *increased* with sample size. Examination of cell means reveals that the increase in bias was trivial in absolute terms. No clear explanation exists for this phenomenon. One could speculate that model 2, being a more complex model than model 1, is subject to greater imprecision and sampling

variability, and that this sampling variability would account for this counter-intuitive finding. However, we would expect sampling variability to decrease with increases in sample size. While this is true, sampling variability will decrease faster as precision increases, and in a simulation, precision increases faster with increasing replications than with increasing sample size per replication. Therefore, if properties of model 2 render it more subject to sampling variability, to counter this sampling variability more replications are required in order to fully understand the nature of estimators of this model. Given the very small differences in bias for parameters as a function of sample size, this seems viable. Nonetheless, more work is needed to understand the specific nature of bias under model 2.

4.3 Estimator Effects

Under model 1, regardless of parameter set, differences in bias and RMSE across estimators were consistent with my hypotheses. In general, FIML based on 7 QP was the least biased estimator, followed by FIML based on 3 QP, with the most biased estimator being WLSMV. However, as I noted above, the absolute magnitude of the bias observed for these estimators was uniformly (and acceptably) low under parameter set 1. Thus, while we can rank the estimators in relative terms, WLSMV did not appear to be meaningfully more biased than any other estimator. This was not the case under the second parameter set for the same model. Though estimators maintained their hypothesized relative ranking, WLSMV exhibited bias values which were much larger than the FIML estimators in absolute terms for certain parameters.

Under model 2 I hypothesized that the FIML estimator (i.e., MCEM) would outperform WLSMV. Given that MCEM is the Bayesian integration analog of adaptive Gauss-Hermite quadrature integration, and that simulation results suggest their equivalence (Schilling & Bock, 2005), MCEM seemed like a reasonable substitute for adaptive quadrature under model 2 for the estimator contrast. However, this was not the case, as the expected relative ranking was reversed, with WLSMV outperforming MCEM in

terms of both bias and RMSE. Because the estimator effects were contained within model without crossover, except for WLSMV, it is unclear whether MCEM is just poorly implemented in Mplus, or whether some aspect of model 2 contaminates the performance of MCEM. To examine MCEM further, MCEM estimation of model 1 was examined and the estimator remained the most biased and least efficient, indicating that the MCEM implementation in Mplus is currently less optimal than the quadrature-based integral approximation routines.

4.4 WLSMV Bias and Contingency Table Sparseness

Bias for WLSMV exceeded that observed for FIML estimators under the first parameter set and model 1, though the difference for most parameters was small and the absolute magnitude of bias for WLSMV was acceptably low. For the second parameter set, irrespective of model, WLSMV had high degrees of bias, and model 2 had high degrees of bias irrespective of parameter set. Examination of the bias observed for the second parameter set and model 2 bias for both parameter sets reveals that while WLSMV bias was high, item parameters for certain items *always* had the highest observed bias. Inspection of the item parameters revealed that in the case of items with even trivially negative thresholds and high discrimination WLSMV exhibited high degrees of bias for threshold estimates. The slope parameters associated with these items were the most biased among the set of slope parameters, though this bias was lower than that associated with the corresponding thresholds. The same was true for positive thresholds, though only when thresholds were large and positive and the corresponding slopes were high.

A logical explanation exists for this phenomenon. In the case of items with negative thresholds and steep slopes the probability of endorsing the item is quite high. For example, in the case of item 7 under the second parameter set $p(item = 1|\boldsymbol{\theta})$ ranges from .97 to .99 across time, under the first parameter set the probability ranged from .86 to .95, thus the one-way table is dominated by ones. In the case of items with large positive thresholds and steep slopes, the opposite is true, with the one-way tables dominated by

zeroes. For example, $p(item = 1|\boldsymbol{\theta})$ for item 4 under the second parameter set ranges from .04 to .28 over time. Examination of two-way tables associated with these items revealed high rates of sparseness, with tables tending toward complete separation of cells. In the case of item 7, two-way tables not only tended toward sparseness, but many exhibited complete separation of cells under any reasonable sample size. For example, cross tabulation under the second parameter set of items 7 and 4 having item parameters $a = 1.68$, and $b = -1.5$ and $a = 1.92$, and $b = 4$ respectively produces a degenerate two-way table when $N = 750$:

$$\begin{bmatrix} 27 & 693 \\ 0 & 30 \end{bmatrix}, \tag{4.1}$$

and the two-way table remains degenerate even with a sample size of $N = 100,000$. Because first and second order moments of the sample proportions are functions of the one and two-way tables, respectively, and WLSMV uses these moments in the calculation of thresholds and polychoric correlations (Olsson, 1979), for such items WLSMV will encounter difficulties which likely produce the observed bias. In the case of two-way tables which degenerate to complete separation of cells, the implementation of WLSMV in Mplus fills in cells with small numbers based on implied fractions from the observed sample size, and while this prevents the estimator from failing, the polychoric correlation estimates obtained must be less than optimal, leading to the extreme bias observed.

4.5 Recommendations for Application

Applied researchers can conclude from this simulation that if convergence can be obtained with 3 quadrature points per dimension of integration, FIML with few quadrature points may be expected to produce little bias in point estimates and minimal dispersion as measured by standard errors. Given the lack of computing burden, we can recommend that this procedure be employed for estimation, though if convergence issues are encountered, FIML with 7 quadrature points is optimal, with WLSMV being employed only as an option of last resort. Though WLSMV is not profoundly more biased or inef-

ficient than FIML estimators, the bias effect is highly dependent on characteristics of the data, and thus accuracy can not be predicted pre-estimation. Examination of WLSMV item parameters, paying particular attention to large slope parameters and thresholds at locations in the neighborhood of ± 2 standard deviations of $\boldsymbol{\theta}_t$, or examination of all contingency tables would give insight into the extent to which WLSMV estimates may be untrustworthy. Lastly, a lower sample size of $N = 750$ was selected because we anticipated that it would be a reasonable and commonly accepted lower bound for sample sizes considered for such models, based on the results observed in this study, it is likely that the minimally sufficient sample size for fitting these types of models is much lower than 750. As such, applications which have sample sizes in the neighborhood of $N = 750$ may be expected to perform very well. In conclusion, these models may be estimated with relatively modest sample sizes, with limited computing burden, and still produce very accurate estimates in expectation when FIML estimators are employed.

