

ABSTRACT

Title of Dissertation: THE MULTIDIMENSIONAL GENERALIZED GRADED UNFOLDING MODEL FOR ASSESSING CHANGE IN REPEATED MEASURES

Weiwei Cui Doctor of Philosophy, 2008

Directed by: Associate Professor James S. Roberts
Department of Psychology

A multidimensional extension of the generalized graded unfolding model for repeated measures (GGUM-RM) is introduced and applied to analyze attitude change across time using responses collected by a Thurstone or Likert questionnaire. The model conceptualizes the change across time as separate latent variables and provides direct estimates of both individual and group change while accounting for the dependency among latent variables. The parameters and hyperparameters of GGUM-RM are estimated by fully Bayesian estimation method via WinBUGS. The accuracy of the estimation procedure is demonstrated by a simulation study, and the application of the GGUM-RM is illustrated by the analysis of attitude change toward abortion among college students.

THE MULTIDIMENSIONAL GENERALIZED GRADED UNFOLDING MODEL
FOR ASSESSING CHANGE IN REPEATED MEASURES

By
Weiwei Cui

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:
Associate Professor James Roberts, Chair
Professor Chan Dayton
Professor Robert Mislevy
Professor Robert Lissitz
Professor Paul Hanges

DEDICATION

This dissertation is dedicated to:
My parents, my husband, and my daughter.

ACKNOWLEDGEMENTS

I would like to thank the following individuals and the Department of Measurement, Statistics, and Evaluation. To Dr. James Roberts for his patient and intelligent supervising through this study; to Dr. Chan Dayton for his encouragement and support through my doctoral program; to Dr. Robert Mislevy for his insightful comments and suggestions through the study; to Dr. Robert Lissitz for bringing me to this program and his thoughtful and experienced guidance through my doctoral program; Dr. Paul Hanges for his thoughtful comments and suggestions for this study.

I would also like to thank Otto Gonzalez for allowing me to run the simulation using the computer lab; and Hi Shin Shim, Vanessa Thompson, J. Daniel Gordon, Zachary Morford, Alisha Monteiro, Casey Bowden and Sheliza Bhanjee for collecting the responses in the Georgia Institute of Technology sample.

Table of Contents

List of Table.....	v
List of Figures	vi
I. Research Objective	1
II. Introduction	2
Problematic Issues in Classical Test Theory Approaches	2
Assessing Individual Change over Time Using Item Response Theory	5
Unidimensional IRT Approach	8
Measuring Change With Multidimensional IRT Models	10
III. Unfolding IRT Approach to Measure Change across Repeated Assessments	20
Modeling Proximity-based Response Processes	20
Directly Assessing Change in Repeated Measures Designs Using the GGUM... ..	24
IV. A Parameter Recovery Simulation	32
Simulation Design	33
Data Analysis and Evaluation of Results.....	37
V. A Real Data Example	42
VI. Results	44
Accuracy of Parameter Recovery	44
Real Data Analyses	61
VII. Discussion and Conclusions	76
Discussion	76
Limitations	81
Conclusion	82
Appendix A. Time-Series Plots for Alpha, Delta and Taus for Two Chains of 10000 Iterations for 10 Items, and Thetas for First Five Persons for Two Chains of 10000 Iterations	84
Bibliography	97

List of Tables

Table 1	ANOVA Effect for the Analysis of the RMSD of Estimates of Item Parameters and Person Parameters	45
Table 2	ANOVA Effect for the Analysis of the Absolute Bias of Parameter Estimates	46
Table 3	ANOVA Effect for the Analysis of the Variance Ratio (<i>VR</i>) of Estimates of Item Parameters and Person Parameters	46
Table 4	ANOVA Effect for the Analysis of the Correlation (<i>r</i>) of Estimates of Item Parameters and Person Parameters	47
Table 5	GGUM-RM Item Parameter Estimates ($\hat{\delta}_i$, $\hat{\alpha}_i$, and $\hat{\tau}_{ik}$) for 19 Abortion Attitude Statements	64
Table 6	Responses for 4 Respondents with Absolute Change Estimates Greater than 1	75

List of Figures

Figure 1	Expected Value of an Observed Response to a Hypothetical Four-Category Item as a Function of α_i and τ_{ik}	23
Figure 2	Mean Accuracy Measures for Alpha Estimates for Same Form and Alternate Forms	53
Figure 3	Mean Accuracy Measures for Delta Estimates for Same Form and Alternate Form	54
Figure 4	Mean Accuracy Measures for Delta Estimates for 10, 20, and 30 Items	55
Figure 5	Mean Accuracy Measures for Tau Estimates for Same Form and Alternate Form	56
Figure 6	Mean Accuracy Measures for Tau Estimates for 10, 20, and 30 Items	57
Figure 7	Mean Accuracy Measures for Theta Estimates for 10, 20, and 30 Items	58
Figure 8	Interaction of and Test Length for RMSD and Correlation of Individual Change Estimates	59
Figure 9	RMSD of Group Change Estimate 1000 and 2000 Respondents	60
Figure 10	RMSD of Estimated Variance of Latent Distribution for 10, 20, and 30 Items	60
Figure 11	Scatter Plot of Estimates of Item Location Parameters at Baseline against Estimates of Item Location Parameters at the Second Assessment Time	62
Figure 12	Scatter Plot of Estimates of Item Discrimination Parameters at Baseline against Estimates of Item Discrimination Parameters at the Second Assessment Time	62
Figure 13	Scatter Plot of Estimates of Item Threshold Parameters at Baseline against Estimates of Item Threshold Parameters at the Second Assessment Time	63
Figure 14	Estimated Item Locations for 19 Abortion Items	67
Figure 15	The ICC for Item 14.....	67
Figure 16	Average Observed and Expected Responses by Theta Group for Item 1-4	69
Figure 17	Average Observed and Expected Responses by Theta Group for Item 5-8	70
Figure 18	Average Observed and Expected Responses by Theta Group for Item 9-12	71
Figure 19	Average Observed and Expected Responses by Theta Group for Item 13-16.....	72
Figure 20	Average Observed and Expected Responses by Theta Group for Item 17-19	73
Figure 21	Scatter Plot of Estimate of Individual Change against Their for Initial Level	74

I. Research Objectives

In many areas of educational and psychological measurement, measuring change over time is of great interest. This dissertation focuses on measuring change over time in the context of item response theory (IRT). Specifically, a new multidimensional extension of the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) is developed and explored. The traditional GGUM is a unidimensional IRT model for unfolding binary or graded responses to Likert or Thurstone style attitude, satisfaction or preference questionnaires. The model postulates a proximity-based response process between an individual and an item such that higher item scores are expected to the extent that the individual is located close to the item on a unidimensional latent continuum. The traditional GGUM is extended to the multidimensional case in this project in an effort to quantify changes in these types of proximity-based constructs over time.

This dissertation first reviews the seminal literature on the measurement of change from both a classical test theory and an item response theory perspective. A new multidimensional version of the GGUM is then described for the measurement of change in proximity-based constructs using repeated measures designs. A simulation study is designed and conducted to assess the data demands required for accurate recovery of model parameters using a fully Bayesian estimation method implemented via the WinBUGS computer program. Finally, the model is applied to real data from a self-report questionnaire designed to measure one's attitude to abortion in a repeated measures context.

II. Introduction

Problematic Issues in Classical Test Theory Approaches to the Measurement of Individual Change over Time

The earliest discussions on assessing change over time generally focused on classical test theory (Thorndike, 1924; Garside, 1956; Bereiter, 1963; Lord, 1962, 1963). Within the classical test theory (CTT) framework, change over time is typically measured as the difference between raw test scores from successive test administrations. In the simplest case, this change would be measured as the difference between raw pretest and posttest scores. Pretest and posttest scores are measured at a manifest variable level, as is the change score derived by differencing them. In other words, they are both considered observed scores in CTT. CTT formulates the observed score as a function of a true score and an error term (i.e., measurement error):

$$X_j = T_j + E_j \quad (1)$$

where

X_j is the observed score for the j^{th} person,

T_j is the true score for the j^{th} person, and

E_j is the measurement error for the j^{th} person.

The CTT model assumes that the measurement errors are distributed with a mean of zero.

Therefore, the true score is equal to the expected value of the observed test score.

Suppose we denote two observed test scores for person j as X_{j1} and X_{j2} . The CTT model represents the difference between these two observed test scores as:

$$D_j = X_{j2} - X_{j1} = T_{j2} + E_{j2} - (T_{j1} + E_{j1}) = T_{j2} - T_{j1} + E_{j2} - E_{j1} = T_{jD} + E_{jD} \quad (2)$$

where

D_j is the observed difference score for the j^{th} person,

T_{j2} is the true score for the j^{th} person on the second test,

T_{j1} is the true score for the j^{th} person on the first test,

E_{j2} is the error score (i.e., measurement error) for the j^{th} person on the second test,

E_{j1} is the error score for the j^{th} person on the first test,

T_{jD} is the true difference between the two tests for the j^{th} person, and

E_{jD} is the difference in error scores for the j^{th} person on the two tests.

As we will see below, this simple conception of a difference score within CTT can lead to some interesting psychometric problems.

Reliability of Change Scores. The reliability paradox in CTT when measuring change over time is widely recognized. Bereiter (1963) pointed out that the reliability of change scores decreases as the correlation between pretest and posttest scores increases, holding other conditions constant. Lord (1963) mathematically showed that this would happen when using differences between pretest and posttest scores as the measure of change. However, a stable test structure that measures the same content domain over time is generally preferred when measuring change over time, and thus, strong correlations are usually seen as an advantage. This results in a paradox because researchers desire pretest and posttest scores that correlate to a large extent, but reliability of the corresponding change score decreases as the correlation increases.

In contrast to the perspectives of Bereiter (1963) and Lord (1963), some researchers have argued that change scores can be reliable even though the reliability paradox holds mathematically (Zimmerman and Williams, 1982, 1998; Regosa and Willet's 1983; Williams & Zimmerman, 1996, 1999). One can see this by examining

Equation 2. If there were few individual differences in the amount of change in true score (i.e., T_{jD} is fairly constant across examinees), then the variance of the resulting difference score would be primarily a function of the difference between two sources of measurement error (i.e., E_{jD}). In that case, the reliability of the observed difference score would be low. In contrast, if there was substantial individual variation in T_{jD} relative to E_{jD} , then the observed difference score could be quite reliable.

The Meaning of Change Scores Constructed from Different Initial Score Levels.

Bereiter (1963) indicated that measuring and comparing change scores from pretest to posttest implied that equal raw score changes at various points on the scale correspond to equal changes in the trait being assessed. In other words, the pretest and posttest scores must be measured at the interval scale level in order to correctly interpret change scores. However, classical test theory makes no presumptions about the level of measurement (i.e., ordinal, interval, etc.) that is achieved for either the pretest or posttest, and thus, the corresponding change score (or some linear transformation of it) will not necessarily represent an interval scale. This is especially the case in educational and psychological testing practice. Fischer (2003) argued that a compression of the scale is bound to occur near the boundaries of the score (i.e. floor and ceiling effects are common when using raw scores). Consequently, change scores for individuals with different initial score levels will generally have different meanings. For example, a small change score for an individual with a high initial score may have a different meaning than the same amount of change for an individual with a moderate initial score.

Negative Correlation of the Change Score and Initial Score. Some researchers (Thorndike, 1924; Bereiter, 1963; Lord, 1963) observed that difference scores are

negatively related to the pretest scores. Two main reasons have been used to explain the occurrence of the negative correlation between baseline measures and change estimates. One is ceiling and floor effects (Wilder, 1967). As scores approach the boundaries of the scale, there is not enough scale space for individual to express his or her response. This may produce smaller changes for individuals at the one extreme end than those at the other extreme end of the scale. Another reason is the regression effect caused by errors of measurement (Lord, 1963; Cronbach & Furby, 1970). Lord (1963) showed that this negative correlation is due to the fact that the difference score has a measurement error component that is opposite in sign but identical in absolute value to the measurement error in the pretest score. This is apparent in Equation 2 where E_{jD} is equal to $E_{j2} - E_{j1}$. Thus, when considering Equations 1 and 2, both the initial observed score and the observed difference score are a function of E_{j1} , although the signs of these functions are reversed. The CTT model suggests that all manifest test scores generally contain some degree of measurement error. Thus, when measurement error is not negligible, this problem is unavoidable for measures of change based on CTT.

Assessing Individual Change over Time Using Item Response Theory

Item response theory (IRT) is a widely used modern psychometric technique in educational and psychological assessment. It describes the interaction between a person and a categorically scored item assuming that the characteristics of the person can be represented by one or more hypothetical constructs. The hypothetical construct(s) is modeled as latent variable(s) within IRT and it can be unidimensional or multidimensional depending on the underlying theory. IRT models represent item and

person characteristics as model parameters that are calibrated simultaneously. This yields item characteristics and person locations (i.e., scores) that are measured on latent scale level.

An IRT model offers several benefits when it fits a given set of item response data. It provides invariant interpretations of person parameters regardless of the distributional characteristics of test or survey items. It also provides interpretations of item parameters that are invariant to the distribution of respondents on the latent continuum. Finally, it allows one to approximate the precision that has been achieved when estimating a given model parameter. This, in turn, enables one to estimate how well each individual has been measured using a given test or survey.

In testing practice, repeated measures designs are often used to assess individual change over time. The same test/questionnaire or alternate tests/questionnaires are repeatedly administered to the same individuals at each time point. When the same test/questionnaire is used, there is some potential for invalidity due to memory or learning effects caused by repeated presentation of the same items. In order to alleviate these contaminating effects, alternate test/questionnaire forms may be used at different assessment administrations. If alternate forms are used, they must be linked to common metric. A common metric is often achieved by placing common items across alternate forms, and then using responses to the common items to link the origin and scale of item parameters across forms. Two types of alternate forms are typical in testing practice. With one type, items on all forms are constructed to possess similar locations on the latent continuum. With the other type, items are constructed to have systematically

different locations across forms. For example, attitude items may be constructed to assess more extreme attitudinal positions across forms.

Both unidimensional and multidimensional IRT models have been used to assess change in repeated measures designs, though they conceptualize and characterize change over time in different ways. When an IRT model is used to assess individual change over time based on responses to a repeatedly administered test or survey, it is usually assumed that a person's position on the latent continuum varies over time, whereas the characteristics of test or survey items are generally held fixed across the repeated administrations.

IRT approaches to measuring individual change avoid some of the problems addressed by Lord (1963) and Bereiter (1963). When specific IRT model parameters are used to represent individual change, these parameters do not depend on the correlation of successive pairs of raw test scores (Embretson, 1991). Instead, the precision of the individual change estimate is evaluated by the inverse of information matrix. Consequently, the reliability paradox is a moot issue when IRT is used to directly measure individual change. Additionally, the nonlinear relationship that is typically assumed between raw scores and latent traits in most IRT models suggests that the same change in raw score may be associated with different amounts of change in the latent trait depending on the magnitude of the initial raw score (Embretson, 1991; Roberts & Ma, 2006). Appropriate use of IRT models also helps to remove ceiling/floor effects, which are both possible reasons for the negative correlation between initial levels and change estimates. However, measurement errors still exist even when IRT models are appropriately used. Thus, unless we have perfect measures, this negative correlation

between initial score and change estimates will also be observed when IRT approaches are used to estimate individual change. Some unidimensional and multidimensional IRT approaches will be introduced in the following two sections.

Unidimensional IRT Approach

Unidimensional IRT models assume that a person's response to a test item is determined by only one underlying theoretical construct. For a repeated measures design, either the same test/questionnaire or alternate test/questionnaire forms are given to the same individuals at each assessment time point. Then, person parameters based on an IRT model are estimated at each assessment time point and the intra-subject difference between estimates provides a measure of each individual's change. This approach requires that the model should hold for the data at each time point. The approach also presumes that item parameters remain stable over time.

When the unidimensional IRT approach is used, the metric of the model parameters must be linked through common items across test/questionnaire forms. Two methods can be used to establish a common metric when using IRT models: separate calibration (Kolen & Brennan, 2004) and simultaneous calibration (Lord, 1980). These two methods are briefly described below.

Separate calibration. The model parameters are estimated separately at each time point. The θ scale for one time point, such as the starting point, is chosen as the baseline metric, and then responses to common items are used to place estimates of model parameters derived from responses to other forms onto the baseline metric using standard linking methods (e.g., mean/mean, mean/sigma, or item characteristic curve methods).

After the parameter estimates for the separate calibrations are placed on the same scale, then measures of individual change are derived simply by differencing a given person's θ estimates at successive time points.

Separate calibration estimates all model parameters independently at each time point. Consequently, estimates of item parameters for common items can drift over time because there is nothing in the method that constrains parameters for common items to be equal across calibrations. This violates the assumption that common item parameters must be stable across time when using this technique.

Simultaneous calibration. With simultaneous calibration (a.k.a. concurrent calibration), individual responses from all time points are used to estimate all model parameters in a single calibration run. Responses from each assessment time point are treated as though different individuals responded to each form. The responses from each assessment period are then combined into one data set that is subsequently analyzed. Responses to common items are available for all time points, but responses to the unique items are only available for the single corresponding assessment time. Responses are coded as “not reached” or “missing” for the unique items not taken at a given assessment time. When these data are subsequently analyzed with a unidimensional IRT model, the resulting estimates of item parameters and person parameters will be on the same metric. As in the separate calibration method, individual change is measured simply by differencing a person's θ estimates at successive time points.

Simultaneous calibration allows the prior distribution of the latent trait to differ at each time point if responses from each time point are treated as though they are from multiple groups. Additionally, the parameters characterizing common items are forced to

be equal across time points, and thus, there is no apparent item drift. However, it does not account for the correlations of latent trait distributions across time.

Measuring Change With Multidimensional IRT Models

Multidimensional IRT (MIRT) models assume that a person's response to a test item is determined by more than one underlying theoretical construct and represent all relevant constructs within a single IRT model. MIRT has been used to assess change over time in two ways. One way is to conceptualize the person's latent trait at each time point as one dimension, and thus, each individual has a latent profile across time. As with the unidimensional IRT approach, change is assessed by the differencing the latent trait estimates between successive time points. However, the multivariate approach provides a benefit that the unidimensional approach does not. Specifically, the multivariate approach provides direct estimates of the correlation among the latent variables in the profile. A second way of implementing a multivariate IRT approach to change assessment is to conceptualize the change between two adjacent time points as a separate dimension. A respondent's "composite" latent trait at each time point is then calculated as the sum of the baseline latent trait and all subsequent changes between latent traits at preceding adjacent time points. This composite latent trait can be used to examine traditional IRT formulations such as the item characteristic function or the item information function, etc. An advantage of this second IRT approach is that change in the latent trait is parameterized directly in the model rather than deriving it after the fact by subtracting latent trait profile estimates. Additionally, as the first approach, the correlation matrix can be estimated directly.

Several MIRT models have been developed to assess change over time by either conceptualizing the θ at each time point as a separate dimension or the change over adjacent time points as a separate dimension (Andersen, 1985; Fischer and Pazer 1991; Embretson 1991; Fischer and Ponocny 1994; Fischer 2003; Wang, Wilson & Adams, 1998; Wang & Chyi-In, 2004; Reckase & Martineau, 2004; Roberts and Ma 2006; te Marvelde, Glas, Van Landeghem & Van Damme, 2006). Some of the models are briefly described and compared below.

Andersen's model. A multidimensional Rasch type IRT model was proposed by Andersen (1985) to estimate change in latent trait scores for repeated measures designs. Like all the multidimensional approaches discussed in this section, this method estimates the correlation among latent trait scores across time. Assume that same test/questionnaire is given to same individuals at each time point and the following conditions hold:

- 1) The probability of a response to any test item follows Rasch model.
- 2) An individual's response vector at a given time point is conditionally independent from the individual's response vectors at other time points given the latent traits associated with the given time point.
- 3) Item parameters are invariant across time points.
- 4) Response vectors for different individuals are independent.

The probability of a "correct" response from the j^{th} individual to the i^{th} item at time t can be defined as:

$$P(X_{j i t} = 1 | \theta_{j t}) = \frac{\exp(\theta_{j t} - b_i)}{1 + \exp(\theta_{j t} - b_i)} \quad (3)$$

where

θ_{jt} is the latent trait parameter for the j^{th} individual at assessment time t (i.e., the person parameter) ,

b_i is the location parameter of the i^{th} item on the latent continuum (i.e., the item parameter).

Because Andersen's model is a Rasch type IRT model, sufficient statistics exist for both item and person parameters and these parameters can be calculated using a conditional maximum likelihood estimation procedure. The multiple latent trait estimates for a given individual constitute a profile. Change is then assessed by calculating the difference of latent trait estimates between adjacent time points. Andersen (1985) also showed that correlations among latent trait densities could be estimated using this model.

Andersen's model is only appropriate for repeated measures designs that use the same test/questionnaire at each assessment point, not for alternate test/questionnaire forms, which are often used in testing practice, too. This property may limit the application of the model. Moreover, users of this model would assess change over time by constructing differences between latent trait scores in a given profile, which implies that the change over time is not directly estimated. Whether this is a reasonable strategy depends on the researcher's primary interest. For example, if the main purpose of the research is to estimate a respondent's latent trait score at each time point while accounting for the correlation among latent trait distributions, this model is appropriate and preferred. If the researcher is interested in assessing individual change over time, then this model suffers from the fact that it does not provide a direct estimate of such change. Estimating change by constructing differences between adjacent latent trait

scores suffers from the same reliability paradox that was previously mentioned for the gain score approach from CTT (Lord, 1963; Roberts & Ma, 2006).

Roberts & Ma (2006) also pointed out that, although Andersen's model is multivariate in form, it is conceptually univariate in nature. They argued that Andersen's model is mathematically multivariate in form, but focuses on only one single variable, as in the multivariate approach to repeated measure in analysis of variance (ANOVA) model.