4.6 Limitations and Future Directions

Because estimators of model 1 were behaving as we would expect asymptotically even at $N = 750$, the presented simulation results prevent making conclusions regarding the lower limits of optimal performance as a function of sample size. However, information related to expected performance of the model in real-data applications is the paramount goal of any simulation. As such, in order to augment the utility of this work additional sample size cells should be added to the design in order to understand at what point below a sample size of 750 the estimators begin to degenerate and exhibit poor performance. This is important because it will both augment our understanding of the finite sample performance of the estimators, and reflect more modest sample sizes encountered in some domains of applied research, such as studies of rare events. It is likely the case that even minimal reductions in sample size will impact WLSMV heavily due to the contingency table issues described above. However, the robustness of the FIML estimators may be quite substantial, and large reductions in sample size may be required in order to observe

poor behavior on their part.

Though the second parameter set was introduced to test the robustness of the estimators to difficulty range effects, an alternative (and more realistic) process common to longitudinal data collection which induces sparseness in contingency tables is missing data. Missing data were not considered here so that focus could be placed on other elements of the design. However, this is a limitation considering the omnipresence of missing data in longitudinal data collections. The inclusion of missing data are important to maximize generalizability and external validity, but could also serve as a second contingency table sparseness condition of interest. Likely consequences of the introduction of missing data to the simulation include increased separation of estimators in terms of bias as WLSMV would most assuredly degenerate as it encountered increasingly sparse contingency tables resulting from missingness. In addition, already sparse contingency tables occurring in the second parameter set would be rendered more sparse, likely resulting in higher rates of NCV for FIML estimators based on few quadrature points. Though this might not have an impact on bias, given that the rates of NCV for FIML based on 3 QP were not associated with substantial increases in bias relative to 7 QP, increased rates of NCV are a substantial liability in application. Thus, the expected impact of missing data on the 3 QP estimator and WLSMV would likely permit stronger conclusions regarding optimizing estimation of these models under real-data conditions.

A final limitation relates to measures of model fit. Though inference for model fit under the FIML estimators remains an area for development given the sparseness of the multidimensional contingency table giving rise to the response patterns, and, in fact, fit indices are not currently reported for FIML estimators, model fit is provided under limited information estimators. Given that this information was only available for one of four estimators considered in this dissertation and interest centered around estimation, the decision was made to defer the study of this facet of modeling to a separate and subsequent project.

Despite these limitations, the key strength of this study is that it permitted an understanding of fundamental aspects of the models considered. Results provide unique insights into the nature of bias in WLSMV as it relates to contingency table sparseness, the robustness of Gauss-Hermite adaptive quadrature-based FIML estimators, the limited effect of sample size, the added complexity of the correlated error model, and the impact of difficulty range on contingency table sparseness. No single study is capable of encapsulating all aspects of a given model, as such, there are two key future directions which will help augment our understanding of this model in addition to examining missing data and smaller sample sizes.

First, in this simulation change in item behavior was assumed to be wholly accounted for by the trends in the moments of $\boldsymbol{\theta}$ over time. Given a small number of closely spaced repeated measures this is a reasonable structure under which to simulate data. An alternative, which would be reflective of more repeated measures, or widely spaced measures, or poorly constructed scales, would be data in which even after accounting for the trends in the moments of $\boldsymbol{\theta}$ the relationship between the items and the latent construct changed over time. Whether we embed this issue in the domains of differential item functioning (DIF), or measurement non-invariance with respect to time, these frameworks exist and are devoted to the examination and testing of such phenomena. Each provide standardized methods for the detection and resolution of time-varying item to construct relations. Consequently, there is no reason to believe that new procedures need be developed in order to examine this issue. I did not consider this issue here because it was deemed outside the scope of an initial study for the model of interest. However, in future work, the model developed and examined herein will be extended to examine ways in which the model integrates into the architecture of DIF and measurement non-invariance, and issues unique to the model which may require augmenting existing protocols for the detection of time-varying item to construct relations.

Second, while the scale simulated in this study was unidimensional within time (which

is representative of many well constructed scales) there do exist a number of scales which are multidimensional. For even a small number of repeated measures multidimensional scales pose a substantial problem for the proposed model, one with no clear solution given present-day computing limitations. However, it is worth emphasizing that the problem with extending the proposed model to more complex factor structures is not indicative of a liability in the model, which fits into current statistical consensus on models for repeated measures (Demidenko, 2004). Instead, the limitation is strictly related to computing power. As computing power increases or new estimation algorithms are developed, the restrictions which prevent expanding into more complex factor structures will diminish or disappear.

4.7 Conclusion

In this study I described a model for longitudinal item response data. Two variations of the model were examined: one where the correlation induced by repeated sampling was accounted for exclusively by a linear polynomial trend with fixed and random components, and one where the correlation induced by repeated sampling was accounted for by both the linear polynomial *and* time-specific error correlations for a subset of items. For each model limited and full information estimators were contrasted across two sample size conditions and two response pattern conditions. Whereas full information estimators outperformed limited information estimators in the first model, the opposite was true in the second model. No meaningful differences were observed for bias and RMSE as a function of sample size, though convergence and NPD solutions diminished as sample size increased. Wide difficulty ranges impacted bias, mostly as a function of sparse contingency tables in limited information estimation, though full information estimators also had some difficulty with items having extreme parameters. The goal of my project was to explicate and study the feasibility of fitting a model for longitudinal item response data. This model was a hybrid of two existing model types: The item response theory model and the latent growth curve model for repeated measures. I have empirically demon-

strated the utility and feasibility of this model. Future work will focus on extensions and applications, but this first step unambiguously establishes the second-order growth model as a promising statistical approach to modeling longitudinal item response data.