Multidimensional Rasch Model for Measuring Learning and Change (MRMLC). Embretson's (1991) developed the MRMLC model to directly assess individual change over time and account for the correlation between latent trait scores across time. One latent dimension is postulated to represent the initial (a.k.a. baseline) latent trait, and one or more additional dimensions are used to represent latent change over time. Embretson referred to these latent change scores as "modifiabilities". The probability function of MRMLC can be mathematically defined as:

$$P(X_{ji(t)} = 1 | \theta_{j1}^*, \dots, \theta_{jT}^*) = \frac{\exp\left(\sum_{q=1}^t \theta_{jq}^* - b_i\right)}{1 + \exp\left(\sum_{q=1}^t \theta_{jq}^* - b_i\right)} \quad (4)$$

Where

$X_{ji(t)}$ represents a response from the j^{th} individual to the i^{th} item where that item may be administered at one or multiple assessment times, each of which is indexed by t , $q = 1, 2, \dots, T$ is a given assessment time point (i.e., q is simply a counter that indexes time points between 1 and t), $t = 1, 2, \dots, T$ is the assessment time point for the response in question,

$\theta_{j1}^* = \theta_{j1}$ is latent person parameter of the j^{th} respondent for the baseline (time 1) level,
 $\theta_{j2}^* = \theta_{j2} - \theta_{j1}$ is the latent change parameter of the j^{th} respondent from time 1 to time 2,
 \dots ,
 $\theta_{jt}^* = \theta_{jt} - \theta_{j(t-1)}$ is the change parameter of the j^{th} respondent from time $t-1$ to time t with t
 $= 3, \dots, T$, and

b_i is the location parameter of the i^{th} item on the latent continuum.

In practice, items may be unique to a given form or may be administered on several, if not all, forms.

The MRMLC is appropriate for binary responses to either the same form or alternate forms with common items among them. Because the model is a Rasch-type IRT model, Embretson (1991) was able to derive conditional maximum likelihood estimates of item and person parameters. She showed that in addition to the person and item parameters, the MRMLC also provides direct estimates of the hyperparameters of latent variable distributions. Estimates of the associated mean vector and variance-covariance matrix for θ_{jt}^* can be derived using traditional maximum likelihood techniques. These quantities are denoted here as:

$$\mu = \left[\mu_{\theta_{j1}^*}, \mu_{\theta_{j2}^*}, \dots, \mu_{\theta_{jt}^*} \right], \text{ and}$$

$$\Sigma_{\theta} = \begin{bmatrix} \sigma^2_{\theta_{11}^*} & \sigma_{\theta_{12}^*} & \dots & \sigma_{\theta_{1T}^*} \\ \sigma_{\theta_{21}^*} & \sigma^2_{\theta_{22}^*} & \dots & \sigma_{\theta_{2T}^*} \\ \dots & \dots & \dots & \dots \\ \sigma_{\theta_{T1}^*} & \sigma_{\theta_{T2}^*} & \dots & \sigma^2_{\theta_{TT}^*} \end{bmatrix}.$$

The variance-covariance matrix can be transformed into a correlation matrix, in which the off diagonal elements represent the correlation among latent variables.

When using this model, the sum of the initial latent trait level and all the latent change scores up to the t^{th} time point can be used to construct a composite latent trait variable at time t :

$$\theta_{jt} = \sum_{q=1}^t \theta_{jq}^* \quad (5)$$

where

$q = 1, 2, \dots, t$ is a given assessment time point (i.e., q is simply a counter that indexes time points between 1 and t),

θ_{jt} is the composite latent trait parameter for the j^{th} respondent at assessment time point t , when $t=1$; it equals the value of initial latent trait, and

θ_{jq}^* is the latent change parameter of the j^{th} respondent from time $q-1$ to time q .

This composite variable can be used to produce traditional IRT formulations like item characteristic curves, item information curves, etc. Within a given time point, an individual's response to all items depends on the composite latent trait at that time point. That is, a unidimensional IRT model based on the composite variable holds for all items within a given time point. Embretson (1991) demonstrated how the MRMLC solved Bereiter's (1963) three psychometric problems by conceptualizing the latent changes as separate dimensions within multidimensional IRT.

Unlike the Andersen model, the MRMLC directly estimates changes in the latent trait for an individual across assessment periods and this overcomes the fundamental problems associated with taking differences between either successive raw scores or successive latent score estimates. If the researcher is interested in the latent trait level of

an individual at any assessment time, then it can be expressed as composite latent trait at time t , which is the sum of the initial latent trait level and all the latent changes up to the time t . However, the MRMLC is somewhat limited. It is only appropriate for dichotomous responses and, like all Rasch models, the discrimination parameters are constrained to be equal across all items. Again, Roberts and Ma (2006) pointed out that the MRMLC is multivariate in its mathematical form, but univariate in nature, just like the Andersen model.

Generalizations of the MRMLC. The MRMLC is an additive model that is designed only for dichotomous responses. This limits the application of the model in testing practice because there are situations in which polytomous item responses are used and there are also situations in which a Rasch-type model does not fit the item responses well. In these situations, a more general IRT model maybe needed. For example, the generalized partial credit model (GPCM; Muraki, 1992) is used to model responses to test items in the National Assessment of Educational Progress (NAEP; Allen, Donoghue & Schoeps, 2001). Several extensions of MRMLC to accommodate polytomous responses and varying discrimination parameters have been proposed using the same conceptualization of change over time (Wang, Wilson & Adams, 1998; Wang & Chyi-In, 2004; Roberts and Ma, 2006). Wang, Wilson & Adams (1998) generalized the MRMLC to assess change over time when polytomous item responses are used on test administrations. In essence, they extended Embretson's MRMLC to the case where a partial credit model (Masters, 1982) would be appropriate at a given time point. Wang & Chyi-In (2004) subsequently demonstrated that the gain score estimated by this

polytomous model could be used as an estimate of effect size along with a standard error estimate.

The Wang et al. (1998) generalization of the MRMLC retains the additive property of Rasch models, which makes it inappropriate for situations in which not all items have the same discrimination parameter. Roberts and Ma (2006) further generalized the MRMLC to accommodate varying discrimination parameters across items. In their model, changes over time are conceptualized as separate dimensions and items within each assessment time are treated as unidimensional, just as in the MRMLC. The IRT model used for each assessment administration is the GPCM, and thus, the model is called GPCM for repeated measures (GPCM-RM). The category probability function of GPCM-RM is defined as:

$$P(X_{ji(t)} = x \mid \theta_{j1}^*, \dots, \theta_{jT}^*) = \frac{\exp\left(\sum_{k=0}^x \alpha_i \left[\sum_{q=1}^t \theta_{jq}^* - \beta_{ik}\right]\right)}{\sum_{w=0}^{M_i} \exp\left(\sum_{k=0}^w \alpha_i \left[\sum_{q=1}^t \theta_{jq}^* - \beta_{ik}\right]\right)} \quad (6)$$

where

$X_{ji(t)} = 0, 1, \dots, M_i$ is the observed response of person j to item i at assessment time t ,

β_{ik} is the k^{th} step location of the i^{th} item on the latent continuum,

α_i is the discrimination of the i^{th} item on the latent continuum, and

θ_{jt}^* are defined as in Equation 4.

Note that β and α_i may be repeated across forms. Model identification and estimation is complex for GPCM-RM. Roberts and Ma (2006) constrained the first step location and the discrimination parameter for one common item to identify the model and used a fully Bayesian technique to estimate model parameters and hyperparameters of the latent

variable distributions. These hyperparameters represent the means of the baseline latent variable and latent change score distributions, and the corresponding variance-covariance matrix among these latent variables. Again, the covariance matrix can be transformed easily into a correlation matrix.

Multidimensional GPCM for Repeated Measures. te Marvelde, Glas, Van Landeghem & Van Damme (2006) generalized the GPCM to the multidimensional case with T dimensions. The authors showed how such a model could be used to analyze longitudinal survey data when repeated measures are used. This multidimensional GPCM model can be expressed as:

$$P(X_{jik} = 1 | \alpha_i, b_i, \theta_j) = \frac{\exp\left(\sum_{t=1}^T \alpha_{it} k \theta_{jt} - b_{ik}\right)}{1 + \sum_{h=1}^{m_i} \exp\left(\sum_{t=1}^T \alpha_{it} h \theta_{jt} - b_{ih}\right)} \quad (7)$$

where

$k = 1, \dots, m_i$ is the response to a given item,

$h = 1, \dots, m_i$ is a counter index,

m_i is the total number of categories for the i th item,

$\alpha_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iT})$ is a T -dimensional vector of discrimination parameters,

$b_i = (b_{i1}, b_{i2}, \dots, b_{im_i})$ is a m_i -dimensional vector of location parameters (i.e., step parameters),

$\theta_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jT})$ is a T -dimensional vector of person parameters,

X_{jik} equals 1 if the j^{th} person gives a response in category k of item i ; otherwise X_{jik} equals to zero.

As indicated by te Marvelde et al. (2006), the model described by Equation 7 is a very general model that will often be constrained when measuring latent change over time. Specifically, constraints will generally be placed on discrimination parameters so that item characteristics are fixed across assessment points. With such constraints in place, the model is similar to the GPCM-RM:

$$P(X_{jik} = 1 | \alpha_i, b_i, \theta_j) = \frac{\exp\left(\alpha_i \sum_{t=1}^T k \theta_{jt} - b_{ik}\right)}{1 + \sum_{h=1}^{m_i} \exp\left(\alpha_i \sum_{t=1}^T h \theta_{jt} - b_{ih}\right)} \quad (8)$$

The model above could easily be denoted to incorporate items that are nested within a test form administered at a specific time point, t . Therefore, it differs from the GPCM-RM primarily because it estimates a composite latent trait score at each time point; specifically, it yields latent profile scores over time. In contrast, the GPCM-RM parameterizes individual differences using an initial (baseline) latent trait and subsequent latent change between successive time points. Te Marvelde et al. (2006) used a marginal maximum likelihood technique with a multivariate normal prior distribution to estimate item parameters, latent growth profiles, and the covariance matrix for the latent traits. This is different from the fully Bayesian estimation technique used by Roberts & Ma (2006) to estimate parameters of the GPCM-RM.

III. Unfolding IRT Approach to Measure Change across Repeated Assessments

Modeling Proximity-based Response Processes

The unidimensional and multidimensional IRT models discussed in the preceding sections all assumed that item responses resulted from a dominance-based response process. In this case, a cumulative (i.e., monotone) IRT model is generally consistent with the data. A cumulative IRT model suggests that higher item scores are more likely to the extent that the respondent possesses higher levels of the latent trait(s). Although cumulative IRT models are often appropriate for tests of achievement and proficiency, they may not adequately represent questionnaire data designed to measure constructs such as attitudes, satisfaction, and preference. These constructs are often measured with questionnaires constructed in the Likert (1932) and Thurstone (1927, 1928) traditions. Moreover, several researchers have argued that responses from such questionnaires are generally more consistent with the notion of proximity-based response process (Andrich, 1996; Roberts, Laughlin, Wedell, 1999; van Schuur & Kiers, 1994). In a proximity-based response process, an individual will agree with or more readily select an item to the extent that the content of the item matches the individual's own ideal level with respect to the construct of interest. This process generally results in different response patterns than those arising from a dominance-based response process, and therefore, different models are often required to adequately represent the data.

Unfolding IRT models have been developed to analyze responses resulting from a proximity-based response process. Unfolding IRT models imply that higher item scores are more likely when the individual is located close to an item on the latent

continuum as opposed to more distant locations. One of the most general unfolding IRT models is the generalized graded unfolding model (GGUM, Roberts, Donoghue & Laughlin, 2000). The GGUM is a unidimensional unfolding model that is appropriate for either binary or graded responses. If the model fits the response data, the GGUM offers the same advantages as any other parametric IRT model. These advantages include sample invariant interpretation of item parameters, item invariant interpretation of person parameters, and estimates of precision at the parameter level.

The GGUM is mathematically defined by its category probability function:

$$P[Z_i = z | \alpha_i, \delta_i, \tau_{ik}, \theta_j] = \frac{\exp(\alpha_i[z(\theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik}]) + \exp(\alpha_i[(M - z)(\theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik}])}{\sum_{w=0}^C [\exp(\alpha_i[w(\theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik}]) + \exp(\alpha_i[(M - w)(\theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik}])]} \quad (9)$$

where

$z = 0, 1, 2, \dots, C$; represents individual's response to an item,

C = the number of observed response categories minus 1,

$M = 2 * C + 1$,

θ_j is the location parameter of the j^{th} individual on the latent continuum,

δ_i is the location parameter of the i^{th} item on the latent continuum,

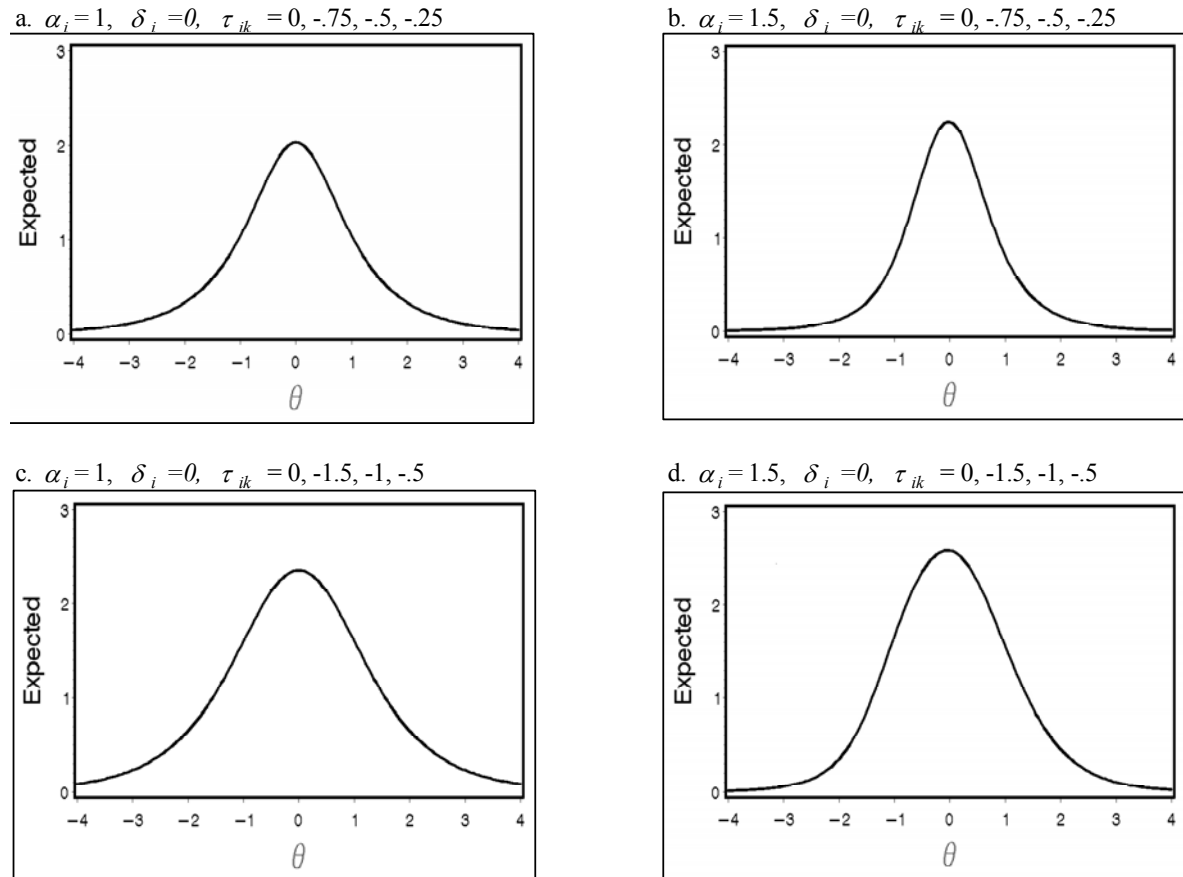
α_i is the discrimination parameter of the i^{th} item,

τ_{ik} is the k^{th} subjective response category threshold parameter for the i^{th} item.

The unidimensional GGUM allows both discrimination and threshold parameters to vary across items. The item characteristic curve (ICC) under the GGUM is always single peaked (Donoghue, 1999) unless the corresponding discrimination parameter is

equal to zero. The shape of the ICC is influenced by both discrimination and threshold parameters. Figure 1 illustrates how the expected value of an observed response to a hypothetical item with four response categories is influenced by the value of discrimination and threshold parameters, respectively, when holding values of other model parameters constant. A comparison of panels 1a to 1b and 1c to 1d shows that the maximum expected value is greater and the shape of the ICC is more peaked as α_i increases from 1 to 1.5. Additionally, the effect of increasing the distance between subjective response category thresholds can be seen by comparing panels 1a to 1c and 1b to 1d. The interthreshold distance is increased from .25 to .5 across these comparisons. Consequently, the maximum expected value increases with greater interthreshold distance, but the shape of ICC becomes less peaked. This second effect of increasing the interthreshold distance is opposite from that found when increasing the value of the discrimination parameter. Consequently, these two types of parameters have distinguishing features that may prove useful when modeling a given set of data.

Figure1: Expected Value of an Observed Response to a Hypothetical Four-Category Item as a Function of α_i and τ_{ik}



Roberts et al. (2000) used a marginal maximum likelihood (MML) technique to estimate item parameters in the GGUM along with an expected a posteriori (EAP) method to estimate person parameters. Roberts, Donoghue and Laughlin (2002) subsequently examined the accuracy of these methods and found that very accurate estimates of item parameters could be obtained with the MML technique when using 750 to 1000 respondents. Similarly, accurate estimates of person parameters were obtained using responses to 15-20, equally spaced items with 6 response categories per item. Additional simulation studies have been performed to explore the data demands under

different sample sizes, test lengths, and with different numbers of response categories (Cui, Roberts & Bao, 2004). The results suggested that when fewer response categories are used, larger sample sizes (i.e., between 1000 and 1250) would be required to obtain accurate estimates of item parameters. However, 20 items still appeared sufficient to produce reasonably accurate estimates of person parameters, regardless of the number of response categories.

Directly Assessing Change in Repeated Measures Designs Using the GGUM

In addition to assessing growth of knowledge, skills and cognitive levels, assessing the change in attitude toward an academic subject, the change in satisfaction with instruction, or the change in preference of instruction styles may be both relevant and important to educational practice. A number of researchers have suggested that these psychological characteristics follow an ideal point process (Andrich, 1996; Roberts, Laughlin, Wedell, 1999; van Schuur & Kiers, 1994), and thus, they can be modeled more appropriately with an unfolding IRT model. Given the emphasis on measuring change over repeated measures, an unfolding model that directly models changes in attitude, satisfaction and preference and the dependency of latent traits is needed. This study will recast the GGUM into a multidimensional form in order to assess change using the same logic implemented in the GPCM-RM.

One advantage of the MRMLC and the GPCM-RM for assessing change over time is that they conceptualize change itself as a separate dimension and model it directly using an appropriate parametric IRT model. This same idea can be applied to conceptualize changes in attitude, satisfaction and preference while incorporating an

appropriate parametric unfolding IRT model, such as the GGUM, to analyze responses collected from Likert or Thurstone questionnaires at each assessment administration. Indeed, this idea is used herein to develop a new IRT model. This new model is referred as the generalized graded unfolding model for repeated measures (GGUM-RM).

The GGUM-RM is defined by its category probability function:

$$P(Z_{ji(t)} = z | \theta_{j1}^*, \theta_{j2}^*, \dots, \theta_{jT}^*) = \frac{\exp(\alpha_i [z(\sum_{q=1}^t \theta_{jq}^* - \delta_i) - \sum_{k=0}^z \tau_{ik}]) + \exp(\alpha_i [(M-z)(\sum_{q=1}^t \theta_{jq}^* - \delta_i) - \sum_{k=0}^z \tau_{ik}])}{\sum_{w=0}^C [\exp(\alpha_i [w(\sum_{q=1}^t \theta_{jq}^* - \delta_i) - \sum_{k=0}^w \tau_{ik}]) + \exp(\alpha_i [(M-w)(\sum_{q=1}^t \theta_{jq}^* - \delta_i) - \sum_{k=0}^w \tau_{ik}])]} \quad (10)$$

for $z = 0, 1 \dots C$,

where

$Z_{ji(t)}$ represents a response from the j^{th} individual to the i^{th} item where that item may be administered at multiple assessment times, each of which is indexed by t ,

$q = 1, 2, \dots, T$ is a given assessment time point (i.e., q is simply a counter that indexes time points between 1 and t),

$t=1, 2, \dots, T$ is the assessment time point for the response in question,

C is the number of response categories minus 1,

M is equal to $2*C+1$,

$\theta_{j1}^* = \theta_{j1}$ is the initial (i.e., baseline) level of the latent trait for the j^{th} person,

$\theta_{j2}^* = \theta_{j2} - \theta_{j1}$ is the change in the level of the latent trait from baseline to time point 2,

.....

$\theta_{jt}^* = \theta_{jt} - \theta_{j(t-1)}$ is the change in the level of the latent trait from time $t-1$ to time point t ,

δ_i is the location parameter for item i ,

τ_{ik} is the k^{th} threshold for item i , and

α_i is the discrimination parameter for item i .

A unidimensional GGUM is presumed to hold for all items administered at each assessment time point, and the construct measured by the items is held constant across time points. Thus, the parameters of α_i , δ_i and τ_{ik} are held constant for common items on successive test administrations. This model treats the baseline level of the latent trait and successive latent changes over time as separate dimensions in a mathematical sense, but the underlying latent trait represents the same psychological construct over assessment periods. An individual's composite latent trait level at the t^{th} time point can be derived by the sum of the initial latent trait level and all the latent change scores up to the t^{th} time point:

$$\theta_{jt} = \sum_{q=1}^t \theta_{jq}^* \quad (11)$$

where

q , θ_{jt} , and θ_{jq}^* are defined as in equation 5.

This composite score can be used to generate traditional IRT quantities, such as the item characteristic curve, the test characteristic curve, the item information function, etc. Since responses are from repeated measures of the same individuals, the latent variables are dependent. This dependency among latent variables can be ascertained via direct estimates of the variance-covariance matrix corresponding to the multivariate normal distribution. (The variance-covariance matrix can be transformed into a correlation matrix if desired). The associated centroid of that distribution can also be directly estimated.

Benefits of the GGUM-RM. If model fits the data, the new GGUM-RM has all advantages associated with any parametric IRT model, and it generally provides a more adequate representation of responses from a Likert-type or Thurstone-type scale than do cumulative IRT models. In addition, the GGUM-RM provides direct estimates of the latent change over time while accounting for the correlation between latent variables. Parameterizing change as alternative latent dimensions resolves the reliability paradox mentioned earlier because the precision associated with a given latent change estimate is not dependent on the correlation between raw test scores at successive points in time (Embretson, 1991). Thus, practitioners can choose tests/questionnaires with high correlations at different assessment administrations without degrading the reliability of the latent change estimates. Also, te Marvedle et al. (2006) pointed out that taking into account of the dependency between latent variables would increase the accuracy of estimates of change.

Model Identification and Parameter Estimation in GGUM-RM. The GGUM-RM is not identifiable without further constraints. The constraint strategy implemented by Roberts and Ma (2006) in the GPCM-RM was to fix the first step location and discrimination parameter for a common item to arbitrary values (i.e., 0 and 1, respectively). This provided a unique scale origin and unit required to estimate the remaining model parameters. However, a preliminary study of the GGUM-RM showed that this method did not provide stable estimates of model parameters. We suspect that this problem was due to the greater amount of metric uncertainty encountered when identifying the scale by fixing the location and discrimination parameters for one item as compared to fixing the mean and variance of the latent trait. Additionally, Roberts,

Donoghue and Laughlin (2000) originally removed the indeterminacy of GGUM parameter estimates by fixing the mean and variance of the latent trait distribution. Thus, the behavior of GGUM parameter estimates using other constraints to remove model indeterminacies was never investigated. It could be that fixing the location and scale parameters for a single item leads to less stable estimation when there are local maxima in the likelihood function like those that have been reported with unfolding models such as the GGUM. In any event, removing the indeterminacies inherent in the GGUM-RM by restricting the mean and variance of θ_{jI}^* appears to be a feasible solution. Specifically, for this model, the mean and variance of the latent initial level are set to 0 and 1, respectively. But the mean and variance of other latent variables, and the covariances among all latent variables are free parameters to be estimated. Thus, in addition to item and person parameters, the model can also provide a direct estimate of latent group change and the covariance of latent change:

$$\mu = \left[0, \mu_{\theta_{j2}^*}, \dots, \mu_{\theta_{jT}^*} \right], \text{ and}$$

$$\Sigma_{\theta} = \begin{bmatrix} 1 & \sigma_{\theta_{12}^*} & \dots & \sigma_{\theta_{1T}^*} \\ \sigma_{\theta_{21}^*} & \sigma_{\theta_2^*}^2 & \dots & \sigma_{\theta_{2T}^*} \\ \dots & \dots & \dots & \dots \\ \sigma_{\theta_{T1}^*} & \sigma_{\theta_{T2}^*} & \dots & \sigma_{\theta_T^*}^2 \end{bmatrix}.$$

After the model is identified, model parameters and hyperparameters can be estimated using a fully Bayesian estimation technique. This technique combines information about the prior distributions of model parameters and hyperparameters along with the traditional likelihood function to produce a joint posterior probability distribution of all parameters. This distribution can be used to calculate the posterior mean of each model

parameter. In this study, a fully Bayesian solution is obtained using a Markov chain Monte Carlo (MCMC) procedure implemented in the WinBUGS computer program (Spiegelhalter, Thomas, Best, & Lunn, 2003). The MCMC procedure is an iterative sampling scheme in which values of model parameters are drawn from an approximate posterior probability distribution and repeatedly corrected to better represent the target posterior distribution. Eventually, the iterative process (i.e., Markov chain) converges to a unique stationary distribution; namely, the posterior distribution of model parameters. Multiple draws can be made from the stationary distribution and used to formulate the sampling distribution of each model parameter. Parameter estimates are derived by taking the mean of a given sampling distribution. The default sampler (*current point Metropolis sampling method*) provided by WinBUGS is used for this study.

Prior Distributions for Model Parameters and Hyperparameters. When a fully Bayesian technique is used, specification of a prior distribution is required for each model parameter. In this study, the prior distributions for the item locations were chosen to be identical $N(0,1)$ distributions, and those for item discrimination parameters were set to be identical lognormal distributions with means of 0 and variances equal to .25. The prior distributions for discrimination parameters are the same as those used in PARSCAL computer program (Muraki & Bock, 1997), but those for location parameters are a slightly less variable than those used in PARSCAL though identical in form (i.e., NIID). Considering GGUM-RM is more complex than GPCM, this change to the variance of the prior distribution seemed warranted. The prior distributions for threshold parameters were set to identical uniform distributions with ranges of $[-4,1]$. The uniform prior distribution guaranteed that estimates of threshold parameters would remain within a

reasonable range of values, though it did not favor any particular values in that range. Given the sparse information about the distribution of GGUM thresholds in the literature, this flat prior distribution seemed reasonable.

The prior distribution for latent θ_{jt}^* , was assumed to be a multivariate normal distribution with mean vector of μ and a covariance matrix $\Sigma_{(\theta)}$. As pointed out earlier, the elements of vector μ are the estimates of the mean latent level at the initial assessment point and the mean latent change between successive assessment points. The diagonal elements of matrix $\Sigma_{(\theta)}$ correspond to the variances for each latent variable and off-diagonal elements are the estimates of the linear dependency between pairs latent variables. The first element of vector μ , corresponding to the mean of the initial latent variable, was set to 0 for identification purposes, whereas the remaining means were estimated using identical $N(0,100)$ prior distributions. The large variance associated with these distributions leads to an extremely non-informative prior distribution for the corresponding hyperparameters. It is not possible to directly constrain the matrix $\Sigma_{(\theta)}$ using WinBUGS because this software does not allow users to constrain the elements of a multivariate distribution that might be used as a prior distribution for $\Sigma_{(\theta)}$ (i.e., a Wishart prior distribution). Consequently, constraining the first diagonal element of $\Sigma_{(\theta)}$ to be equal to 1 for model identification purposes is not possible for the general case.

However, it can be accomplished quite easily in the case of two time points by specifying separate prior distributions for the free elements in the $\Sigma_{(\theta)}$. Specifically, the first diagonal element of $\Sigma_{(\theta)}$ was fixed to one, and the inverse of the second diagonal element was modeled with a gamma (.5, .5) prior distribution. The single unique off-diagonal covariance was modeled as a function of the correlation between the two latent traits

times the standard deviation of the second latent trait. The correlation was, in turn, modeled using a uniform $[-1,+1]$ prior distribution. With these specifications in place, a variance-covariance matrix was formed and subsequently inverted and used as a hyperparameter for the multivariate normal distribution as required in WinBUGS. (Inversion of the covariance matrix is peculiar requirement of WinBUGS which parameterizes the multivariate normal distribution by its corresponding centroid and inverse variance-covariance matrix.) These priors are appropriate only for the situations involving two time points. When individuals are surveyed at more than two time points, then the variance-covariance matrix cannot be modeled in this fashion because there is no guarantee that the method will produce a positive definite matrix.

IV. A Parameter Recovery Simulation

The main purpose of the simulation study was to demonstrate the feasibility of calculating parameter estimates using the fully Bayesian technique implemented in WinBUGS. All other things being equal, sample size is the key factor that affects the accuracy of the item parameter estimates, while the accuracy of person parameter estimates will increase as test length increases. Furthermore, the proportion of common items shared by test forms in a repeated measures design will also affect parameter estimation across assessment periods. Thus, this study varied three independent variables: sample size, test length, and the proportion of common items shared by alternate forms.

Ten replications were simulated under each combination of levels for the three factors. This small number of replications was undertaken because WinBUGS runs very slow for complex model like GGUM-RM. In some cases, an analysis of data from a single replication required up to three days of computing time on a fast Pentium-based processor. On every replication, each survey item always had four response categories.

Simulation Design

Number of assessment points. Two assessment points were simulated in each condition for this study. One reason is that it is the simplest situation for repeated measures designs. The other reason is that the strategy used to model the variance-covariance matrix is only suitable for situations with two time points, and can not be directly generalized to repeated measures designs with more than two assessment points. As mentioned previously, this is a limitation of the WinBUGS program rather than a typical feature of Markov Chain Monte Carlo methods.

Number of response categories. GGUM-RM is appropriate for either binary or graded responses. Questionnaires constructed in Likert or Thurstone tradition usually have graded response items. Some research has suggested that, because of the ambiguity of the neutral or undecided option, an odd number of response options should be avoided when constructing questionnaires using Likert or Thurstone methods, if exploring the ambiguity is not particularly of interest (Bock & Jones, 1968; Dubois & Burns, 1975; Andrich, de jong, & Sheridan, 1997). Additionally, a previous simulation study (Cui, Roberts & Bao, 2004) suggested that somewhat larger samples were required to obtain item parameter estimates of similar accuracy as the number of response categories decreased from six categories to two categories. These two lines of research suggested that four response categories per item would be a reasonable choice in this simulation study.

Sample size. Previous research has shown that very accurate estimates of item parameters for a unidimensional GGUM model can be obtained with a marginal maximum likelihood (MML) technique when using 750 respondents along with a items that have 6 response categories (Roberts, Donoghue and Laughlin 2002). When the response categories are reduced to 4 and 3 categories, respectively, approximately 1000-1250 respondents are required to achieve similar levels of accuracy (Cui, Roberts & Bao, 2004). The model used in this simulation is more complex than the GGUM. Thus, we expected that more simulees would be needed to obtain accurate estimates of item parameters. Therefore, sample sizes of 1000 and 2000 were studied in an effort to describe the characteristics of parameter estimates within an informed sample size range.

Test length. Test length will influence the precision of person parameter estimates. Estimates will be more precise when more informative items are used. Test lengths of 10, 20 and 30 items were simulated in this study. Roberts, Donoghue and Laughlin (2002) suggested that at least 15 items were required in order to obtain accurate estimates of person parameters for a unidimensional GGUM when using items with 6 response categories. In light of their suggestion and the additional complexity of the GGUM-RM, it seemed reasonable to initially explore test lengths of 20 and 30 items with 4 response categories per item. Considering that fewer items are often used in evaluation research studies, a third condition which utilized 10 items was also included in the simulation study.

The proportion of common items shared by alternate forms. When measuring individual differences using a repeated measure design, either the same test/questionnaire or alternate forms of a test/questionnaire is/are administered at each time point. When alternate forms are used, common items are embedded across forms to maintain the same metric. In this study, the proportion of common items shared by alternate forms was either 30% or 100%. These two conditions corresponded to situations in which an alternate form with common items or an identical form is used over time, respectively. When alternate forms are used in testing practice, the number of common items is an important issue that should be considered. Larger numbers of common items led to less random linking error. Some researchers have suggested that a common item set should be at least 20% of the length of the total test (Angoff, 1971; Kolen & Brennan 2004). As the GGUM-RM is a complex multidimensional IRT model, more items might be required to

estimate stable model parameters. Thus, 30% of common items shared across alternate forms were chosen for this simulation study.

Response Generation Procedure. The true item parameters used on a given replication were re-sampled from a list of 47 item parameter estimates derived from an analysis of an abortion attitude questionnaire reported by Roberts, Lin and Laughlin (2001). The items on this list were originally associated with 6-category response formats. However, the data from their study were recoded into four response categories and item parameters were estimated from the recoded data based on a new maximum marginal a posteriori (MMAP) estimation program that is currently under study (Roberts & Thomson, 2007). The MMAP parameter estimates associated with these items served as the generating (i.e., true) item parameters on a given replication. A general principle in Thurstone (1928) attitude scale construction is that items should be explicitly chosen to represent the affective continuum in an approximately uniform pattern. Therefore, most, if not all, of the respondents will be close to some item on the scale. In an effort to obtain a relatively uniform distribution across a reasonable portion of the attitude continuum, only 35 items with location estimates between [-2.5 2.5] on the latent continuum were chosen to serve in the item pool. The interval was then divided into five segments: [-2.5 - 1.5], (-1.5 -0.5], (-0.5 0.5], (0.5 1.5] and (1.5 2.5]. An equal number of items were randomly drawn with replacement from each segment.

The true person parameters were randomly sampled from a multivariate normal distribution with $\mu = [0, .5]$ which indicated that the true average change over the two assessment points was .5 units on the latent continuum. The variances for composite θ at each assessment point were set to 1 and 1.5625 respectively. As mentioned earlier, the

variance of the latent variable at the initial assessment time was set to 1 in order to identify the model, whereas the larger variance for the composite θ represented a characteristic that is often seen in practice (Embretson, 1991). The correlation between the composite θ at the two time points was randomly drawn from a uniformly distributed closed interval of [.36 .64] on each replication. This interval represented low to moderate correlations, and it was characteristic of a range of values reported in the literature (Embretson, 1991). Consequently, the true value of the correlation between θ_{j1}^* and θ_{j2}^* on a given replication was equal to:

$$r_{\theta_{j1}^*, \theta_{j2}^*} = \text{cov}(\theta_1, \theta_2 - \theta_1) / [\text{std}(\theta_1) \text{std}(\theta_2 - \theta_1)]. \quad (12)$$

One can then use expectation algebra to yield the following identities:

$$\text{cov}(\theta_1, (\theta_2 - \theta_1)) = r_{\theta_1, \theta_2} \text{Std}(\theta_1) \text{std}(\theta_2) - \text{var}(\theta_1) \quad (13)$$

and

$$\text{std}(\theta_2 - \theta_1) = \sqrt{\text{var}(\theta_1) + \text{var}(\theta_2) - 2r_{\theta_1, \theta_2} \text{std}(\theta_1) \text{std}(\theta_2)}. \quad (14)$$

Recall that $\text{std}(\theta_1)$ was constrained to be 1 to achieve identifiability of model parameters, $\text{std}(\theta_2)$ was set equal to 1.25, and the correlation between θ_1 and θ_2 was sampled from a uniform distribution on the interval [.36 .64]. Therefore, the true correlation between θ_{j1}^* and θ_{j2}^* on a given replication was equal to:

$$r_{\theta_{j1}^*, \theta_{j2}^*} = \frac{r_{\theta_1, \theta_2} (1.25) - 1}{\sqrt{1 + 1.25^2 - 2r_{\theta_1, \theta_2} (1.25)}} \quad (15)$$

The observed response of each individual was generated with a unidimensional GGUM model for each time point using the simulated true item parameters and the true composite θ at that point. After an observed response to each item was generated for all subjects, the data were used to estimate GGUM-RM parameters. The process of

generating data and subsequently estimating parameters was replicated 10 times in each condition defined by sample size, test length, and the proportion of common items.

Model parameter estimation. Fully Bayesian estimates were obtained by implementing an MCMC algorithm with the WinBUGs computer program. Prior distributions for each model parameter were specified as described earlier in this report. An adaptive MCMC algorithm, specifically the Metropolis sampling algorithm (Spiegelhalter, Thomas, Best, & Lunn, 2003; Gelman, Carlin, Stern, & Rubin, 2003), was used to produce a series that converged to the joint posterior distribution of model parameters. The default number of adaptive iterations that was built within WinBUGS was 4000 for Metropolis sampling. However, the GGUM-RM is a complex model and may need more burn-ins to converge. Therefore, following the procedures described by Roberts and Ma (2006), 9000 burn-in iterations were conducted. A preliminary analysis showed that similar estimates of model parameters and hyperparameters could be obtained when running the model with different initial values for model parameters. Time series plots for two chains (Appendix A) also showed that the series stabilized long before 9000 iterations. This suggests that the algorithm can converge to the joint posterior distribution within 9000 adaptive iterations. An additional 1000 iterations were obtained following the burn-in iterations, and these 1000 iterations were used to calculate *expected a posteriori* estimates of model parameters.

Data Analysis and Evaluation of Results

In all, there were 12 different combinations of simulated assessment conditions (two sample sizes x three test lengths x two common item levels). For each combination,

10 replications were generated. After the estimates of model parameters were obtained, the accuracy of the estimates was assessed for each type of model parameter estimated in a given replication.

Measures of accuracy. Four measures of estimation accuracy were used to evaluate parameter recovery. All of the four measures have been used in previous recovery studies of IRT model parameters (Andrich 1988; Andrich & Luo, 1993; Kim, Cohen etc., 1994; Roberts, & Laughlin, 1996; Roberts, Donoghue & Laughlin, 2000). The degree of the accuracy of model estimation depends on three types of discrepancies between estimated and true parameter distributions: the difference of the mean and the variance between the estimated and true parameter distributions, and the covariance between the estimated and the true parameter distributions. Three measures were used to evaluate each individual discrepancy separately, and a fourth measure reflected all three types of discrepancies.

Root mean squared deviation (RMSD). The RMSD was calculated as follows:

$$RMSD(H) = \sqrt{\sum_{i=1}^I (H_i - \hat{H}_i)^2 / I} \quad (16)$$

where

H_i is the true value of model parameters for a particular parameter type (i.e., α , δ , τ , μ , σ , and θ) in a given replication,

\hat{H}_i is the corresponding estimate of the model parameters in that replication,

I is the number of parameters of a particular type within a given replication.

The RMSD index is sensitive to three types of discrepancies between estimated and true parameters, and it will increase as either one of the three discrepancies becomes larger.

Specifically, Roberts and Laughlin (1996) illustrated that the RMSD could be decomposed as:

$$RMSD = \sqrt{S^2_H + S^2_{\hat{H}} - 2S_{H\hat{H}} + (\bar{X}_{\hat{H}} - X_H)^2} \quad (17)$$

where

S^2_H is the sample variance of the true model parameters for a particular parameter type (i.e., α , δ , τ , μ , σ , and θ) in a given replication,

$S^2_{\hat{H}}$ is the sample variance of the estimate of the model parameters in that replication,

$S_{H\hat{H}}$ is the sample covariance between H and \hat{H} ,

\bar{X}_H is the average of the H s,

$\bar{X}_{\hat{H}}$ is the average of the \hat{H} s.

Since the RMSD reflects the entire deviation in model estimation, it is used as the primary accuracy measure for this simulation study.

Absolute Bias. Absolute bias is a measure of the average absolute difference of estimated and true parameter distributions and it was calculated for each replication as follows:

$$Absolute \quad Bias \quad (H) = \frac{\sum_{i=1}^I |H_i - \hat{H}_i|}{I} \quad (18)$$

where

H_i , \hat{H}_i , and I are defined as in equation 16. The value of absolute bias will increase as the absolute difference between each pair of estimated and true parameter increases.

Variance ratio (VR). The difference of the variances between the estimated and true parameter distributions was evaluated by variance ratio and it was calculated for each replication as follows:

$$VR = s^2_{\hat{H}_i} / s^2_{H_i} \quad (19)$$

where

$s^2_{H_i}$ is the variance of the true model parameters for a particular type of parameters (i.e., α , δ , τ and θ) in a given replication,

$s^2_{\hat{H}_i}$ is the corresponding variance of the estimated model parameters in that replication.

The value of variance ratio will be greater or smaller than 1 as the difference of the variances of the two distributions becomes larger, depending which distribution has larger variation.

Pearson correlation (r). A Pearson product-moment correlation between estimated and true parameters was computed across model parameters (i.e., α , δ , τ and θ) within each replication to evaluate the linear association between the two distributions. Strong linearity between the estimated and true parameters will result in higher value of r .

Effect of key factors. This simulation represents a 2x3x2 fixed-effects factorial design. A univariate ANOVA procedure was used to analyze the bias, variance ratio, correlation and $RMSD$ associated with a given type of parameter. Sample size, test length, and the proportion of common items shared across alternate forms served as three between-replication factors in the analysis. There were 10 replications in each cell of the corresponding factorial design. Given that the $RMSD$ index for 8 parameters, and absolute bias, VR , and r indices for 5 parameters were analyzed in this study, the Type I error rate was set to $\alpha = .00625$ for $RMSD$, and $\alpha = .01$ for absolute bias, VR , and r in

ANOVA to control for the fact that the same ANOVA model was run 8 times for RMSD, 5 times for absolute bias, VR , and r . The proportion of the total sums of squares associated with each ANOVA effect (η^2) was also calculated. Interpretations were limited to those effects that are both statistically significant and had an associated $\eta^2 > .05$, which indicated that an ANOVA effect accounted for at least 5% of the total sum of squares in the dependent variable. The latter criterion limited interpretations to those effects that were of sufficient magnitude to be meaningful in practice.

V. A Real Data Example

An analysis of graded *agree-disagree* responses to abortion attitude items was performed using data from 750 University of South Carolina (USC) and 428 Georgia Institute of Technology (GT) undergraduate students. Students from USC responded to each of 50 items using one of six response categories: *strongly disagree*, *disagree*, *slightly disagree*, *slightly agree*, *agree*, and *strongly agree* at one assessment time. For each respondent, items were presented in a random order, and they were not allowed to skip any item. Students from GT responded to 40 questionnaire items. Of these 40 items, 19 of them were from the questionnaire administered to USC students. The response categories were identical in both the USC and GT samples. Items were also presented to GT students in random order, but students were allowed to skip any item they did not want to answer. GT students were also asked to respond to the same attitude questionnaire at two separate assessment times that were approximately three weeks apart.

Previous research suggested that the 19 items that appeared on both questionnaires were unidimensional and represented all portions of attitude continuum to approximately the same degree (Roberts, Donoghue, & Laughlin, 2000). These items were also fit reasonably well by the GGUM. Thus, the responses to these 19 items were analyzed in this study.