CHAPTER 5

Appendices

5.1 Appendix 1: Parameter set 1 RMSE

Table 5.1: Item and Structural Parameter RMSE for Model 1, Set 1, N=750

N=750							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_1	0.46	0.46(0.05)	0.045	0.46(0.04)	0.044	0.50(0.05)	0.065
a_2	0.69	0.70(0.07)	0.068	0.69(0.07)	0.067	0.69(0.07)	0.073
a_4	0.92	0.94(0.08)	0.083	0.93(0.08)	0.078	0.96(0.10)	0.107
a_5	1.15	1.16(0.09)	0.094	1.15(0.09)	0.093	1.20(0.11)	0.119
a_6	1.37	1.38(0.12)	0.118	1.37(0.12)	0.116	1.41(0.13)	0.136
a_7	1.68	1.72(0.17)	0.172	1.69(0.16)	0.164	1.66(0.21)	0.206
a_8	1.76	1.75(0.13)	0.134	1.76(0.14)	0.140	1.78(0.17)	0.172
a_9	0.30	0.30(0.04)	0.038	0.30(0.04)	0.038	0.31(0.05)	0.048
b_1	2.30	2.31(0.12)	0.122	2.31(0.12)	0.123	2.26(0.12)	0.121
b_2	-0.50	-0.50(0.19)	0.193	-0.53(0.20)	0.198	-0.55(0.22)	0.221
b_4	3.00	3.00(0.14)	0.138	3.00(0.14)	0.138	2.94(0.14)	0.152
b_5	1.50	1.51(0.07)	0.073	1.50(0.07)	0.073	1.48(0.07)	0.071

Continued on next page

Table 5.1 – continued from previous page

N=750							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
b_6	1.00	1.01(0.07)	0.074	1.00(0.07)	0.074	1.00(0.08)	0.076
b_7	-0.30	-0.28(0.13)	0.133	-0.31(0.13)	0.131	-0.28(0.15)	0.147
b_8	2.00	2.00(0.08)	0.078	2.00(0.08)	0.080	1.97(0.07)	0.080
b_9	-1.00	-1.05(0.41)	0.415	-1.09(0.42)	0.426	-1.09(0.45)	0.460
μ_α	1.39	1.40(0.07)	0.074	1.39(0.07)	0.075	1.38(0.07)	0.071
μ_β	0.50	0.50(0.04)	0.039	0.50(0.04)	0.038	0.48(0.04)	0.044
τ_α	0.67	0.69(0.13)	0.130	0.69(0.13)	0.131	0.66(0.14)	0.140
$\tau_{\alpha\beta}$	0.05	0.04(0.04)	0.044	0.04(0.04)	0.045	0.04(0.05)	0.051
τ_β	0.05	0.05(0.03)	0.028	0.05(0.03)	0.029	0.05(0.03)	0.032
ψ_1	0.67	0.64(0.15)	0.150	0.67(0.15)	0.152	0.63(0.15)	0.155
ψ_2	0.81	0.79(0.14)	0.145	0.81(0.14)	0.144	0.76(0.15)	0.158
ψ_3	1.05	1.05(0.19)	0.191	1.08(0.19)	0.194	0.99(0.19)	0.199
ψ_4	1.39	1.38(0.28)	0.283	1.42(0.28)	0.276	1.31(0.28)	0.289

Table 5.2: Item and Structural Parameter RMSE for Model 1, Set 1, N=3000

N=3000							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_1	0.46	0.46(0.02)	0.023	0.46(0.02)	0.022	0.50(0.03)	0.049
a_2	0.69	0.70(0.03)	0.034	0.69(0.03)	0.033	0.69(0.03)	0.035
a_4	0.92	0.93(0.04)	0.042	0.92(0.04)	0.040	0.95(0.04)	0.056
a_5	1.15	1.16(0.05)	0.048	1.15(0.05)	0.047	1.20(0.05)	0.075
a_6	1.37	1.38(0.06)	0.061	1.37(0.06)	0.060	1.41(0.07)	0.079
a_7	1.68	1.71(0.09)	0.093	1.68(0.09)	0.088	1.60(0.09)	0.124
a_8	1.76	1.76(0.07)	0.073	1.76(0.08)	0.078	1.79(0.09)	0.096
a_9	0.30	0.30(0.02)	0.020	0.30(0.02)	0.020	0.32(0.02)	0.029
b_1	2.30	2.30(0.07)	0.066	2.30(0.07)	0.067	2.25(0.06)	0.080
b_2	-0.50	-0.48(0.09)	0.094	-0.51(0.09)	0.095	-0.53(0.10)	0.105
b_4	3.00	3.00(0.08)	0.076	3.00(0.08)	0.075	2.93(0.07)	0.104
b_5	1.50	1.50(0.04)	0.036	1.50(0.04)	0.036	1.48(0.03)	0.040
b_6	1.00	1.01(0.03)	0.034	1.00(0.03)	0.034	1.00(0.03)	0.033
b_7	-0.30	-0.27(0.07)	0.072	-0.30(0.07)	0.069	-0.30(0.07)	0.074
b_8	2.00	2.00(0.04)	0.043	2.00(0.04)	0.043	1.96(0.04)	0.057
b_9	-1.00	-0.99(0.19)	0.191	-1.03(0.19)	0.195	-1.01(0.21)	0.210
μ_α	1.39	1.40(0.04)	0.037	1.39(0.04)	0.037	1.38(0.03)	0.037
μ_β	0.50	0.50(0.02)	0.021	0.50(0.02)	0.020	0.48(0.02)	0.032
τ_α	0.67	0.67(0.07)	0.075	0.67(0.07)	0.074	0.64(0.07)	0.082
$\tau_{\alpha\beta}$	0.05	0.04(0.02)	0.024	0.04(0.02)	0.024	0.04(0.02)	0.026
τ_β	0.05	0.05(0.01)	0.015	0.05(0.02)	0.015	0.05(0.02)	0.016

Continued on next page

Table 5.2 – continued from previous page

N=3000							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
ψ_1	0.67	0.64(0.07)	0.076	0.67(0.08)	0.076	0.63(0.07)	0.086
ψ_2	0.81	0.79(0.07)	0.077	0.82(0.08)	0.077	0.75(0.08)	0.095
ψ_3	1.05	1.03(0.10)	0.102	1.06(0.10)	0.100	0.97(0.10)	0.127
ψ_4	1.39	1.36(0.13)	0.137	1.40(0.14)	0.137	1.26(0.13)	0.185

Table 5.3: Item and Structural Parameter RMSE for Model 2, Set 1

RMSE									
		$N = 750$				$N = 3000$			
		$WLSMV$		$MCEM$		$WLSMV$		$MCEM$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_1	0.46	0.49(0.06)	0.066	0.59(0.12)	0.178	0.49(0.03)	0.042	0.60(0.07)	0.151
a_2	0.69	0.69(0.07)	0.073	0.86(0.17)	0.246	0.69(0.04)	0.038	0.86(0.1)	0.2
a_4	0.92	0.95(0.1)	0.103	1.21(0.24)	0.376	0.94(0.05)	0.052	1.21(0.13)	0.317
a_5	1.15	1.25(0.11)	0.151	1.46(0.27)	0.411	1.25(0.06)	0.116	1.46(0.16)	0.354
a_6	1.37	1.47(0.13)	0.167	1.72(0.33)	0.483	1.47(0.07)	0.121	1.73(0.18)	0.404
a_7	1.68	1.71(0.2)	0.198	2.12(0.41)	0.601	1.67(0.1)	0.098	2.12(0.24)	0.503
a_8	1.76	1.87(0.19)	0.217	2.20(0.42)	0.614	1.86(0.09)	0.136	2.20(0.24)	0.503
a_9	0.3	0.33(0.05)	0.056	0.38(0.08)	0.111	0.33(0.02)	0.038	0.38(0.04)	0.094
b_1	2.3	2.16(0.12)	0.178	2.19(0.18)	0.21	2.15(0.06)	0.161	2.16(0.09)	0.168
b_2	-0.5	-0.51(0.21)	0.213	-0.13(0.32)	0.488	-0.50(0.11)	0.11	-0.09(0.19)	0.448
b_4	3	2.81(0.14)	0.23	2.73(0.25)	0.369	2.80(0.07)	0.21	2.69(0.13)	0.335
b_5	1.5	1.42(0.07)	0.106	1.53(0.11)	0.109	1.41(0.03)	0.095	1.51(0.06)	0.057
b_6	1	0.95(0.07)	0.088	1.12(0.12)	0.17	0.95(0.03)	0.061	1.12(0.07)	0.14
b_7	-0.3	-0.15(0.13)	0.197	0.18(0.24)	0.537	-0.16(0.07)	0.155	0.20(0.15)	0.526
b_8	2	1.88(0.07)	0.138	1.93(0.13)	0.149	1.87(0.04)	0.131	1.91(0.07)	0.111
b_9	-1	-1.06(0.45)	0.454	-0.56(0.46)	0.638	-0.99(0.21)	0.207	-0.47(0.24)	0.58
μ_α	1.39	1.30(0.07)	0.114	1.42(0.11)	0.982	1.30(0.03)	0.099	1.42(0.06)	0.992
μ_β	0.5	0.47(0.04)	0.048	0.41(0.07)	0.178	0.47(0.02)	0.037	0.40(0.04)	0.114
τ_α	0.67	0.59(0.13)	0.153	0.47(0.18)	0.644	0.58(0.07)	0.109	0.44(0.1)	0.644
$\tau_{\alpha\beta}$	0.05	0.04(0.05)	0.05	0.03(0.03)	0.027	0.03(0.02)	0.025	0.03(0.02)	0.023
τ_β	0.05	0.04(0.03)	0.031	0.03(0.02)	0.476	0.04(0.01)	0.016	0.03(0.01)	0.419
ψ_1	0.67	0.62(0.15)	0.157	0.48(0.19)	0.275	0.61(0.07)	0.095	0.46(0.1)	0.258
ψ_2	0.81	0.67(0.13)	0.187	0.45(0.17)	0.297	0.67(0.07)	0.153	0.43(0.09)	0.263
ψ_3	1.05	0.9(0.18)	0.236	0.63(0.24)	0.391	0.88(0.09)	0.194	0.58(0.13)	0.309
ψ_4	1.39	1.22(0.27)	0.319	0.87(0.35)	0.114	1.17(0.13)	0.26	0.80(0.18)	0.063
ρ_1	0.3	0.28(0.08)	0.087	0.35(0.14)	0.242	0.29(0.04)	0.044	0.36(0.07)	0.201

Continued on next page

Table 5.3 – continued from previous page

RMSE									
		$N = 750$				$N = 3000$			
		$WLSMV$		$MCEM$		$WLSMV$		$MCEM$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
ρ_2	0.3	0.25(0.11)	0.121	0.29(0.22)	0.337	0.26(0.06)	0.07	0.34(0.1)	0.225
ρ_3	0.3	0.27(0.13)	0.136	0.34(0.22)	0.302	0.26(0.07)	0.078	0.32(0.12)	0.253
ρ_4	0.3	0.27(0.11)	0.116	0.35(0.19)	0.28	0.27(0.05)	0.059	0.35(0.09)	0.212