Because USC students only responded to the 19 items at one assessment time without skipping any single item, all the responses for the second assessment time were treated as missing for these students (and coded as “NA” in WinBUGS). GT students were asked to respond to all items at two assessment times, but only 113 students

returned for the second assessment. Also, there were missing data within an individual's response vector at a given assessment time because GT students were allowed to skip items. In this analysis, missing responses within each assessment time and missing responses at the second assessment time were all coded as "NA" in WinBUGS.

Analysis of the 38 responses from 1178 students to the repeated assessments (19 responses at each of 2 assessment times) was performed using the same prior distributions for item parameters, person parameters and hyperparameters as those used in simulation study. Also, because three weeks is an extremely short period to expect much naturalistic change in attitudes to abortion, no significant mean change was expected in this analysis.

VI. Results

Accuracy of Parameter Recovery

A univariate ANOVA was performed for four measures associated with each given parameter. The 2 levels of form, 3 levels of test length, and 2 levels of sample size served as the three between-subject factors, and there were 10 replications within each cell. Only those effects that met both of the two criteria were interpreted as having impact on the accuracy of model estimation. Recall that these two criteria were $p \leq .00625$ for RMSD or $p \leq .01$ for absolute bias, variance ratio and correlation along with $\eta^2 \geq .05$. The p -values and η^2 values associated with each ANOVA effect are shown in Tables 1-4. Entries given in bold correspond to the effects that meet both interpretation criteria.

Accuracy of $\hat{\alpha}_i$. ANOVA results showed statistically significant main effects of form and of sample size on the RMSD, absolute bias, and correlation for $\hat{\alpha}_i$. None of the three key variables or their interactions showed any statistically significant influence on the variance ratio for $\hat{\alpha}_i$. In addition, the interaction between form and sample size had significant influence on the correlation between $\hat{\alpha}_i$ and true α_i . Figures 2a-2e show the mean accuracy measures of RMSD, absolute bias, and correlation associated with the $\hat{\alpha}_i$ as a function of form and sample size.

As shown in Figure 2a-2b, the average difference in RMSDs of $\hat{\alpha}_i$ was primarily influenced by form and sample size. Specifically, the average values of the RMSD for $\hat{\alpha}_i$ increased when alternate forms, instead of same form, were used across time points. This was presumably due to the fact that there were more responses per item with the same form was used, and consequently, the RMSD was lower in that condition. As expected,

the RMSD for $\hat{\alpha}_i$ also decreased with sample size. Again, this was due to having more responses for each item. Although these differences emerged using the predefined interpretation criteria, the average values of RMSD for $\hat{\alpha}_i$ were reasonably small in every form and sample size condition.

Table 1: ANOVA Effect for the Analysis of the RMSD of Estimates of Item Parameters and Person Parameters

Item Parameter	α_i		δ_i		τ_{ik}	
	<i>p-value</i>	η^2	<i>p-value</i>	H^2	<i>p-value</i>	η^2
Form	0.0001	0.1381	0.0001	0.1090	0.0001	0.1451
Test Length	0.0023	0.0456	0.2363	0.0178	0.2529	0.0155
Form x Test Length	0.7408	0.0021	0.0271	0.0454	0.0433	0.0360
Sample Size	0.0001	0.3937	0.0089	0.0432	0.0001	0.0885
Form x Sample Size	0.0111	0.0236	0.5056	0.0027	0.6223	0.0014
Test Length x Sample Size	0.3394	0.0077	0.0007	0.0937	0.0019	0.0738
Form x Test Length x Sample Size	0.3781	0.0069	0.0787	0.0317	0.0360	0.0382
Person Parameter	θ_1^*		θ_2^*			
	<i>p-value</i>	η^2	<i>p-value</i>	η^2		
Form	0.9583	0.0000	0.0001	0.0494		
Test Length	0.0001	0.9483	0.0001	0.6245		
Form x Test Length	0.0082	0.0042	0.0001	0.0508		
Sample Size	0.6834	0.0001	0.6536	0.0005		
Form x Sample Size	0.5512	0.0001	0.4300	0.0016		
Test Length x Sample Size	0.0800	0.0022	0.4848	0.0036		
Form x Test Length x Sample Size	0.7473	0.0002	0.6848	0.0019		
Hyperparameter	$\mu_{\theta^*_2}$		$\sigma^2_{\theta^*_2}$		$\sigma_{\theta^*_1\theta^*_2}$	
	<i>p-value</i>	η^2	<i>p-value</i>	η^2	<i>p-value</i>	η^2
Form	0.2175	0.0120	0.8220	0.0004	0.8633	0.0002
Test Length	0.5016	0.0108	0.3843	0.0140	0.6225	0.0080
Form x Test Length	0.1257	0.0330	0.1329	0.0299	0.9100	0.0016
Sample Size	0.0010	0.0887	0.0202	0.0405	0.0116	0.0550
Form x Sample Size	0.6446	0.0017	0.1510	0.0152	0.1538	0.0172
Test Length x Sample Size	0.7288	0.0050	0.0034	0.0871	0.4252	0.0144
Form x Test Length x Sample Size	0.7120	0.0053	0.1670	0.0265	0.9039	0.0017

Note: The bold numbers denoted the effect that met both interpretation criteria.

Table 2: ANOVA Effect for the Analysis of the Absolute Bias of Parameter Estimates

Item Parameter	α_i		δ_i		τ_{ik}	
	<i>p-value</i>	η^2	<i>p-value</i>	η^2	<i>p-value</i>	η^2
Form	<.0001	0.1200	<.0001	0.1085	<.0001	0.1690
Test Length	0.0015	0.0452	0.0346	0.0398	0.0624	0.0288
Form x Test Length	0.7626	0.0018	0.0612	0.0328	0.1643	0.0186
Sample Size	<.0001	0.4390	<.0001	0.1111	<.0001	0.1771
Form x Sample Size	0.0132	0.0208	0.7891	0.0004	0.8592	0.0002
Test Length x Sample Size	0.0989	0.0155	0.0021	0.0750	0.0161	0.0435
Form x Test Length x Sample Size	0.6223	0.0031	0.3098	0.0136	0.2142	0.0158
Person Parameter	θ_1^*		θ_2^*			
	<i>p-value</i>	η^2	<i>p-value</i>	η^2		
Form	0.8216	0.0000	0.2533	0.0005		
Test Length	<.0001	0.9413	<.0001	0.9555		
Form x Test Length	0.0096	0.0046	0.2940	0.0009		
Sample Size	0.7041	0.0001	0.1995	0.0006		
Form x Sample Size	0.7687	0.0000	0.6423	0.0001		
Test Length x Sample Size	0.0809	0.0024	0.2233	0.0012		
Form x Test Length x Sample Size	0.5377	0.0006	0.8590	0.0001		

Note: The bold numbers denoted the effect that met both interpretation criteria.

Table 3: ANOVA Effect for the Analysis of the Variance Ratio (*VR*) of Estimates of Item Parameters and Person Parameters

Item Parameter	α_i		δ_i		τ_{ik}	
	<i>p-value</i>	η^2	<i>p-value</i>	η^2	<i>p-value</i>	η^2
Form	0.5385	0.0033	0.7479	0.0009	0.0265	0.0394
Test Length	0.1682	0.0311	0.8437	0.0028	0.4219	0.0136
Form x Test Length	0.7918	0.0040	0.0631	0.0468	0.2429	0.0223
Sample Size	0.8392	0.0004	0.9871	0.0000	0.7579	0.0007
Form x Sample Size	0.0865	0.0256	0.1583	0.0167	0.5789	0.0024
Test Length x Sample Size	0.7174	0.0057	0.4561	0.0131	0.0646	0.0438
Form x Test Length x Sample Size	0.7608	0.0047	0.1960	0.0273	0.1000	0.0366
Person Parameter	θ_1^*		θ_2^*			
	<i>p-value</i>	η^2	<i>p-value</i>	η^2		
Form	0.4850	0.0028	0.0133	0.0293		
Test Length	<.0001	0.3458	<.0001	0.4480		
Form x Test Length	0.4727	0.0085	0.5380	0.0058		
Sample Size	0.8438	0.0002	0.9442	0.0000		
Form x Sample Size	0.1330	0.0129	0.0964	0.0130		
Test Length x Sample Size	0.3567	0.0117	0.7860	0.0022		
Form x Test Length x Sample Size	0.4135	0.0100	0.8179	0.0019		

Note: The bold numbers denoted the effect that met both interpretation criteria.

Table 4: ANOVA Effect for the Analysis of the Correlation (r) of Estimates of Item Parameters and Person Parameters

Item Parameter	α_i		δ_i		τ_{ik}	
	<i>p-value</i>	η^2	<i>p-value</i>	η^2	<i>p-value</i>	η^2
Form	<.0001	0.2702	<.0001	0.1439	<.0001	0.2054
Test Length	0.0713	0.0139	0.9537	0.0006	0.5920	0.0057
Form x Test Length	0.4524	0.0041	0.0832	0.0311	0.0137	0.0484
Sample Size	<.0001	0.3618	0.1659	0.0119	0.0130	0.0346
Form x Sample Size	<.0001	0.0705	0.7270	0.0008	0.9319	0.0000
Test Length x Sample Size	0.8832	0.0006	0.0006	0.0961	0.0011	0.0791
Form x Test Length x Sample Size	0.8834	0.0006	0.0114	0.0569	0.0253	0.0413
Person Parameter	θ_1^*		θ_2^*			
	<i>p-value</i>	η^2	<i>p-value</i>	η^2		
Form	0.7087	0.0001	0.0012	0.0357		
Test Length	<.0001	0.9306	<.0001	0.5496		
Form x Test Length	0.1853	0.0020	0.0005	0.0524		
Sample Size	0.1360	0.0013	0.5928	0.0009		
Form x Sample Size	0.9689	0.0000	0.5207	0.0013		
Test Length x Sample Size	0.0551	0.0034	0.2099	0.0102		
Form x Test Length x Sample Size	0.5887	0.0006	0.9412	0.0004		

Note: The bold numbers denoted the effect that met both interpretation criteria.

As shown in Figure 2c-2d, the average values of the absolute bias of $\hat{\alpha}_i$ were a function of both form and sample size main effects. The mean value of absolute bias for $\hat{\alpha}_i$ increased when alternate forms were administered across assessment times instead of same form. Mean absolute bias of $\hat{\alpha}_i$ decreased as the sample size increased. However, the average values of absolute bias were small in these conditions.

The average correlations between $\hat{\alpha}_i$ and α_i were greater than .97 in every condition, which indicated that the $\hat{\alpha}_i$ was a strong linear function of the true α_i in each condition. The small variation among these average correlations was influenced by form, sample size, and the interaction of these two factors (Figure 2e). The correlation became weaker when alternate forms, instead of same form, were administered across assessment times. As the sample size increased, $\hat{\alpha}_i$ s showed stronger linear correlation with α_i s. Additionally, though average correlations increased as sample size increased in both

same form and alternate form conditions, larger increases appeared in alternate form condition.

Accuracy of $\hat{\delta}_i$. The average RMSD $\hat{\delta}_i$ was primarily a function of the main effect of test form, and this relationship is shown in Figure 3a. Smaller values of average $\hat{\delta}_i$ RMSD were observed in the same form condition than in the alternate form condition. Additionally, the RMSD for $\hat{\delta}_i$ was also a function of the interaction between sample size and test length (Figure 4a). Larger sample size led to a small reduction in RMSD for $\hat{\delta}_i$, but only for tests of 20 or 30 items. With the smaller 10-item test, the RMSD was similar regardless of sample size. Results of *t*-test showed that average RMSDs significantly decreased as the sample size increased in 20 or 30 items condition, but not in the 10 items condition (*t*-test: $p > .05$ for 10 items condition, $p < .001$ for both 20 items and 30 items conditions). Thus, it appears that test length was a limiting factor that mitigated the effect of sample size. Test length generally affects the precision of individual latent traits estimates, and this, in turn, seems to have moderated the sample size effect.

The absolute bias of $\hat{\delta}_i$ was influenced by sample size, form, and the interaction of test length and sample size (Figure 3b, 3c, and Figure 4b). Larger average values of the absolute bias of $\hat{\delta}_i$ were observed in the alternate form condition. The mean values of the absolute bias of $\hat{\delta}_i$ decreased as the sample size increased, also, the impact of sample size showed different patterns at different levels of test length. *t*-test results indicated that the average values of the absolute bias of $\hat{\delta}_i$ significantly decreased as sample size increased from 1000 to 2000 when 20 or 30 items were administered (*t*-test: $p < .001$ for

20 items condition and $p < .0001$ for 30 items condition), but not when 10 items were used (t -test: $p > .05$).

The average correlations between $\hat{\delta}_i$ and true δ_i were near to 1 (all greater than .99) in every condition, which indicated that the $\hat{\delta}_i$ and the δ_i were highly collinear. The small differences in mean correlations were influenced by form, and the interaction of test length and sample size. As shown in Figure 3c, the correlation was stronger in the same form condition than in the alternate form condition. The sample size showed statistically significant impact on the correlations for tests of 20 and 30 items, but not for smaller tests of 10 items (Figure 4c; t -test results suggested $p > .05$ for the 10 item condition and $p < .001$ for both the 20 and 30 item conditions).

Accuracy of $\hat{\tau}_i$. The RMSD for $\hat{\tau}_i$ was influenced by form, sample size and the interaction of test length and sample size. The pattern of these effects mimicked those seen with the RMSD for $\hat{\delta}_i$. As shown in Figure 5a and 5b, the RMSD had smaller average value when the same form, instead of the alternate forms, was used across assessment times. Also, the average RMSD decreased as the sample size increased. Furthermore, as the sample size increased, the average RMSDs significantly decreased in both the 20 and 30 item conditions, but not in the 10-item condition (Figure 6a; the t -test yielded $p > .05$ for the 10 item condition and $p < .001$ for the 20 item and 30 item conditions).

The average values of the absolute bias of $\hat{\tau}_i$ s were influenced by both form and sample size (Figures 5c-5d). When the same form was administered across time points,

the average values of the absolute bias of $\hat{\tau}_i$ were smaller. Also, the mean values of the absolute bias of $\hat{\tau}_i$ decreased as sample size increased.

The average correlations between $\hat{\tau}_i$ and true τ_i were greater than .97 in every condition, which suggested that the $\hat{\tau}_i$ is quite linearly related to τ_i . The small differences in average correlations between $\hat{\tau}_i$ and true τ_i were mainly influenced by form, and the interaction of test length and sample size. As shown in Figure 5c, the average correlation was higher when the same form, instead of the alternate forms, was administered across assessment times. The main effect of the sample size did not exhibit a significant main effect on the correlation, but it did interact with test length (Figure 6b). Specifically, the mean values of correlations increased for larger samples when either 20 or 30 items were used, but there was no statistically significant difference between sample size conditions when only 10 items were used. (The corresponding *t*-test results revealed that $p > .05$ for the 10 item condition, and $p < .001$ for both the 20 item and 30 item conditions).

*Accuracy of $\hat{\theta}^*_{j1}$ and $\hat{\theta}^*_{j2}$.* All four accuracy measures, RMSD, absolute bias, variance ratio, and correlation, for $\hat{\theta}^*_{j1}$ and $\hat{\theta}^*_{j2}$ were primarily a function of test length. As shown in Figures 7a, 7b, and 7d, the mean values of RMSD and absolute bias of $\hat{\theta}^*_{j1}$ and $\hat{\theta}^*_{j2}$ both decreased as the number of items increased, while the mean correlations of $\hat{\theta}^*_{j1}$ and $\hat{\theta}^*_{j2}$ increased as the number of items increased. *post hoc* test (Tukey HSD) showed that the mean values of variance ratio decreased when test length increased from 10 items to 20 items, but with longer test of 30 items, had similar value of variance ratio as 20 item test (Figure 7c).

The RMSD and correlation for $\hat{\theta}^*_{j2}$ were also both a function of a form by test length interaction. Statistically smaller average RMSD was observed when alternate forms with 20 items were used (Figure 8a), but superiority of alternative forms disappeared with there were 10 or 30 items on each test. The same pattern was observed for the correlation of $\hat{\theta}^*_{j2}$ and true θ^*_{j2} (Figure 8b). One might speculate that the additional unique test items associated with alternate forms increased the accuracy of $\hat{\theta}^*_{j2}$. This result is not altogether surprising in an unfolding model where repeated administration of a parallel item does not produce a unique global maximum in the likelihood function associated with theta. However, the reversal of this effect with short or long test makes the finding more difficult to understand. Perhaps these cases provide two few or more than enough unique items regardless of the type of test form that is implemented.

The average RMSD, absolute bias, and variance ratio of $\hat{\theta}^*_{j1}$ (the estimate of individual's initial level) were all lower than the corresponding values for $\hat{\theta}^*_{j2}$ (the estimate of individual's change over time) in every condition (Figure 7a, 7b, and 7d), and the reverse was seen with the average correlation. This is presumable due to the fact that only those item responses from the second assessment point provided information about the estimate of $\hat{\theta}^*_{j2}$. In contrast, all item responses from both assessment points provided information about the estimate of $\hat{\theta}^*_{j1}$.

Accuracy of hyperparameter estimates ($\hat{\mu}_{\theta^*_2}$, $\hat{\sigma}_{\theta^*_1\theta^*_2}$, and $\hat{\sigma}^2_{\theta^*_2}$). The obtained $\hat{\mu}_{\theta^*_2}$, $\hat{\sigma}_{\theta^*_1\theta^*_2}$, and $\hat{\sigma}^2_{\theta^*_2}$ correspond to the direct estimate of group change, the covariance between latent variables, and the variance of the latent change score distribution. Since

only one set of estimated hyperparameters was obtained for each replication, only one accuracy measure, RMSD was calculated and evaluated for these estimates.

The ANOVA results indicated that the mean values of the RMSD for $\hat{\mu}_{\theta_2^*}$ were significantly influenced by sample size. As shown in Figure 9, the mean values of the RMSD decreased as the sample size increased. The values of the RMSD were .026 and .015 in sample size of 1000 and 2000 conditions, respectively, which were very small in either case.

The mean values of the RMSD for $\hat{\sigma}_{\theta_2^*}^2$ were significantly influenced by the interaction of test length and sample size (Table 1). The average RMSD was statistically smaller in the larger sample size condition, but this difference only held in the case of a short test consisting of 10 items, not shown in relative longer test of 20 and 30 items (Figure 10; *t*-test results yielded that $p < .0001$ for 10 item condition, $p > .05$ for both 20 and 30 item conditions).

Lastly, none of the design effects met the criteria for interpretation when considering the RMSD for $\hat{\sigma}_{\theta_1\theta_2^*}$. However, it is noteworthy that the effect of sample size was in the same direction as that for the RMSD of $\hat{\mu}_{\theta_2^*}$ and $\hat{\sigma}_{\theta_2^*}^2$. Specifically, the average values of the RMSD of $\hat{\sigma}_{\theta_1\theta_2^*}$ decreased as the sample size increased (0.076 for sample size of 1000, and 0.057 for sample size of 2000, respectively).

Figure 2: Mean Accuracy Measures for Alpha Estimates for Same Form and Alternate Form

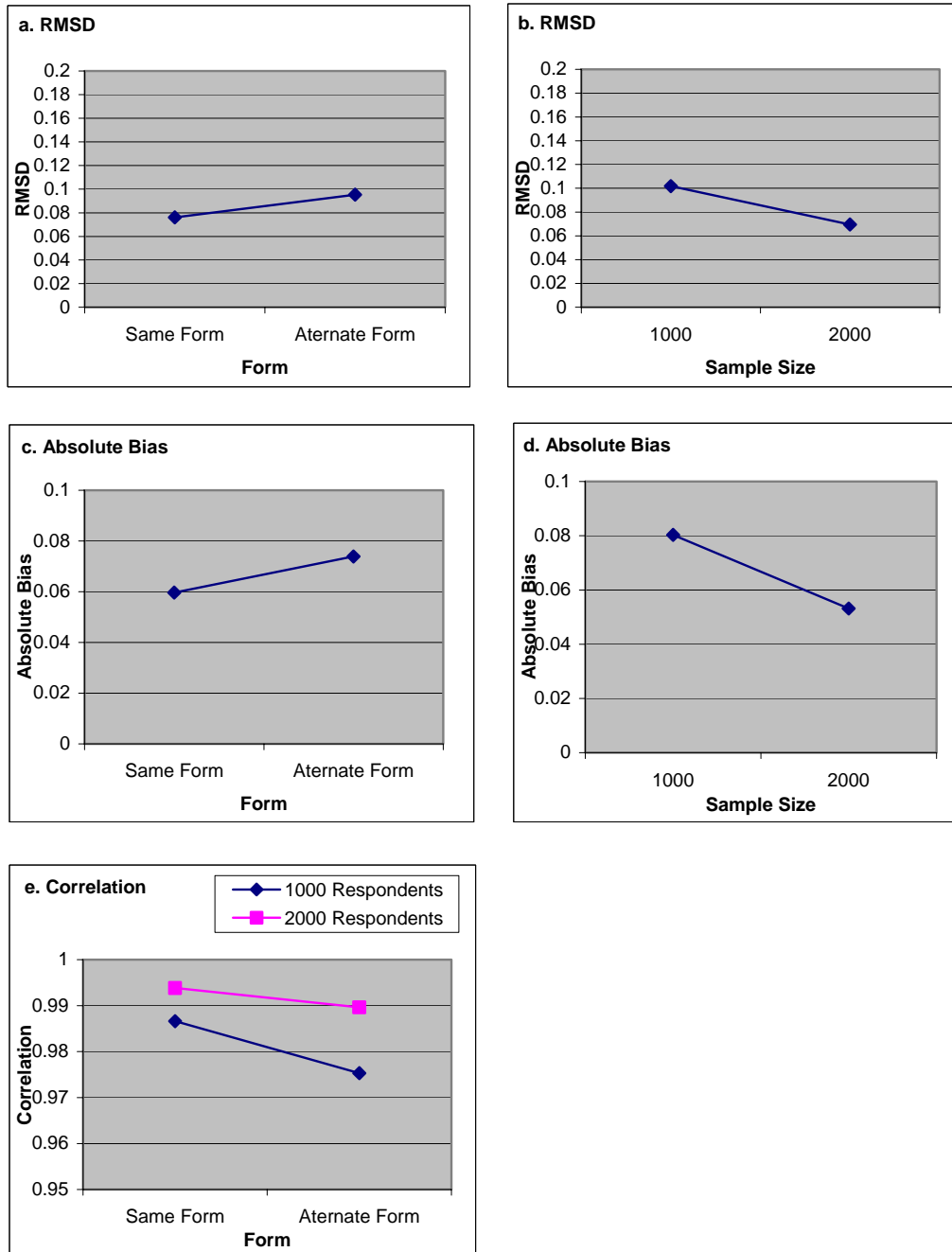


Figure 3: Mean Accuracy Measures for Delta Estimates for Same Form and Alternate Form

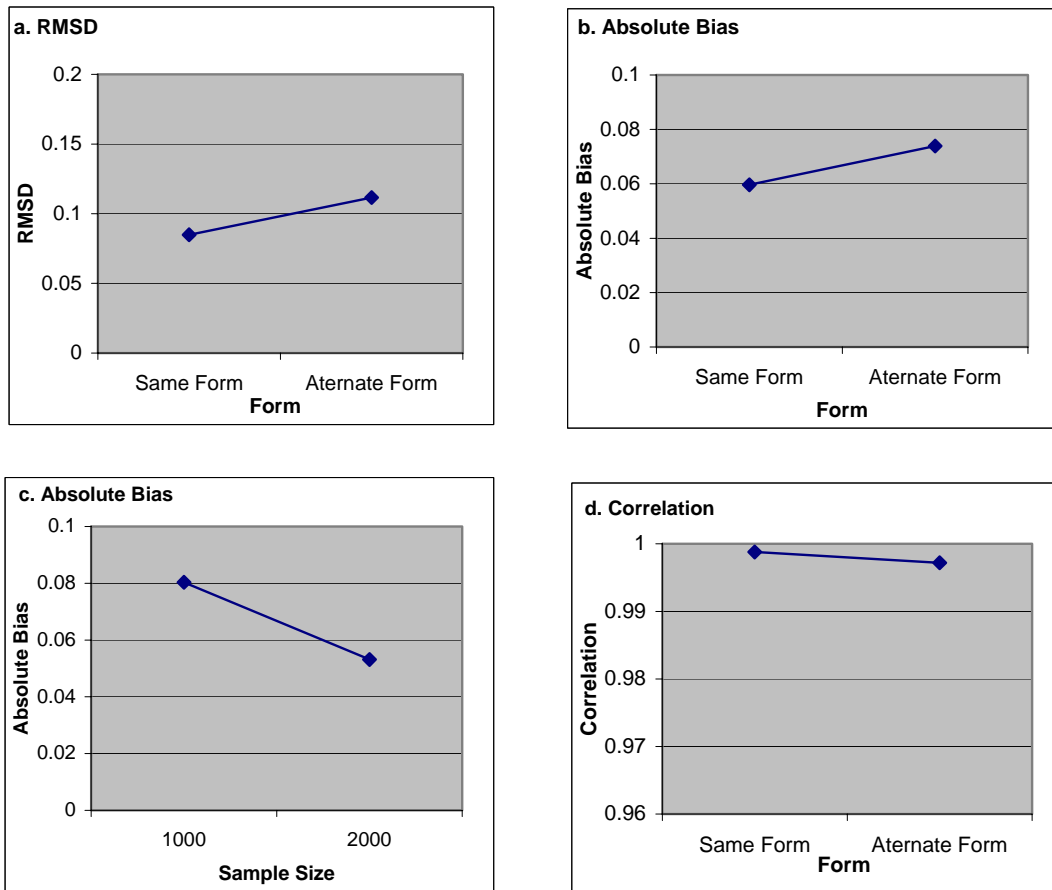


Figure 4: Mean Accuracy Measures of Delta Estimates for 10, 20 and 30 Items

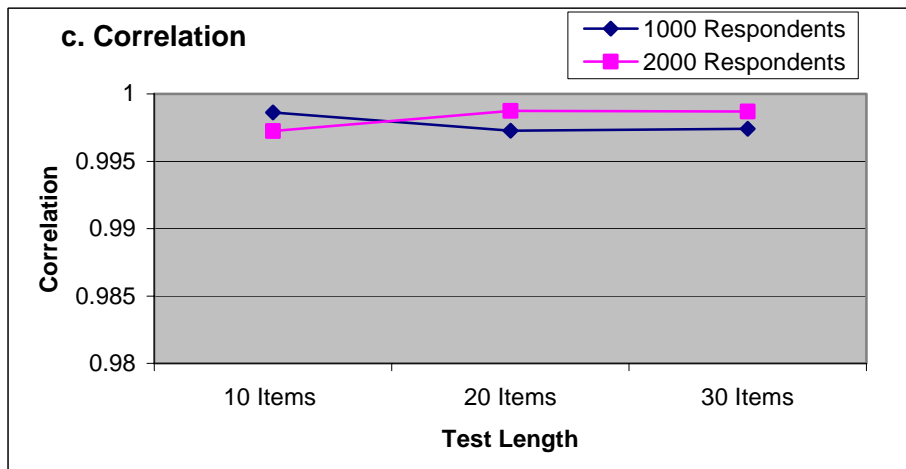
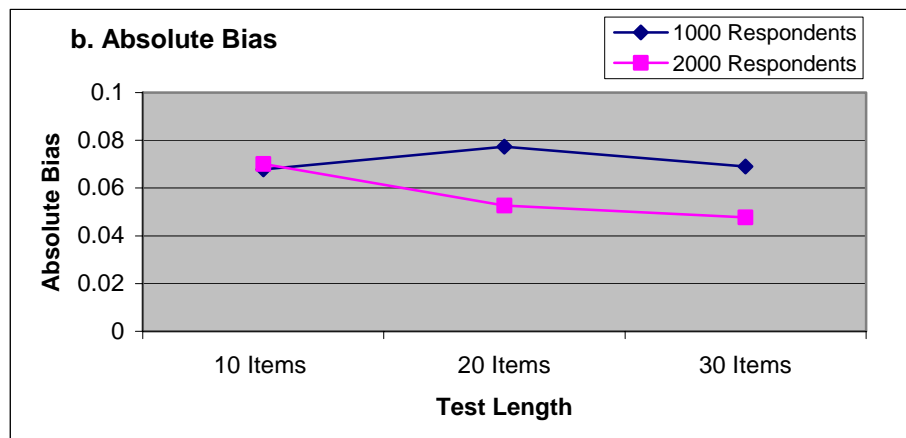
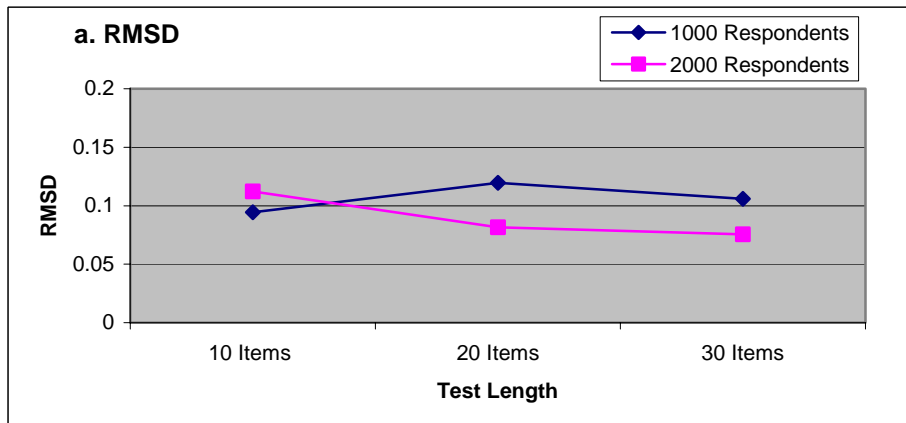


Figure 5: Mean Accuracy Measures for Tau Estimates for Same Form and Alternate Form

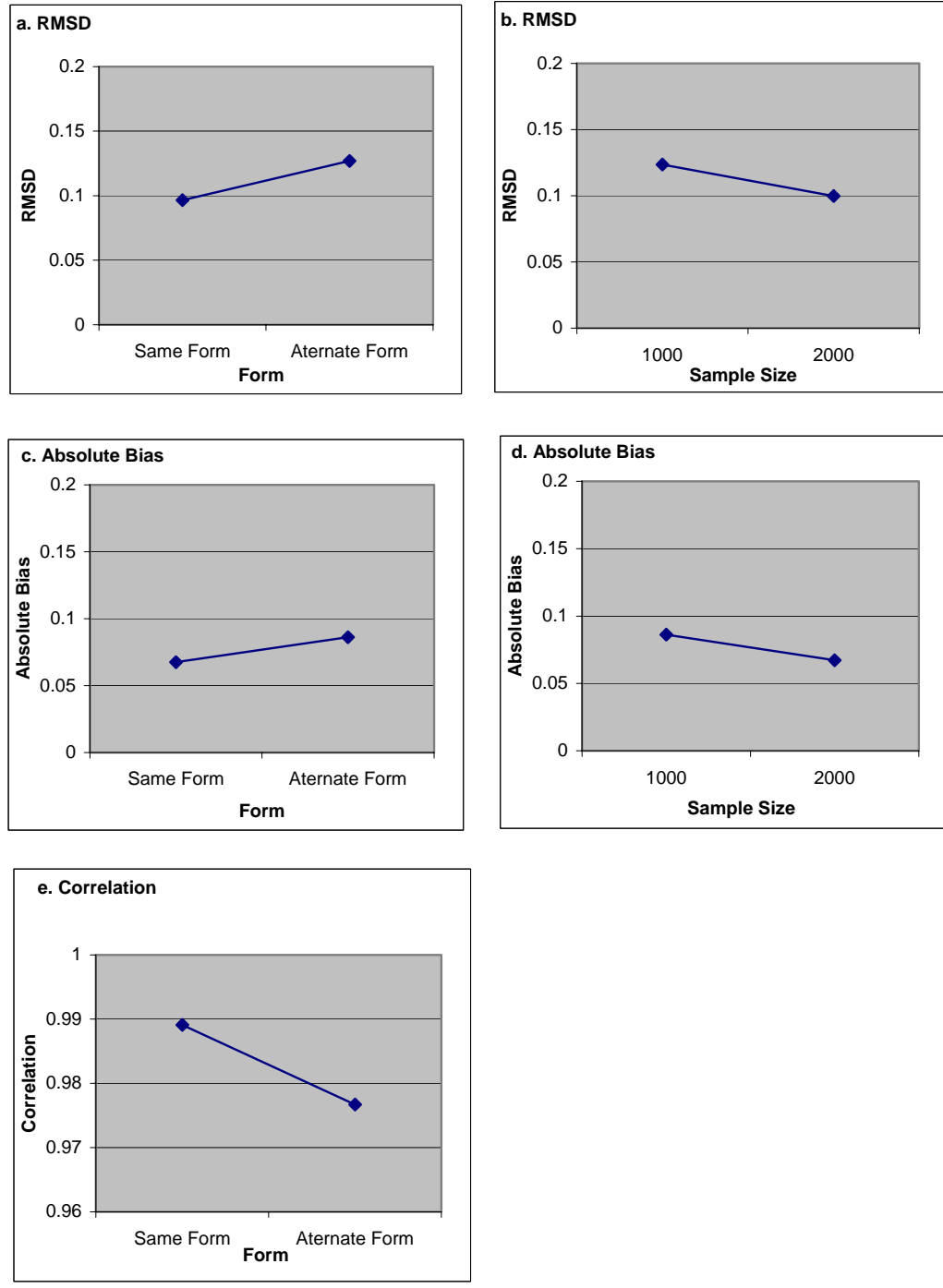


Figure 6: Mean Accuracy Measures of Tau Estimates for 10, 20 and 30 Items

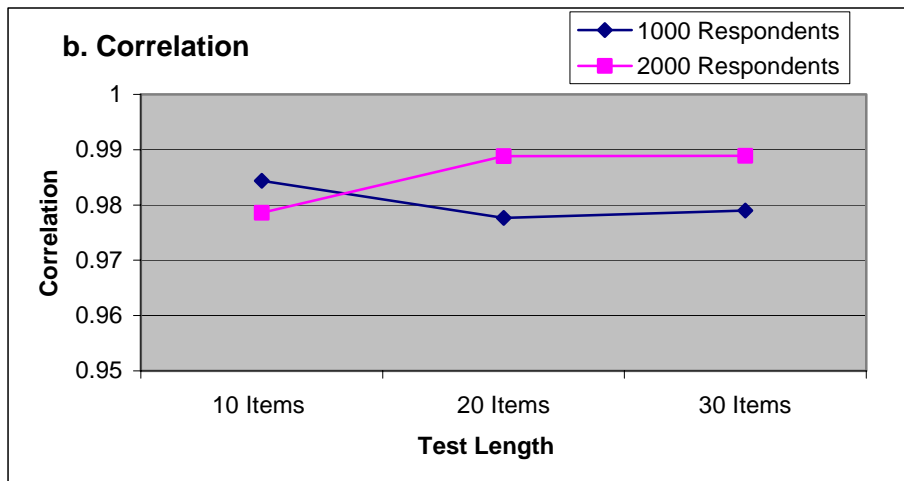
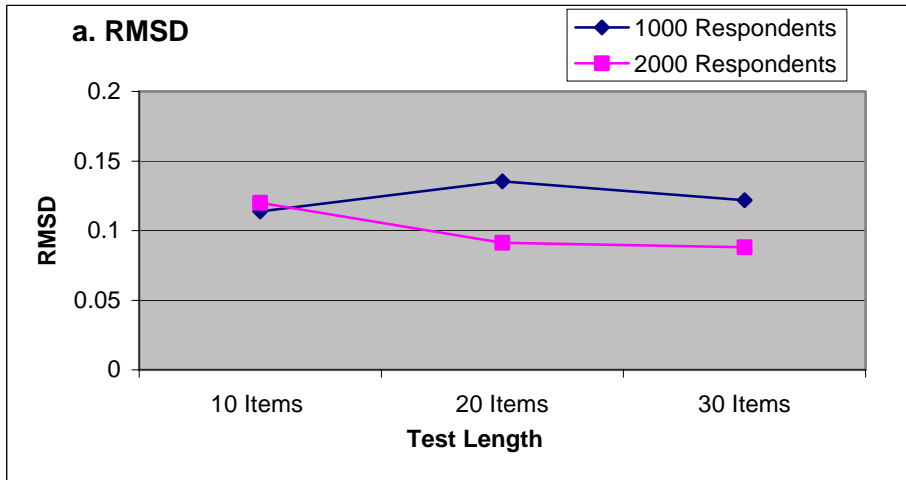


Figure 7: Mean Accuracy Measures of Theta Estimates for 10, 20 and 30 Items

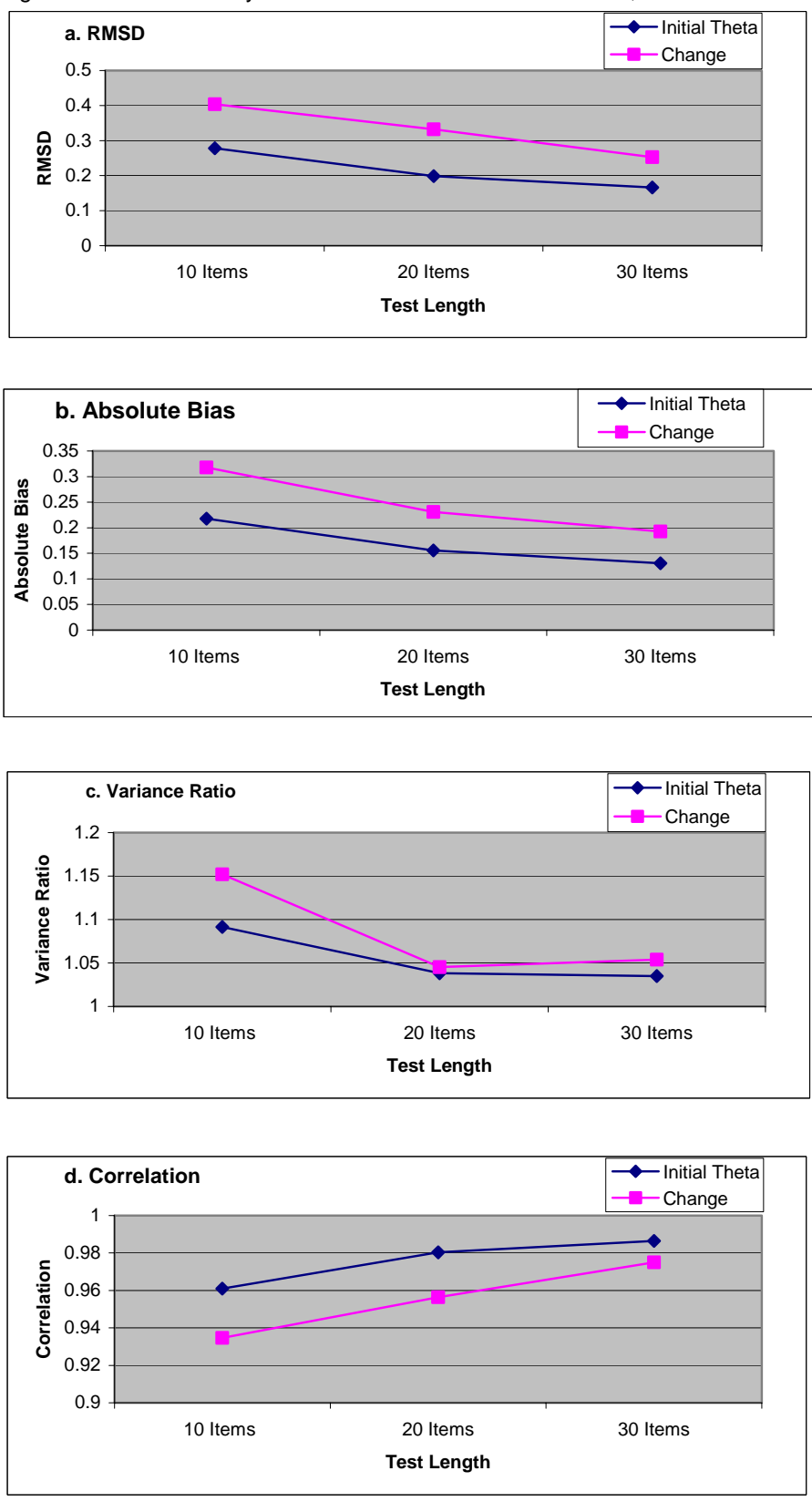


Figure 8: Interaction of Form and Test Length for RMSD and Correlation of Individual Change Estimates