5.2 Appendix 2: Parameter set 2 RMSE

Table 5.4: Item and Structural Parameter RMSE for Model 1, Set 2, N=750

N=750							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_1	0.46	0.48(0.05)	0.053	0.46(0.05)	0.048	0.50(0.07)	0.077
a_2	0.69	0.72(0.07)	0.079	0.70(0.07)	0.069	0.68(0.09)	0.089
a_4	1.92	2.01(0.18)	0.203	1.95(0.20)	0.206	1.65(0.25)	0.366
a_5	1.20	1.24(0.10)	0.113	1.20(0.10)	0.100	1.26(0.14)	0.150
a_6	1.80	1.90(0.18)	0.206	1.82(0.18)	0.178	1.90(0.26)	0.280
a_7	1.68	1.79(0.24)	0.265	1.72(0.26)	0.262	1.00(0.15)	0.700
a_8	1.76	1.83(0.16)	0.174	1.77(0.15)	0.155	1.76(0.21)	0.205
a_9	0.30	0.31(0.04)	0.041	0.30(0.04)	0.038	0.30(0.05)	0.049
b_1	3.30	3.24(0.18)	0.187	3.32(0.18)	0.183	3.26(0.23)	0.233
b_2	-1.00	-0.93(0.23)	0.238	-1.00(0.23))	0.233	-1.06(0.31)	0.312
b_4	4.00	3.90(0.18)	0.208	4.00(0.19)	0.193	4.03(0.26)	0.265
b_5	2.50	2.45(0.10)	0.112	2.50(0.11)	0.105	2.47(0.13)	0.132
b_6	1.20	1.19(0.06)	0.064	1.20(0.07)	0.065	1.20(0.07)	0.065

Continued on next page

Table 5.4 – continued from previous page

N=750							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
b_7	-1.50	-1.42(0.24)	0.254	-1.52(0.27))	0.265	-2.45(0.49)	1.066
b_8	2.80	2.74(0.12)	0.131	2.80(0.12)	0.119	2.76(0.15)	0.156
b_9	-2.00	-1.98(0.50)	0.501	-2.08(0.52))	0.523	-2.22(0.66)	0.699
μ_α	1.39	1.37(0.07)	0.072	1.39(0.07)	0.071	1.38(0.07)	0.071
μ_β	0.50	0.49(0.04)	0.040	0.50(0.04)	0.038	0.49(0.05)	0.049
τ_α	0.67	0.64(0.13)	0.131	0.69(0.14)	0.137	0.68(0.17)	0.170
$\tau_{\alpha\beta}$	0.05	0.04(0.04)	0.042	0.04(0.04)	0.045	0.04(0.05)	0.055
τ_β	0.05	0.05(0.03)	0.026	0.06(0.03)	0.028	0.06(0.04)	0.036
ψ_1	0.67	0.61(0.14)	0.150	0.66(0.15)	0.148	0.60(0.19)	0.198
ψ_2	0.81	0.75(0.13)	0.145	0.81(0.14)	0.139	0.75(0.18)	0.190
ψ_3	1.05	0.98(0.17)	0.184	1.06(0.18)	0.182	1.00(0.24)	0.248
ψ_4	1.39	1.28(0.25)	0.275	1.39(0.26)	0.260	1.33(0.34)	0.347

Table 5.5: Item and Structural Parameter RMSE for Model 1, Set 2, N=3000

N=3000							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_1	0.46	0.48(0.02)	0.030	0.46(0.02)	0.023	0.50(0.03)	0.052
a_2	0.69	0.72(0.04)	0.047	0.69(0.04)	0.036	0.68(0.04)	0.040
a_4	1.92	1.99(0.09)	0.116	1.93(0.10)	0.099	1.64(0.11)	0.297

Continued on next page

Table 5.5 – continued from previous page

N=3000							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_5	1.20	1.24(0.05)	0.068	1.20(0.05)	0.051	1.27(0.06)	0.094
a_6	1.80	1.88(0.10)	0.126	1.80(0.09)	0.093	1.82(0.10)	0.103
a_7	1.68	1.76(0.12)	0.146	1.69(0.12)	0.123	1.19(0.11)	0.498
a_8	1.76	1.83(0.08)	0.105	1.76(0.08)	0.080	1.86(0.11)	0.150
a_9	0.30	0.31(0.02)	0.024	0.30(0.02)	0.021	0.31(0.02)	0.027
b_1	3.30	3.22(0.10)	0.125	3.31(0.10)	0.099	3.19(0.10)	0.150
b_2	-1.00	-0.92(0.12)	0.140	-1.00(0.12)	0.123	-1.06(0.14)	0.152
b_4	4.00	3.89(0.10)	0.144	4.00(0.10)	0.104	3.94(0.12)	0.132
b_5	2.50	2.45(0.05)	0.077	2.50(0.06)	0.056	2.42(0.05)	0.096
b_6	1.20	1.19(0.03)	0.032	1.20(0.03)	0.032	1.19(0.03)	0.034
b_7	-1.50	-1.41(0.13)	0.159	-1.50(0.14)	0.138	-1.92(0.24)	0.485
b_8	2.80	2.73(0.06)	0.092	2.80(0.07)	0.066	2.70(0.07)	0.123
b_9	-2.00	-1.92(0.26)	0.272	-2.03(0.27)	0.271	-2.03(0.29)	0.288
μ_α	1.39	1.37(0.04)	0.040	1.39(0.04)	0.036	1.37(0.03)	0.042
μ_β	0.50	0.48(0.02)	0.026	0.50(0.02)	0.020	0.48(0.02)	0.032
τ_α	0.67	0.62(0.07)	0.084	0.67(0.08)	0.076	0.62(0.08)	0.090
$\tau_{\alpha\beta}$	0.05	0.04(0.02)	0.023	0.04(0.02)	0.024	0.04(0.03)	0.028
τ_β	0.05	0.05(0.01)	0.015	0.05(0.02)	0.015	0.05(0.02)	0.016
ψ_1	0.67	0.63(0.07)	0.083	0.67(0.08)	0.076	0.56(0.09)	0.139
ψ_2	0.81	0.75(0.07)	0.093	0.81(0.07)	0.074	0.69(0.08)	0.145
ψ_3	1.05	0.97(0.09)	0.119	1.06(0.10)	0.096	0.93(0.10)	0.158