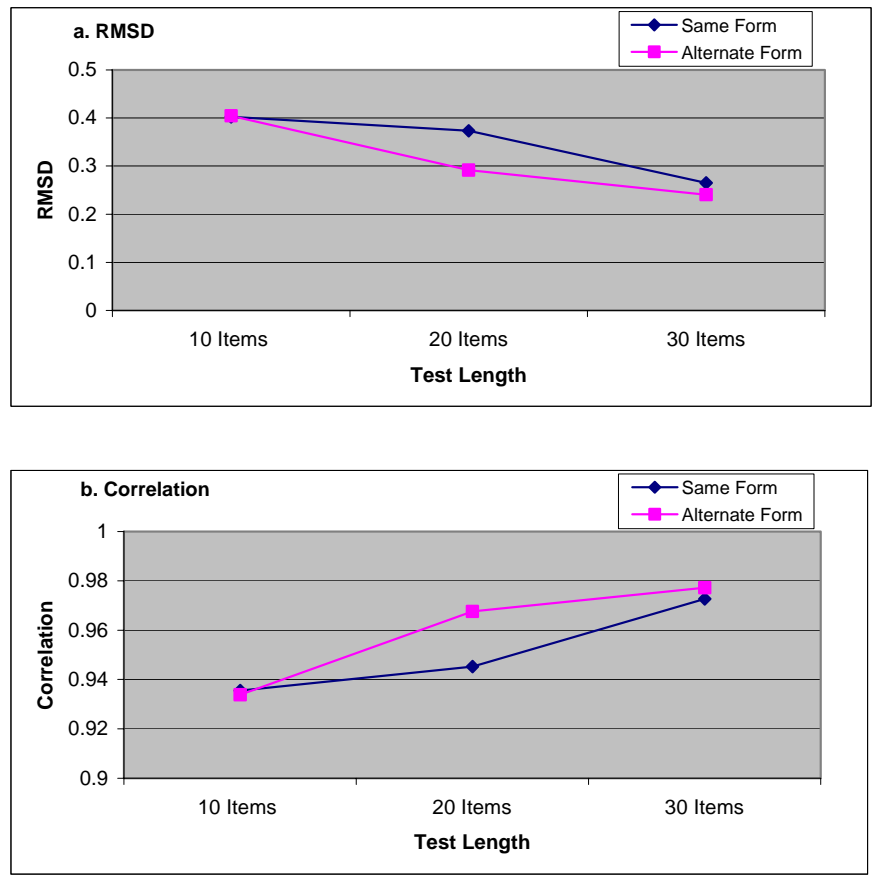


Figure 9: RMSD of Group Change Estimate for 1000 and 2000 Respondents

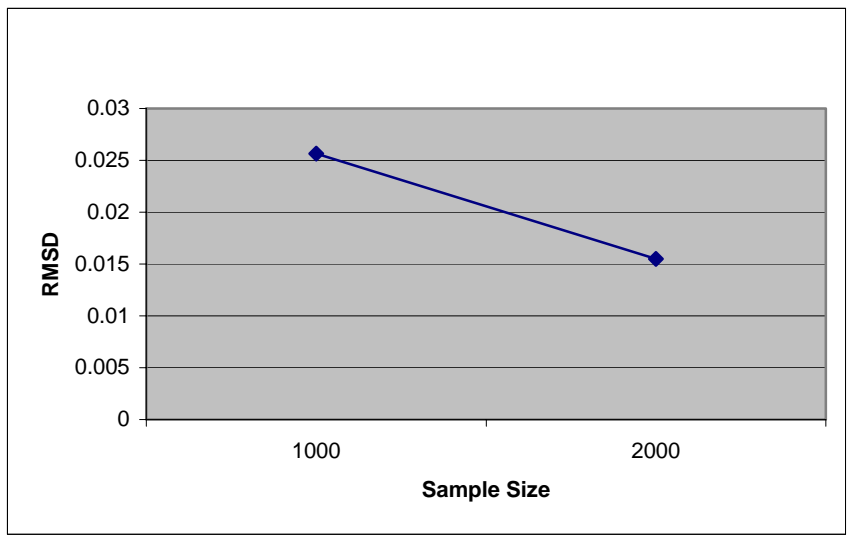
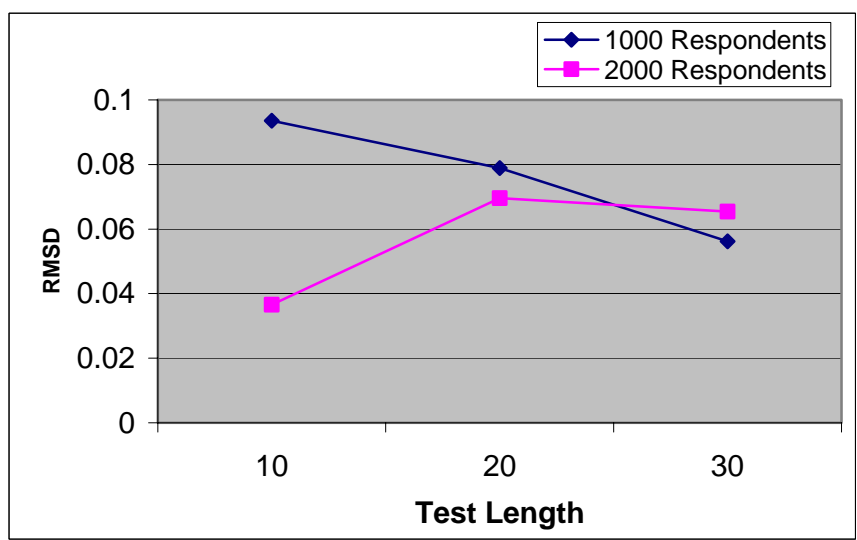


Figure 10: RMSD of Estimated Variance of Latent Distribution for 10, 20, and 30 Items



Real Data Analyses

An analysis was performed to provide at least partial information about the invariance of item parameters across time. Specifically, item parameters were calibrated using 1178 responses collected from the baseline assessment, and then the parameters were calibrated a second time using only the 113 responses collected from the second assessment. Full Bayesian estimation of model parameters was implemented via WinBUGS for both calibrations. Prior distributions for all model parameters were the same as those used in simulation study. Figures 11-13 graphically display the relationships between corresponding parameter estimates derived from the two assessment periods. (i.e., $\hat{\delta}_i$, $\hat{\alpha}_i$, and $\hat{\tau}_{ik}$). As shown in figure 11, the two sets of item location estimates (i.e., $\hat{\delta}_i$) were highly correlated (Pearson $r = .978$), which implies that the estimates of the item location parameters are very stable across assessment times for these abortion statements. However, the two sets of item discrimination and threshold parameter estimates were only moderately correlated. The correlation for discrimination parameter estimates was .613, whereas those for threshold parameter estimates were .519, .745, .720, .649, and .727) for $\hat{\tau}_{i1}$ through $\hat{\tau}_{i5}$, respectively. The relative decrease in these correlations was presumably due to the small sample size at the second assessment time along with the fact that these item parameters are generally harder to estimate than the location parameter (Roberts, Donoghue & Laughlin, 2002).

Figure 11: Scatter Plot of Estimates of Item Location Parameters at Baseline against Estimates of Item Location Parameters at the Second Assessment Time

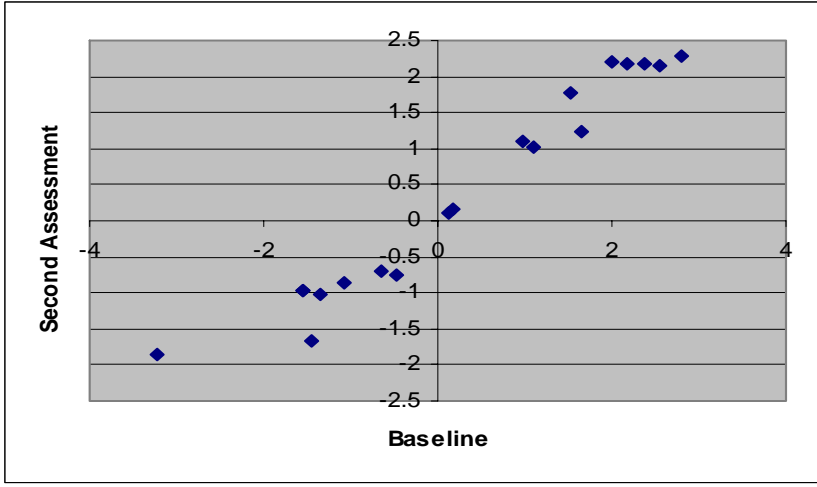


Figure 12: Scatter Plot of Estimates of Item Discrimination Parameters at Baseline against Estimates of Item Discrimination Parameters at the Second Assessment Time

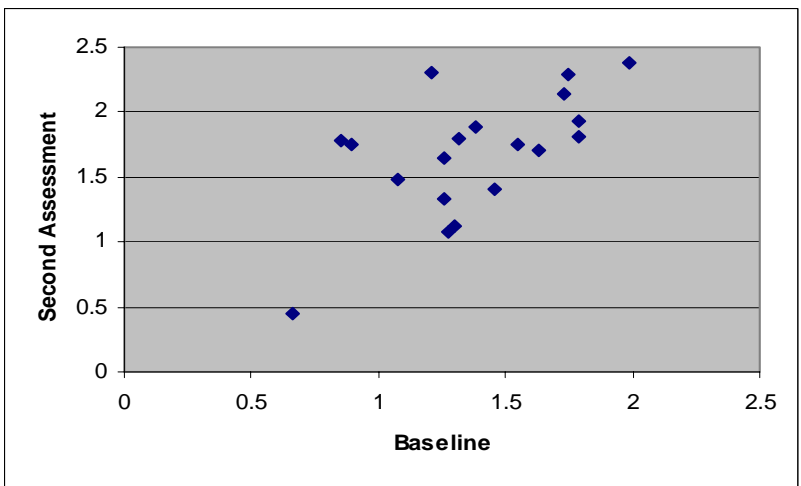
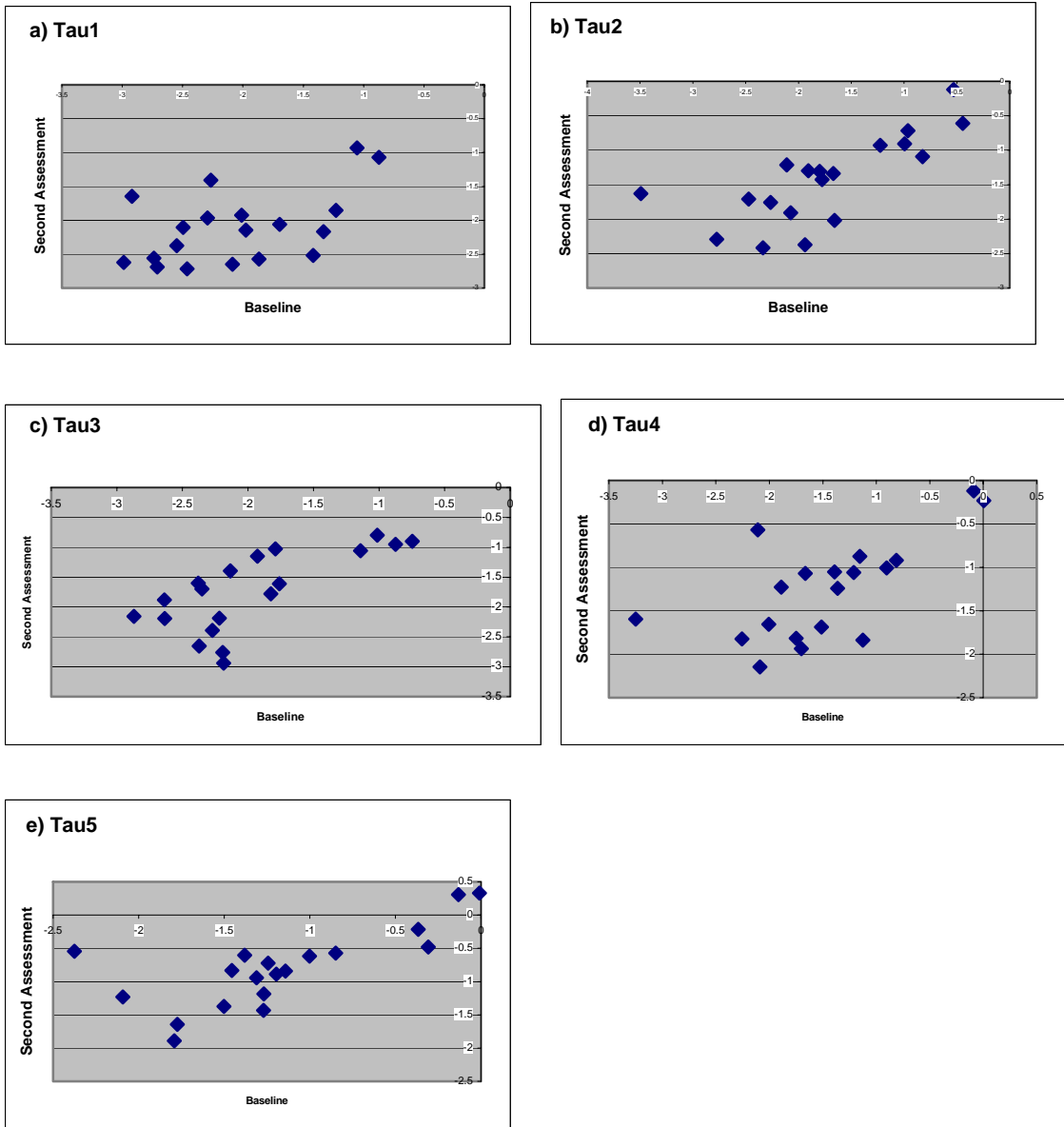


Figure 13: Scatter Plot of Estimates of Item Threshold Parameters at Baseline against Estimates of Item Threshold Parameters at the Second Assessment Time



An analysis of up to 38 responses from 1178 students to the repeated assessments (19 responses at each of 2 assessment times) was performed using the GUUM-RM. As reported above, the 19 responses from the second assessment were missing for all but 113 of these students. The prior distributions used to estimate model parameters and

hyperparameters in the GGUM-RM were identical to those used in the simulation study. The resulting item parameter estimates are given in Table 5 and the statements are listed in order of increasing $\hat{\delta}_i$. The scale was arbitrarily oriented such that the negative pole represented attitudes in favor of abortion, whereas the positive pole represented attitudes against abortion. Aside from this orientation, the calibration results were highly consistent with those obtained using the traditional GGUM in a previous study by Roberts, Donoghue, and Laughlin (2000). Their study was based on the same 750 cases from the University of South Carolina sample used here which may account for some of this consistency. The consistency is still noteworthy given that an additional sample of 428 Georgia Institute of Technology respondents was also included in this calibration.

Table 5: GGUM-RM Item Parameter Estimates ($\hat{\delta}_i$, $\hat{\alpha}_i$, and $\hat{\tau}_{ik}$) for 19 Abortion Attitude Statements

Item	Statement	$\hat{\delta}_i$	$\hat{\alpha}_i$	$\hat{\tau}_{i2}$	$\hat{\tau}_{i3}$	$\hat{\tau}_{i4}$	$\hat{\tau}_{i5}$	$\hat{\tau}_{i6}$
1	Abortion should be legal under any circumstances	-2.63	1.20	-2.30	-2.87	-2.32	-2.63	-1.68
2	A woman should retain the right to choose an abortion based on her own life circumstances	-1.57	1.74	-2.59	-2.08	-2.37	-1.84	-1.36
3	Outlawing abortion violates a woman's civil rights	-1.48	1.76	-2.48	-2.00	-2.30	-1.58	-1.28
4	Society has no right to limit a woman's access to abortion	-1.47	1.27	-2.17	-2.04	-1.81	-1.79	-1.14
5	Regardless of my personal views about abortion, I do believe others should have the legal right to choose for themselves	-1.35	1.23	-2.60	-1.77	-2.66	-1.99	-1.81
6	Although abortion on demand seems quite extreme, I generally favor a woman's right to choose	-1.06	1.77	-2.03	-1.76	-1.83	-1.35	-0.82

7	Abortion should generally be legal, but should never be used as a conventional method of birth control	-0.70	0.91	-1.54	-0.97	-2.13	-1.15	-1.32
8	Abortion should be a woman's choice, but should never be used simply due to its convenience	-0.54	0.88	-1.67	-0.87	-2.31	-1.23	-1.19
9	My feelings about abortion are very mixed	0.12	1.40	-1.07	-0.46	-1.03	-0.01	-0.36
10	I cannot whole-heartedly support either side of the abortion debate	0.17	1.59	-0.91	-0.51	-0.78	-0.09	-0.32
11	Abortion should be illegal except in extreme cases involving incest or rape	1.00	1.23	-1.31	-0.97	-0.89	-0.94	-0.10
12	Abortion is basically immoral except when the woman's physical health is in danger	1.11	1.42	-1.78	-1.22	-1.16	-0.85	0.02
13	Even if one believes that there may be some exceptions, abortion is still generally wrong	1.65	1.51	-2.34	-1.69	-1.78	-1.40	-1.00
14	Abortion should not be made readily available to everyone	1.65	0.60	-2.03	-1.79	-2.42	-1.64	-1.37
15	Abortion could destroy the sanctity of motherhood	2.15	1.08	-2.28	-1.99	-2.38	-1.34	-1.19
16	Abortion can be described as taking a life unjustly	2.54	1.97	-3.12	-2.73	-2.60	-2.10	-1.85
17	Abortion is the destruction of one life for the convenience of another	2.63	1.32	-3.09	-2.87	-2.71	-2.22	-2.13
18	Abortion is inhumane	2.93	1.72	-3.30	-3.00	-2.97	-2.55	-2.34
19	Abortion is unacceptable under any circumstances	3.28	1.24	-2.53	-2.76	-2.29	-2.77	-1.90

The estimated locations ($\hat{\delta}_i$) for these 19 items ranged from -2.63 to 3.28 . As shown in Figure 14, they were nicely spread along the latent attitude continuum, and represented the entire range of attitudinal positions. The locations of statements on the latent continuum corresponded well with the content of the statements. Statements

located near the negative pole corresponded to attitudes strongly in favor of abortion, whereas statements located near the positive pole represented attitudes strongly against abortion. Those statements oriented at the middle of the scale conveyed mixed feelings about abortion. Statements between the middle of the scale and the extremes were more moderate in content, but not neutral.

Items typically exhibited moderate to high $\hat{\alpha}_i$ values. However, there was one item (Item 14) that exhibited a relatively low $\hat{\alpha}_i$ value. The discrimination of a GGUM-RM item is a function of both discrimination parameter and the interthreshold distance, and thus, a low discrimination parameter value does not necessarily lead to low discrimination for a given item. Figure 15 shows the ICC for Item 14. Even with its relatively low estimate for α_i , this item is moderately efficient for determining the location of respondents on the latent continuum in the neighborhood of $\hat{\delta}_i$.

Most of these items (15 out of 19 items) showed disordinal thresholds. The occurrence of disordinal thresholds suggested that for those items, one or more response options were used infrequently. However, Roberts & Ma (2006) pointed out that disordinal thresholds occurred often in IRT analysis of self-report questionnaire data and were generally not a threat to the underlying validity of the model.

Figure 14: Estimated Item Locations for 19 Abortion Items

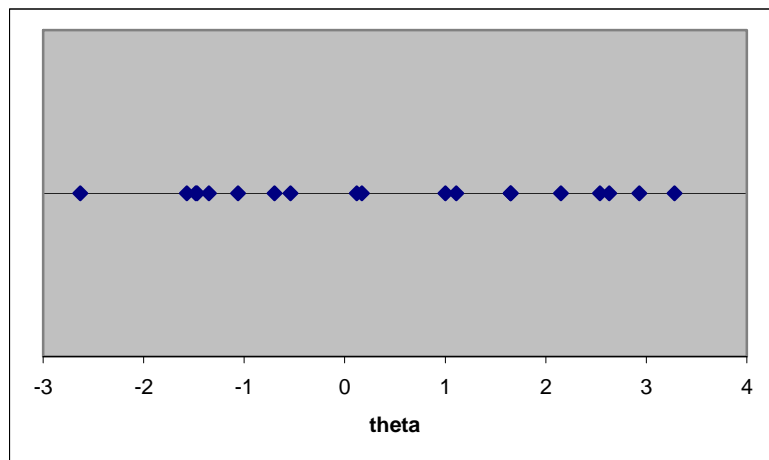
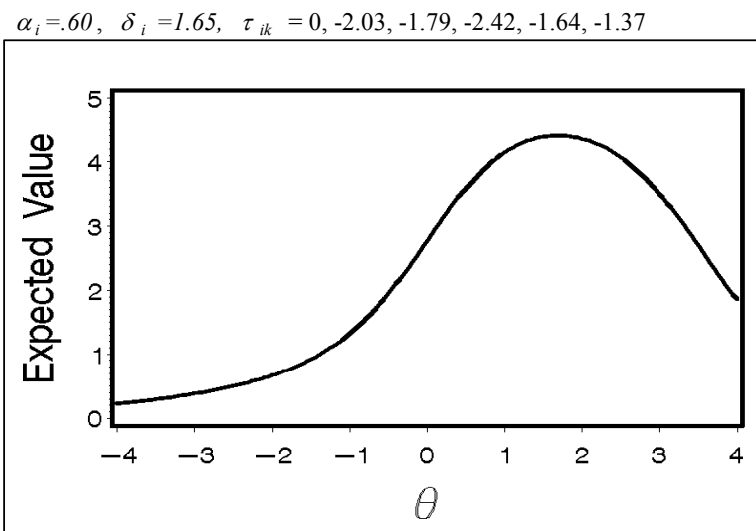


Figure 15: The ICC for Item 14

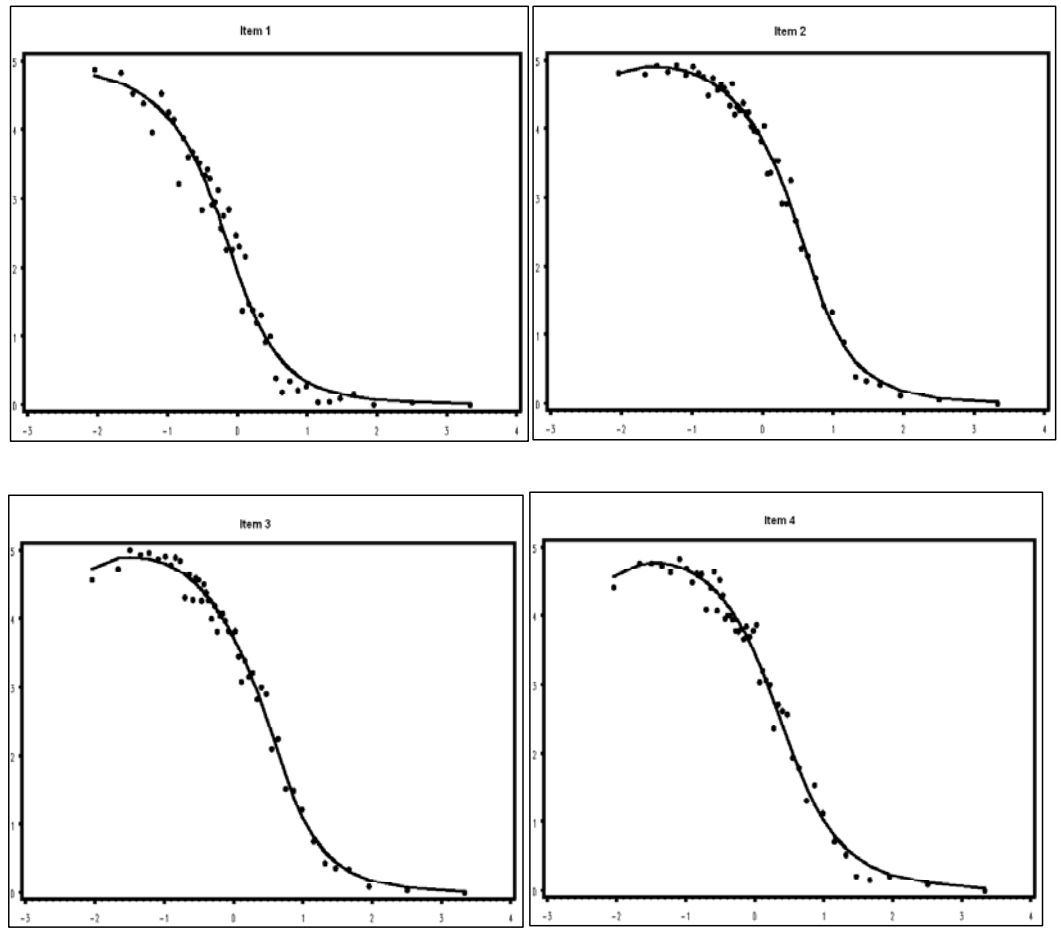


The fit of the GGUM-RM was graphically examined at an item level. The composite latent trait scores were first ranked, and then grouped with 50 respondents in each group but the last group, which was formed by the remaining 6 respondents located at the positive extreme. The average observed item response and the average response predicted by the model were plotted against the average composite theta in each group.