Continued on next page

Table 5.5 – continued from previous page

N=3000							
		$QP = 3$		$QP = 7$		$WLSMV$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
ψ_4	1.39	1.28(0.12)	0.167	1.39(0.13)	0.125	1.23(0.13)	0.211

Table 5.6: Item and Structural Parameter RMSE for Model 2, Set 2

RMSE									
		$N = 750$				$N = 3000$			
		$WLSMV$		$MCEM$		$WLSMV$		$MCEM$	
P	θ	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$	$\hat{\theta}(SD)$	$RMSE_{\hat{\theta}}$
a_1	0.46	0.49(0.07)	0.076	0.65(0.13)	0.236	0.5(0.03)	0.049	0.69(0.1)	0.254
a_2	0.69	0.68(0.09)	0.09	0.91(0.18)	0.283	0.68(0.04)	0.04	0.96(0.13)	0.298
a_4	1.92	1.67(0.24)	0.35	2.71(0.55)	0.96	1.69(0.1)	0.255	2.85(0.38)	1.003
a_5	1.2	1.31(0.14)	0.176	1.64(0.32)	0.544	1.33(0.07)	0.144	1.75(0.24)	0.599
a_6	1.8	1.97(0.27)	0.318	2.31(0.46)	0.686	1.89(0.1)	0.134	2.44(0.34)	0.719
a_7	1.68	1.24(0.17)	0.472	2.43(0.52)	0.912	1.47(0.11)	0.238	2.57(0.38)	0.965
a_8	1.76	1.83(0.22)	0.229	2.41(0.47)	0.801	1.92(0.1)	0.186	2.55(0.36)	0.865
a_9	0.3	0.32(0.05)	0.052	0.4(0.09)	0.136	0.33(0.02)	0.036	0.43(0.07)	0.149
b_1	3.3	3.1(0.22)	0.296	2.81(0.3)	0.573	3.04(0.09)	0.277	2.69(0.19)	0.638
b_2	-1	-1.03(0.32)	0.319	-0.51(0.37))	0.607	-0.99(0.14)	0.142	-0.39(0.24)	0.66
b_4	4	3.86(0.24)	0.281	3.33(0.39)	0.777	3.76(0.11)	0.26	3.16(0.24)	0.871
b_5	2.5	2.37(0.12)	0.18	2.23(0.2)	0.337	2.32(0.05)	0.187	2.15(0.12)	0.371
b_6	1.2	1.14(0.07)	0.087	1.24(0.11)	0.115	1.13(0.03)	0.075	1.24(0.05)	0.063
b_7	-1.5	-0.91(0.26)	0.643	-0.13(0.28)	1.403	-0.65(0.1)	0.852	-0.02(0.19)	1.492
b_8	2.8	2.65(0.14)	0.203	2.45(0.23)	0.422	2.59(0.06)	0.217	2.36(0.14)	0.465
b_9	-2	-2.13(0.63)	0.642	-1.27(0.59))	0.937	-1.94(0.27)	0.276	-1.03(0.37)	1.04
μ_α	1.39	1.26(0.07)	0.146	1.34(0.11)	0.984	1.25(0.03)	0.142	1.32(0.05)	1.013
μ_β	0.5	0.51(0.05)	0.049	0.41(0.07)	0.179	0.49(0.02)	0.022	0.38(0.05)	0.175
τ_α	0.67	0.63(0.16)	0.167	0.43(0.16)	0.654	0.58(0.07)	0.114	0.36(0.1)	0.656
$\tau_{\alpha\beta}$	0.05	0.04(0.05)	0.053	0.02(0.03)	0.032	0.03(0.02)	0.025	0.01(0.01)	0.032
τ_β	0.05	0.04(0.03)	0.035	0.02(0.02)	0.531	0.04(0.01)	0.019	0.01(0.01)	0.43
ψ_1	0.67	0.77(0.2)	0.22	0.54(0.2)	0.36	0.69(0.08)	0.082	0.46(0.12)	0.39
ψ_2	0.81	0.61(0.16)	0.258	0.34(0.15)	0.368	0.57(0.06)	0.253	0.29(0.09)	0.414
ψ_3	1.05	0.86(0.21)	0.285	0.5(0.2)	0.433	0.79(0.08)	0.278	0.42(0.12)	0.475
ψ_4	1.39	1.19(0.29)	0.353	0.72(0.28)	0.128	1.09(0.12)	0.322	0.61(0.17)	0.087
ρ_1	0.3	0.27(0.09)	0.09	0.36(0.14)	0.24	0.28(0.04)	0.046	0.36(0.07)	0.2
ρ_2	0.3	0.23(0.12)	0.144	0.36(0.28)	0.338	0.25(0.06)	0.081	0.4(0.14)	0.203
ρ_3	0.3	0.25(0.15)	0.161	0.33(0.22)	0.309	0.26(0.07)	0.079	0.31(0.1)	0.256
ρ_4	0.3	0.31(0.3)	0.299	0.21(0.46)	0.566	0.35(0.12)	0.132	0.2(0.31)	0.463

5.3 Appendix 3: Cell Means

Sample size effects may be computed by taking the difference between the first two rows for each parameter, using $N = 750$ as the reference. Likewise, the third through fifth rows may be used to calculate all possible estimator main effects. Sample size by estimator interactions may be computed from the remaining rows. To facilitate use of this table I give an example of the computation of all effects for the aggregate a:

- Sample Size $= N_{3k} - N_{750} = .008 - .012 = -.004$
- 3QP-7QP $= 3QP - 7QP = .009 - 0 = .009$
- 3QP-WLSMV $= 3QP - WLSMV = .009 - .02 = -.011$
- 7QP-WLSMV $= 7QP - WLSMV = 0 - .02 = -.02$
- Sample Size by 3QP-7QP $= (3QP_{750} - 7QP_{750}) - (3QP_{3K} - 7QP_{3K}) = (0.009 - 0.001) - (0.008 - 0) = 0$
- Sample Size by 3QP-WLSMV $= (3QP_{750} - WLSMV_{750}) - (3QP_{3K} - WLSMV_{3K}) = (0.009 - 0.024) - (0.008 - .017) = -.006$
- Sample Size by 7QP-WLSMV $= (7QP_{750} - WLSMV_{750}) - (7QP_{3K} - WLSMV_{3K}) = (0.001 - 0.024) - (0 - .017) = -.006$

These estimates match those presented in Table 3.4 within rounding error, and are the explicit equations used for generating the reported contrasts:

```
PROC GLM DATA= GC_ANALYSIS;
CLASS SAMPLE_SIZE QPOINTS ;
MODEL A_DIF = SAMPLE_SIZE QPOINTS SAMPLE_SIZE*QPOINTS /  CLPARM ;

ESTIMATE 'N3000 Vs. N750' SAMPLE_SIZE -1 1 ;

ESTIMATE '3QP  Vs. 7QP' QPOINTS  0 -1  1 ;
ESTIMATE '3QP  Vs. WLSMV' QPOINTS -1  1  0 ;
ESTIMATE '7QP  Vs. WLSMV' QPOINTS -1  0  1 ;
ESTIMATE 'FIML Vs. WLSMV' QPOINTS -2  1  1 / DIVISOR=2;

ESTIMATE 'DIFFERENCE BETWEEN 3QP AND 7QP  WHEN N=3K VERSUS N=750' SAMPLE_SIZE*QPOINTS  0 -1  1  0  1 -1;
ESTIMATE 'DIFFERENCE BETWEEN 3QP AND WLSMV  WHEN N=3K VERSUS N=750' SAMPLE_SIZE*QPOINTS -1  1  0  1 -1  0;
```

```
ESTIMATE 'DIFFERENCE BETWEEN 7QP AND WLSMV  WHEN N=3K VERSUS N=750' SAMPLE_SIZE*QPOINTS -1  0  1  1  0 -1;
ESTIMATE 'DIFFERENCE BETWEEN FIML AND WLSMV WHEN N=3K VERSUS N=750' SAMPLE_SIZE*QPOINTS -2  1  1  2 -1 -1 / DIVISOR=2;

ODS OUTPUT ESTIMATES=PARMS;

RUN;
QUIT;
```

The FIML main effect and interaction may be computed by taking the average of the 3 QP and 7 QP bias and constructing the appropriate contrasts. For the main effect, the difference between that average and the WLSMV cell mean is taken, and the interaction is constructed by averaging the FIML cell means stratified by sample size and taking the appropriate differences with WLSMV means stratified by sample size. These examples may be used to re-construct any contrast. But contrasts need not be constructed to interpret the direction of effect, that may be gleaned from the cell means in Table 5.7 without any further computations, though interactions are a bit harder to interpret in the absence of some calculations. This same procedure may be applied to the cell means for parameter set 2 presented in Table 5.8.

Table 5.7: Item and Structural Parameter Bias Cell Means for Model 1, Set 1

Set 1 Bias Cell Means								
Cell	a	b	μ_α	μ_β	τ_α	$\tau_{\alpha\beta}$	τ_β	ψ
N_{750}	0.012	-0.016	-0.002	-0.004	0.009	-0.008	0.002	-0.022
N_{3K}	0.008	-0.008	-0.003	-0.009	-0.011	-0.005	-0.001	-0.033
$WLSMV$	0.02	-0.03	-0.011	-0.022	-0.023	-0.01	-0.003	-0.069
$3QP$	0.009	0.004	0.005	0	0.008	-0.005	0.002	-0.022
$7QP$	0	-0.01	0	0.002	0.012	-0.003	0.002	0.009
$WLSMV_{750}$	0.024	-0.033	-0.01	-0.019	-0.012	-0.01	-0.001	-0.059
$WLSMV_{3K}$	0.017	-0.027	-0.012	-0.025	-0.033	-0.01	-0.004	-0.079
$3QP_{750}$	0.009	-0.001	0.005	0.002	0.017	-0.007	0.003	-0.018

Continued on next page

Table 5.7 – continued from previous page

Set 1 Bias Cell Means								
Cell	a	b	μ_α	μ_β	τ_α	$\tau_{\alpha\beta}$	τ_β	ψ
$3QP_{3K}$	0.008	0.008	0.004	-0.002	-0.001	-0.003	0	-0.026
$7QP_{750}$	0.001	-0.014	0	0.004	0.021	-0.006	0.004	0.012
$7QP_{3K}$	0	-0.005	-0.001	0.001	0.002	-0.001	0.001	0.006

Table 5.8: Item and Structural Parameter Bias Cell Means for Model 1, Set 2

Set 2 Bias Cell Means								
Cell	a	b	μ_α	μ_β	τ_α	$\tau_{\alpha\beta}$	τ_β	ψ
N_{750}	-0.008	-0.063	-0.013	-0.006	0.002	-0.008	0.005	-0.046
N_{3K}	-0.004	-0.04	-0.014	-0.014	-0.03	-0.005	-0.001	-0.067
$WLSMV$	-0.08	-0.138	-0.02	-0.016	-0.017	-0.008	0.002	-0.096
$QP3$	0.054	-0.011	-0.019	-0.015	-0.036	-0.007	-0.001	-0.075
$QP7$	0.009	-0.007	-0.003	0.002	0.01	-0.005	0.003	0
$WLSMV_{750}$	-0.095	-0.165	-0.015	-0.008	0.014	-0.007	0.005	-0.062
$WLSMV_{3K}$	-0.065	-0.11	-0.024	-0.024	-0.047	-0.009	-0.001	-0.13
$QP3_{750}$	0.058	-0.013	-0.02	-0.013	-0.027	-0.01	0.002	-0.074
$QP3_{3K}$	0.051	-0.008	-0.018	-0.017	-0.045	-0.004	-0.003	-0.075
$QP7_{750}$	0.015	-0.011	-0.004	0.003	0.018	-0.008	0.006	-0.003
$QP7_{3K}$	0.003	-0.004	-0.001	0	0.002	-0.002	0.001	0.003