The degree of item fit was evaluated by how well the average responses predicted by the model matched the corresponding observed averages. Figures 16-20 illustrate these fit plots for each of the 19 items. In each figure, the solid dot represents the average observed response for a given respondent group, whereas the smooth curve represents the average expected response predicted by the model. The observed and expected average values were generally comparable, except for Item 11, 13 and 14. Item 11 had two misfit groups, Item 13 and 14 had one misfit group on the positive extreme side of the scale due to the small size of the last group (6 remaining respondents at the positive extreme). Thus, there was evidence that suggested the model performed reasonably well for the abortion questionnaire.

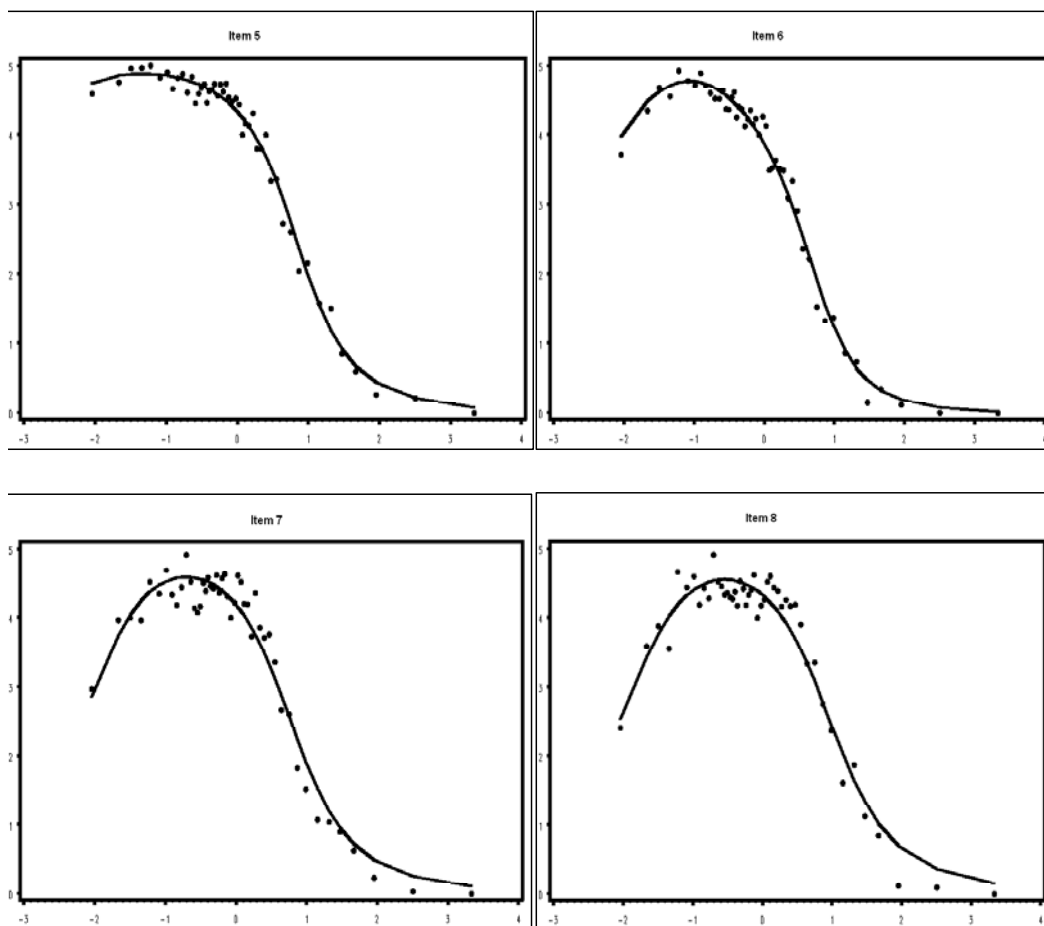
As shown in Figure 16-20, 9 of 19 items had ICCs that were (more or less) monotonically increasing (items 16 through 19) or monotonically decreasing (items 1 through 5). These statements represented opinions that were either very pro-life or very pro-choice, respectively. Six of 19 items had ICCs that exhibited a noticeable amount of folding (i.e., nonmonotonicity) in either the positive (items 13, 14, and 15) or the negative (items 6, 7, and 8) regions of the continuum. These items corresponded to moderately pro-life or moderately pro-choice orientations, respectively. Both types of ICCs were consistent with the assumptions of an unfolding model. The remaining 4 items (items 9 through 12) were somewhat neutral in their content and exhibited markedly folded ICCs. The ICCs of these four items were single-peaked and decreased at both extremes of the scale as predicted as GGUM-RM. In summary, 10 of 19 items showed a nonmonotonic pattern, which indicated that an unfolding model was more appropriate than a cumulative model for responses to this abortion questionnaire.

Figure 16: Average Observed Versus Expected Item Responses by Theta Group for Items 1-4.



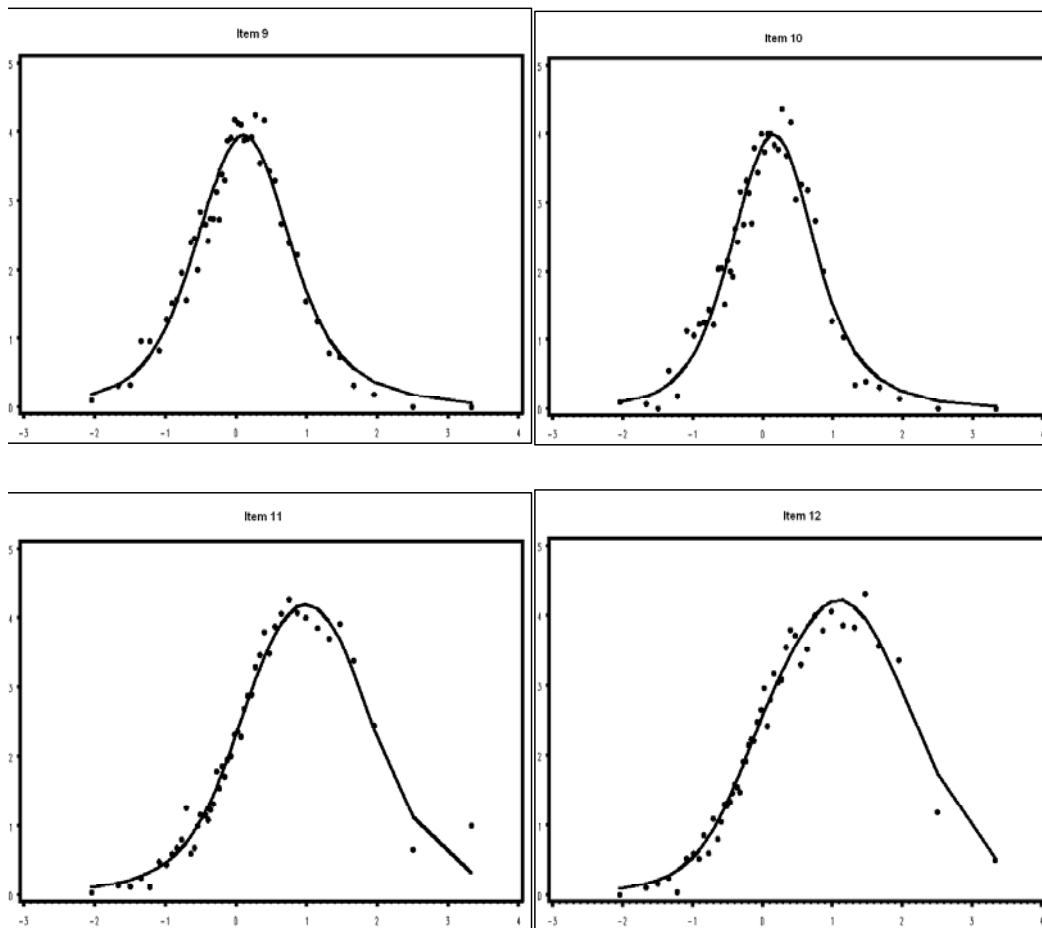
Note: Dots represent average observed responses;
Smooth curve represents average expected responses.

Figure 17: Average Observed Versus Expected Item Responses by Theta Group for Items 5-8.



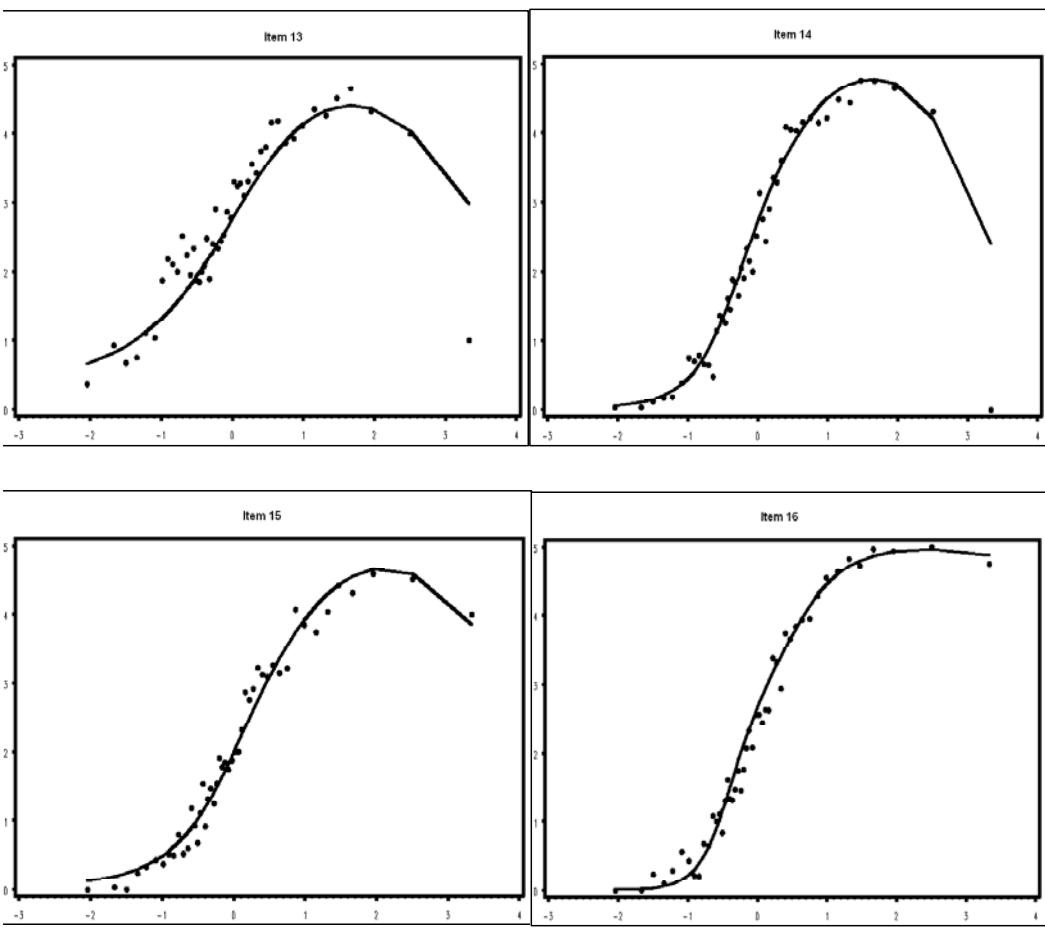
Note: Dots represent average observed responses;
Smooth curve represents average expected responses.

Figure 18: Average Observed Versus Expected Item Responses by Theta Group for Items 9-12.



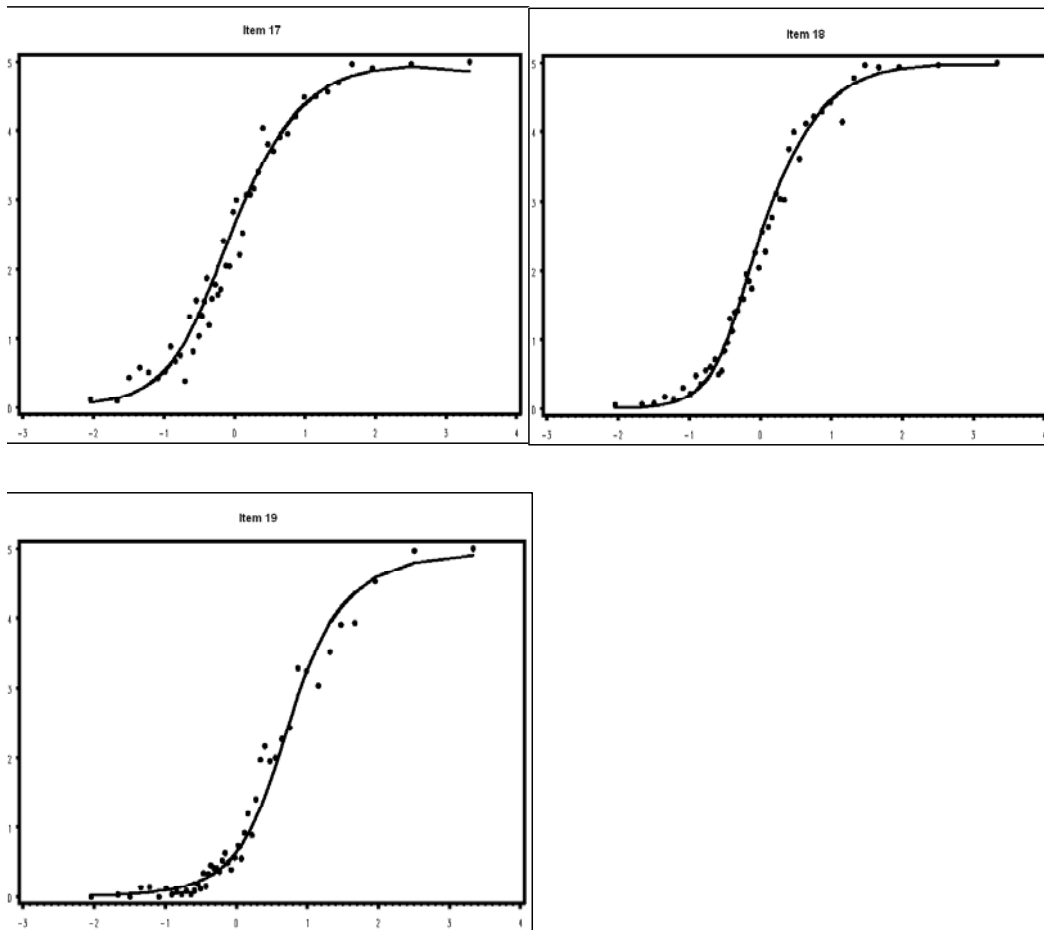
Note: Dots represent average observed responses;
Smooth curve represents average expected responses.

Figure 19: Average Observed Versus Expected Item Responses by Theta Group for Items 13-16.



Note: Dots represent average observed responses;
Smooth curve represents average expected responses.

Figure 20: Average Observed Versus Expected Item Responses by Theta Group for Items 17-19.

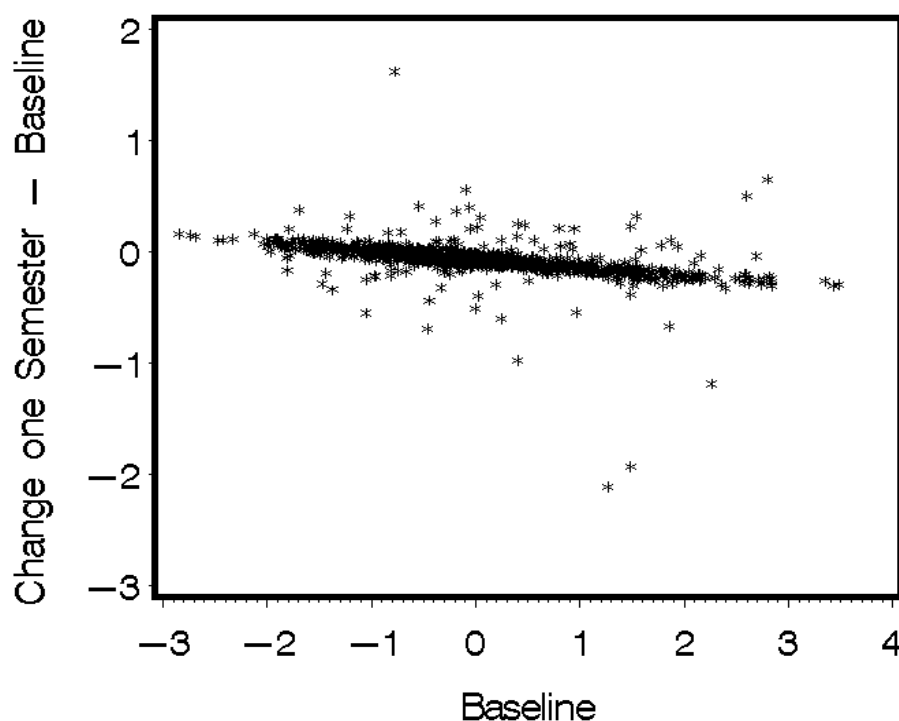


Note: Dots represent average observed responses;
Smooth curve represents average expected responses.

The model also provided direct estimates of hyperparameters of the multivariate normal distribution used to model the latent variables. In this analysis, the mean and variance of the baseline latent trait (i.e., baseline attitude) were constrained to be 0 and 1 for identification purposes, whereas the mean and the variance of the latent attitude change, and the covariance between attitude change and the baseline attitude were estimated. The estimated average abortion attitude change was equal to -0.066 , which implied that there was little, if any, attitude change within the approximate 3-week test-retest interval. This was expected given the short testing interval and the lack of any

experimental attempt to change attitudes. The estimated variance for the attitude change and the covariance between attitude change and baseline attitude were .244 (standard deviation of .494) and -.074, respectively. As expected, a negative correlation between attitude change and the baseline attitude was observed in this analysis ($r = -.151$). This negative correlation can also be seen in Figure 21, in which, the individual change estimates are plotted against their baseline levels. Students with neutral baseline attitudes exhibited little change across testing occasions, whereas students with extreme baseline attitudes for or against abortion showed slightly noticeable change that mitigated their extreme positions somewhat.

Figure 21: Scatter Plot of Estimate of Individual Change against Their Initial Level



As shown in Figure 21, there were 4 respondents whose absolute change estimates were greater than 1 unit. The response patterns for these 4 respondents are

shown in Table 6. Recall that an item score of 5 corresponded to “strongly agree” and a score of 0 represented “strongly disagree”. Visual inspection of the responses for these four respondents suggests that they did change their attitude toward abortion across assessment times.

Table 6: Responses for 4 Respondents with Absolute Change Estimates Greater than 1

Student ID	Assessment Time	Responses
1096	Baseline	5555534440010400000
	Second Time	0000010120054434545
1098	Baseline	5555555550000000000
	Second Time	0000001123334444455
1137	Baseline	13110. 1000055345555
	Second Time	4535554505510300000
1150	Baseline	5555450050000000000
	Second Time	4045254432111111101

VII. Discussion and Conclusions

Discussion

The simulation results suggest that the model estimation procedure is computationally feasible. For the item parameters, accurate estimates can be obtained with a sample size of 1000 via the full Bayesian estimation method when using 4-category items. Previous simulation studies showed that when using marginal maximum likelihood estimation, accurate estimates of item parameters can be obtained with 750 – 1000 respondents when 6 response category items are used (Roberts, Donoghue, and Laughlin, 2002) or with 1000 respondents when 4-category items are used (Cui, Roberts, and Bao, 2004). Although GGUM-RM is more complex than the unidimensional GGUM, the Bayesian estimation technique used by GGUM-RM provided prior information for all model parameters (i.e., item parameters, person parameters and hyperparameters) during model estimation, whereas MMLE used by unidimensional GGUM in previous simulation studies only provided prior information for person parameters. The additional prior distributions used to estimate GGUM-RM parameters apparently provided enough supplementary information in the solution to mitigate the need for larger samples.

For the person parameters, accurate estimates of an individual's initial level (θ_{j1}^*) and change (θ_{j2}^*) can be obtained when using 20 and 30 items, respectively. That is, more items would be needed to obtain equivalently accurate estimates of individual change than an estimate of the individual's initial level. This is due to the model parameterization in the GGUM-RM. The GGUM-RM formulates the category

probability function using a composite score (i.e., $\theta_{j2} = \theta_{j1}^* + \theta_{j2}^*$), thus, during model estimation, all responses collected at both assessment times are used to estimate the individual's initial level (θ_{j1}^*), but only those responses collected from the second assessment time are used to estimate the individual's change (θ_{j2}^*). As a result, the estimates of the individual's initial levels obtained by the GUUM-RM were more accurate than the estimates of individual change.

Roberts, Donoghue, & Laughlin (2000) pointed out that the data demands of 20 items and 1000 respondents for accurate model parameter estimation may exceed the resources of applied researcher; however, if suitable item parameter estimates are available, for example, when published standardized test/questionnaires are available, then accurate estimates of a single individual's change over time can be obtained using GGUM-RM. Thus, this model could be applicable in small scale testing applications.