Table 5.9: Item and Structural Parameter Bias Cell Means for Model 2, Set 1

Set 1 Bias Cell Means									
Cell	a	b	μ_α	μ_β	τ_α	$\tau_{\alpha\beta}$	τ_β	ψ	ρ
N_{750}	0.166	0.031	-0.029	-0.056	-0.136	-0.014	-0.015	-0.251	-0.093
N_{3K}	0.163	0.036	-0.035	-0.065	-0.159	-0.016	-0.016	-0.282	-0.101
$WLSMV$	0.051	-0.059	-0.092	-0.028	-0.081	-0.01	-0.008	-0.14	-0.031
$MCEM$	0.279	0.127	0.027	-0.094	-0.216	-0.019	-0.023	-0.395	-0.162
$WLSMV_{750}$	0.055	-0.062	-0.09	-0.025	-0.076	-0.009	-0.008	-0.13	-0.034
$WLSMV_{3K}$	0.047	-0.057	-0.093	-0.031	-0.085	-0.012	-0.008	-0.149	-0.029
$MCEM_{750}$	0.278	0.123	0.031	-0.087	-0.197	-0.019	-0.022	-0.372	-0.152
$MCEM_{3K}$	0.28	0.13	0.024	-0.099	-0.232	-0.02	-0.024	-0.415	-0.172

Table 5.10: Item and Structural Parameter Bias Cell Means for Model 2, Set 2

Set 2 Bias Cell Means									
Cell	a	b	μ_α	μ_β	τ_α	$\tau_{\alpha\beta}$	τ_β	ψ	ρ
N_{750}	0.394	0.2	-0.092	-0.042	-0.141	-0.016	-0.023	-0.291	-0.072
N_{3K}	0.425	0.206	-0.107	-0.057	-0.188	-0.02	-0.023	-0.346	-0.079
$WLSMV$	0.166	0.146	-0.133	-0.002	-0.068	-0.008	-0.014	-0.165	-0.025
$MCEM$	0.688	0.269	-0.061	-0.104	-0.275	-0.03	-0.033	-0.495	-0.133
$WLSMV_{750}$	0.146	0.132	-0.128	0.006	-0.041	-0.003	-0.014	-0.126	-0.037
$WLSMV_{3K}$	0.184	0.157	-0.138	-0.009	-0.092	-0.012	-0.014	-0.198	-0.016
$MCEM_{750}$	0.642	0.268	-0.056	-0.089	-0.241	-0.029	-0.031	-0.456	-0.107
$MCEM_{3K}$	0.739	0.27	-0.067	-0.119	-0.313	-0.031	-0.035	-0.538	-0.161

References

Bartholomew, D. J., & Knott, M. (1999). *Kendall's library of statistics 7, latent variable models and factor analysis*. London: Arnold Publishers.

Birnbaum, A. (1968). Statistical theories of mental test scores. In F. M. Lord & M. R. Novick (Eds.), (chap. Some latent trait models and their use in inferring an examinee's ability). London: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the em algorithm. *Psychometrika*, *46*, 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, New Jersey: Wiley.

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods and Research*, *29*, 468-508.

Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5-32.

Curran, P. J., Hussong, A. M., Cai, L., Huang, W., Chassin, L., Sher, K. J., et al. (2008). Pooling data from multiple longitudinal studies: The rols of item response theory in integrative data analysis. *Developmental Psychology*, *44*, 365-380.

Demidenko, E. (2004). *Mixed models: Theory and applications*. New York: Wiley.

Gibbons, R., & Hedeker, D. (1997). Random effects probit and logistic models for three-level data. *Biometrics*, *53*, 1527-1537.

Hill, C. D. (2006). *Two models for longitudinal item response data*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill, Chapel Hill, NC.

Johnson, C., & Raudenbush, S. W. (2006). Methodolgical issues in aging research. In C. S. Bergeman & S. M. Boker (Eds.), (chap. A repeated measures, multilevel Rasch model with application to self-reported criminal behavior). Hillsdale, NJ: LEA.

Joreskog, K., & Moustaki, I. (2001). Factor analysis for ordinal variables: a comparison of three approaches. *Multivariate Behavioral Research*, *36*, 347-387.

Liu, L. C. (2008). A model for incomplete longitudinal multivariate ordinal data. *Statistics in Medicine*, *27*, 6299-6309.

Liu, L. C., & Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*, 261-268.

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

McCullough, P., & Nelder, J. (1989). *Generalized linear models*. London: Chapman and Hall/CRC.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*, 23-43.

Muthén, B. O. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551-560.

Muthén, B. O., & Muthén, L. K. (2007). *Mplus user's guide, fifth edition*. Los Angeles, CA.: Muthén and Muthén.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, *31*, 214-225.

Nesselroade, J. R., & Baltes, P. B. (1979). *Longitudinal research in the study of behavioral development*. New York: Academic Press.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*, 443-460.

Pearson, K. (1901). Mathematical contributions to the theory of evolution, vii: On the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London, Series A*, *195*, 1-47.

Schilling, S., & Bock, R. D. (2005). High dimensional maximum marginal likelihood item factor analsysis by adaptive quadrature. *Psychometrika*, *70*, 533-555.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York: Chapman and Hall/CRC.