In practice, alternate forms with anchor items may be used in order to alleviate potential contaminations caused by memory and learning effects. However, using alternate forms across assessment times may bring both advantages and disadvantages to model estimation of the GGUM-RM. Specifically, when holding other conditions the same, using alternate forms across assessment times results in less accurate estimates of item parameters, but more accurate estimate of individual change than using the same form across assessment times. The inaccuracy in item parameter estimates is due to the fact that there are fewer individual responses to all unique (i.e., non-anchor) items, In contrast, the additional accuracy in estimates of individual change results from the fact that each individual responds to more distinct items in the alternate forms condition. Person parameters in the GGUM family of models become more accurate as the number

of distinct items administered increases. If items with identical characteristics were administered to an individual then these models would not indicate whether the individual was located above or below the item on the latent continuum.

For the person parameter estimates, the simulation results showed that whether the same form or the alternate forms were used across assessment times had a negligible impact on the estimate of the individual's initial level, but it had a significant influence on the estimate of the individual's change in the 20 items condition. Specifically, the model provided a better estimate of individual change in the alternative forms with 20 items than in the same form with 20 items. Two possible reasons can be used to explain this. First, there may be sampling fluctuation, since this recovery study only simulated 10 replications in each combined condition. The small number of replications may have led to larger sampling fluctuation and the form effect on individual change estimates observed in this study could have been caused simply by chance. If more replications had been simulated within each combined condition, then the form effect might disappear.

A second possible reason for the alternate form effect encountered in the 20-item condition relates to the fact mentioned earlier that the accuracy of person parameter estimates in the GGUM family of models improves as the number of distinct and informative items increase. The notion of distinct items is key to this explanation and alternate forms provide more distinct items than do common forms. Given this basic tenant, why then, did the alternate form effect only emerge in the 20-item condition? The answer may relate to the minimum number of items required to obtain fairly accurate estimates of person parameters in the GGUM family. Cui, Roberts & Bao (2004) suggested that tests/questionnaires with 20 4-category response items that are evenly

distributed along the entire scale are required to obtain reasonably accurate estimates of person parameters. In the 10-item condition from the present study, the number of distinct items administered across assessment times was less than 20 items in both the common form and the alternate forms conditions; thus, the test may have been too short to provide a good estimate of individual change, regardless of whether the same form or alternate forms were administered across assessment times. In the 30-item condition, the number of distinct items administered across assessment times was more than 20 in both the common form and the alternate forms conditions; thus, the test was long enough to give good estimates of individual change in both form conditions. However, in the 20-item condition from the present study, the number of distinct items in the common form condition was equal to 20, which Cui et al. suggested was the minimum number of items required to achieve reasonably accurate estimates. In contrast, there were 34 distinct items in the alternate forms condition, which clearly exceeded the test length suggestions made by Cui and colleagues. Consequently, more accurate estimation of individual change was obtained in the alternate forms condition than in the same form condition when a test length of 20 items was used.

With respect to the accuracy of item parameter estimation in the GGUM-RM, the simulation results indicated that when alternate forms were used across assessment times, estimates of item parameters were less accurate than those obtained by administering the same form across assessment times. That is because when the same form is used across assessment times, responses collected at all assessment times can be used to calibrate item parameters, but they can only be used to calibrate anchor items when alternate forms are used across times, and those unique items that appeared on one form can only be

calibrated using the responses collected at that specific assessment time. Thus, using the same form across assessment times can provide more accurate calibration of item parameters when using GGUM-RM than using alternate forms across assessment times.

In longitudinal data analysis, researchers are also interested in assessing group change. The GGUM-RM can provide a direct estimate of group change via estimating hyperparameters of the latent distributions. Simulation results suggested that very accurate estimates of group change can be obtained with 1000 respondents using a 10-item test/questionnaire. One advantage of directly estimating the group change across assessment times, instead of averaging the estimates of individual change, is that it avoids the impact of the shrinkage of EAP estimate of individual change when using Bayesian estimation. This implies that if the main purpose of a study is to assess group change over time, then equally or more accurate estimates of group change can be obtained with fewer items compared to the estimate of individual change.

Usually researchers are interested in estimating change over time in longitudinal data analysis, but sometimes the latent trait level at any assessment time may also be of interest. An individual's latent trait level at any assessment time can be easily derived from a composite score (i.e., $\hat{\theta}_{j_2} = \hat{\theta}_{j_1}^* + \hat{\theta}_{j_2}^*$) when using GGUM-RM to analyze responses collected from repeated measures designs. Then, a direct estimate of this composite score along with the standard error of the estimate can be obtained through WinBUGS by simply monitoring the composite ($\hat{\theta}_{j_2}$). This can be an advantage for those applied researchers who are interested in both individual composite scores over time (i.e., profiles) as well as individual change.

Limitations

Findings from this study are encouraging and suggest that the use of GGUM-RM in large scale testing practice is viable. However, generalizations from the results of this study are limited to similar conditions represented herein. The estimation strategy employed in this study is only appropriate for repeated measures designs with two assessment points. It cannot simply be generalized to situations that have more than two assessment times due to the limitations of WinBUGS with respect to constraining the variance-covariance matrix. If individuals are administered the test more than two times, more constraints are needed to make sure that the variance-covariance matrix is positive definite.

The GGUM-RM presumes that within the same assessment administration, a unidimensional GGUM holds and item parameters are invariant across assessment points. That is, only individuals change their position on the latent continuum, but the way in which items measure the construct remains constant over time. This assumption may or may not be true for a given item in practice. For example, the construct being measured may remain the same across assessment times or it might possibly change such that old dimensions disappear or new dimensions emerge. Reckase and Martineau (2004) demonstrated how the dimensions, measured in the context of vertical scaling of education proficiency tests, changed across cross-sectional examinee samples with increasing proficiency levels. The GGUM-RM might possibly be extended to accommodate such dimensionality change(s) by releasing certain constraints implemented in the current model. However, the estimation of model parameters would be quite difficult and time consuming due to the increased complexity. Further study is

needed to see if stable estimates of model parameters over time can be obtained generally in common attitude measurement situations.

Conclusion

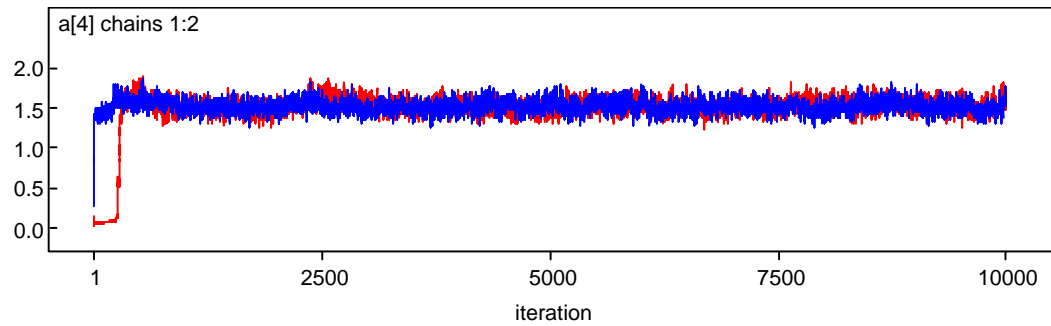
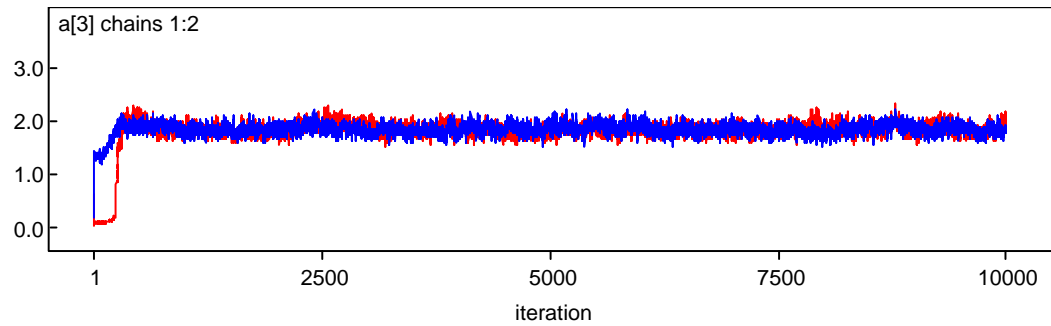
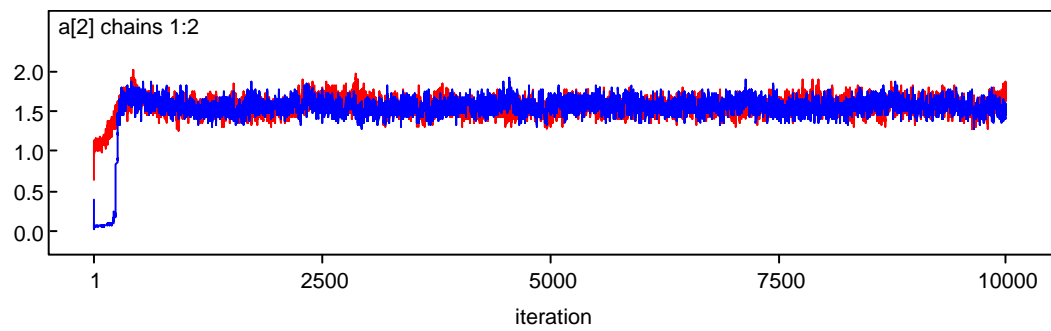
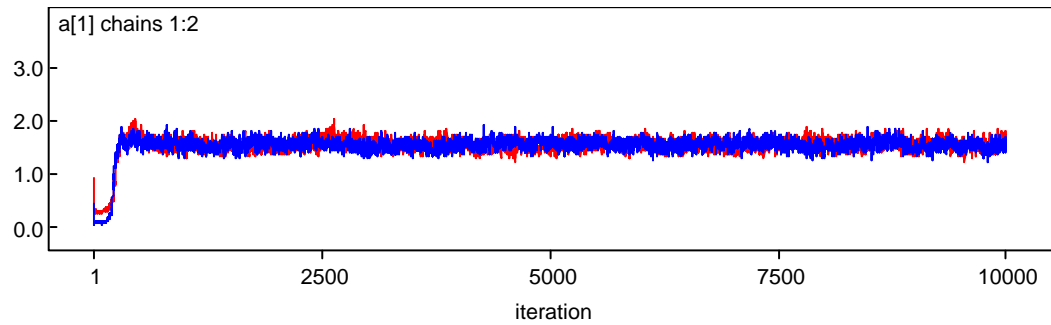
Assessing change in attitudes over time either at the individual level or at the group level using longitudinal data is likely to be of great interest in the future. Both researchers and practitioners are concerned about the appropriateness and the adequacy of psychometric methods available for this purpose. This study has extended unidimensional GGUM to a multidimensional format (GGUM-RM) and used it to directly estimate both individual and group change over time for repeated measures designs, while accounting for the dependency between latent trait variables at multiple assessment points. Simulation results suggest that estimation of model parameters is computationally feasible and the application of the model in real test practice is plausible.

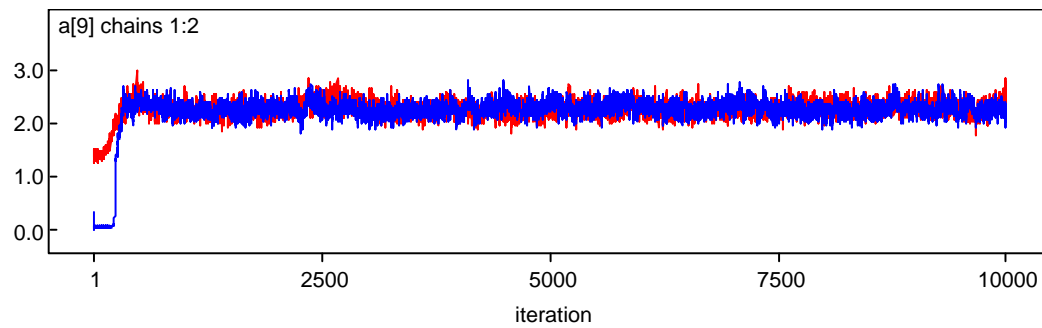
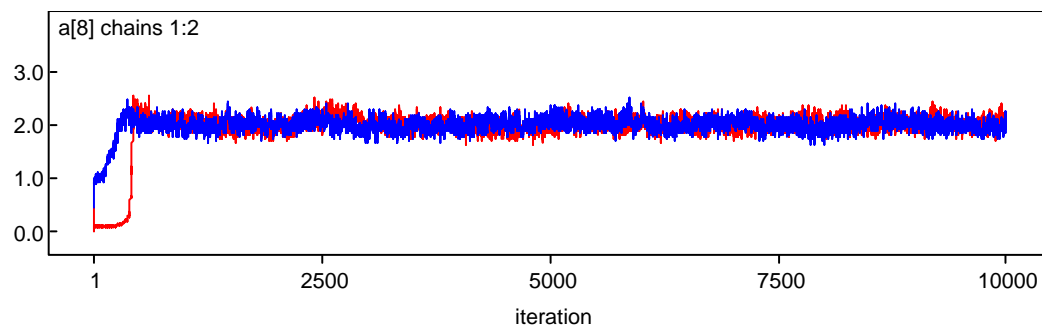
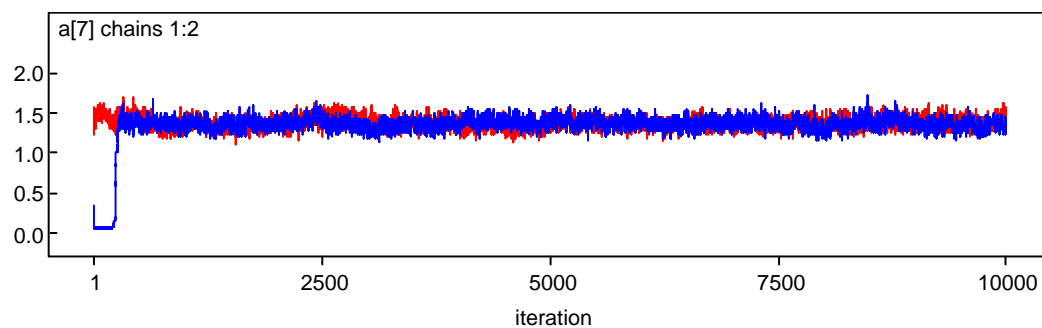
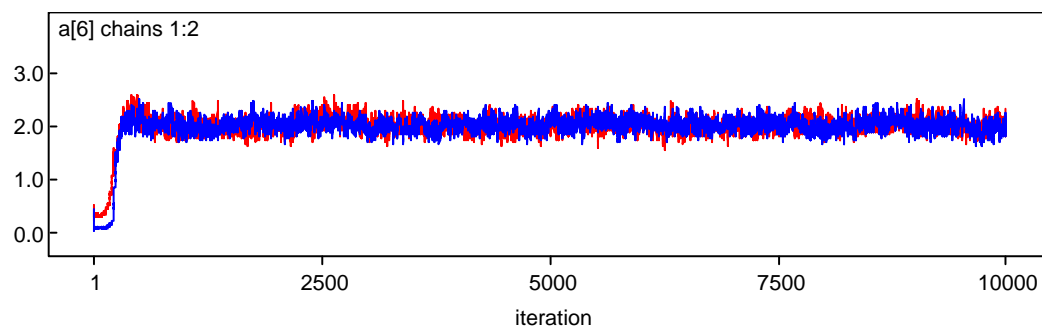
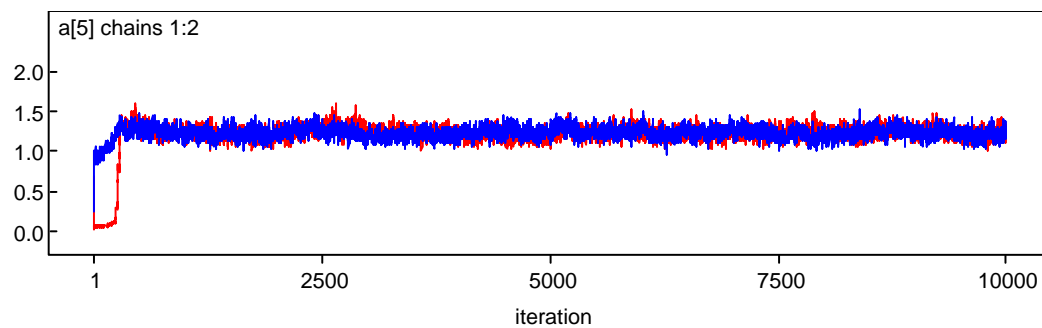
The GGUM-RM fits a non-linear function between latent change and observed scores, consistent with the fact that true changes are not linearly related to gain scores. The GGUM-RM uses standard errors to evaluate the precision of estimates of individual change, and thus, the reliability paradox becomes irrelevant (i.e., the reliability of gain scores decreases as the correlation between pretest and posttest increases). Also, direct estimates of group change provide researchers with a realistic option to avoid the problem of negative correlation between individual change estimates and initial levels when they are more interested in the change of a group. Lastly, the GGUM-RM is a parametric unfolding IRT model and is more appropriate for constructs measured by the

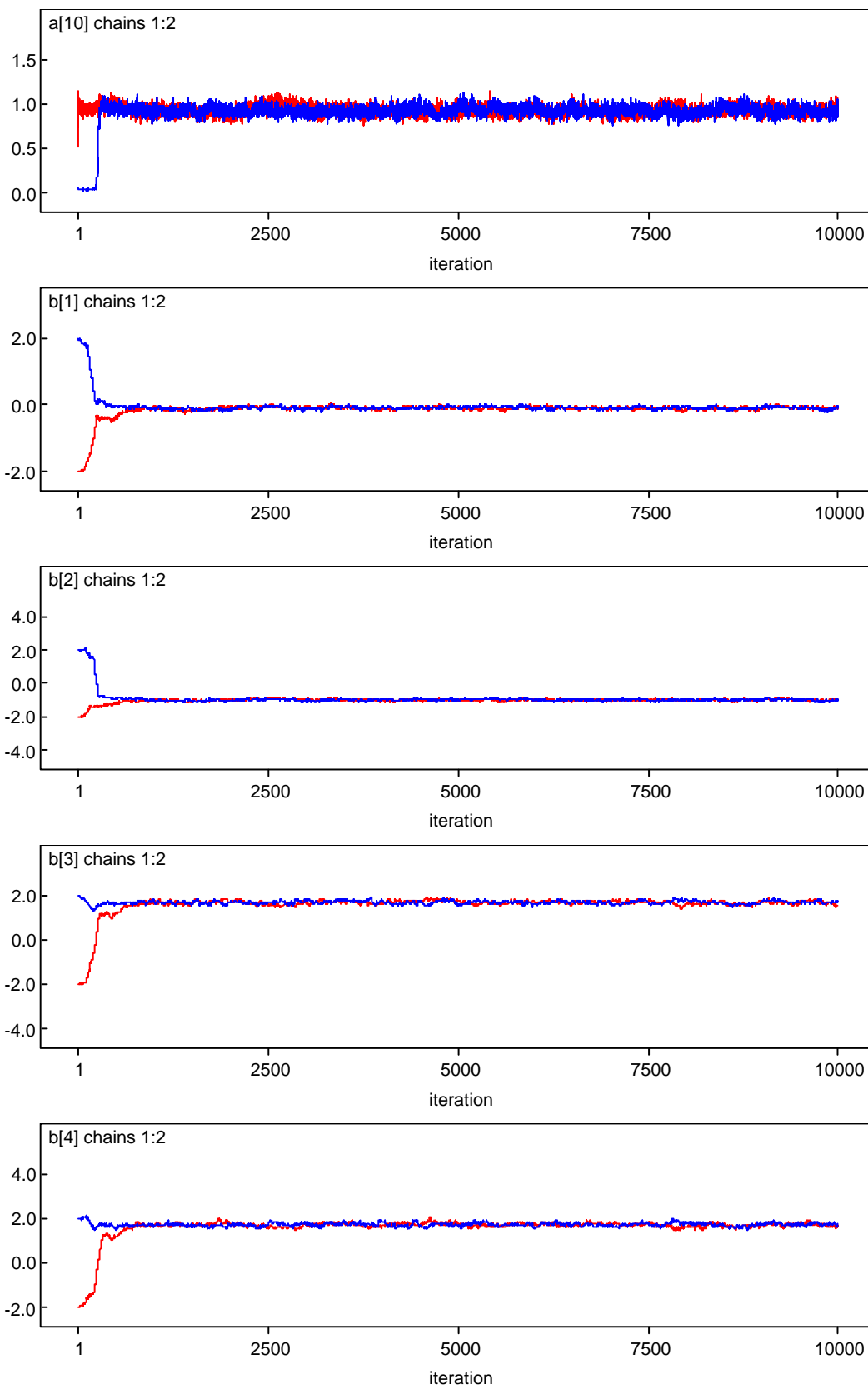
proximity-based response processes, as is typically the case with traditional attitude questionnaires.

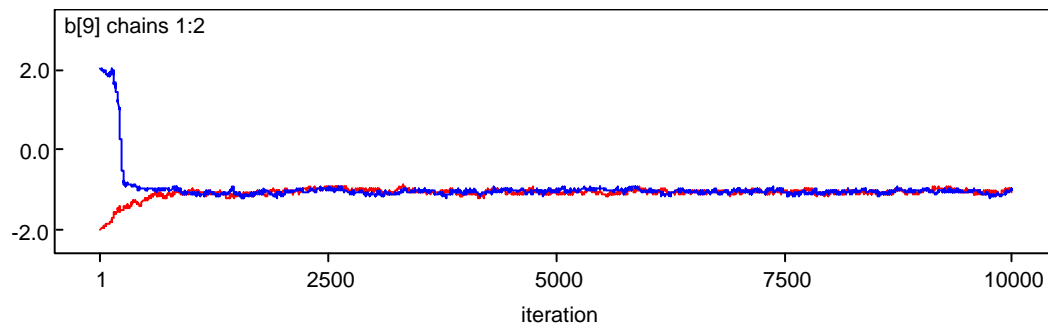
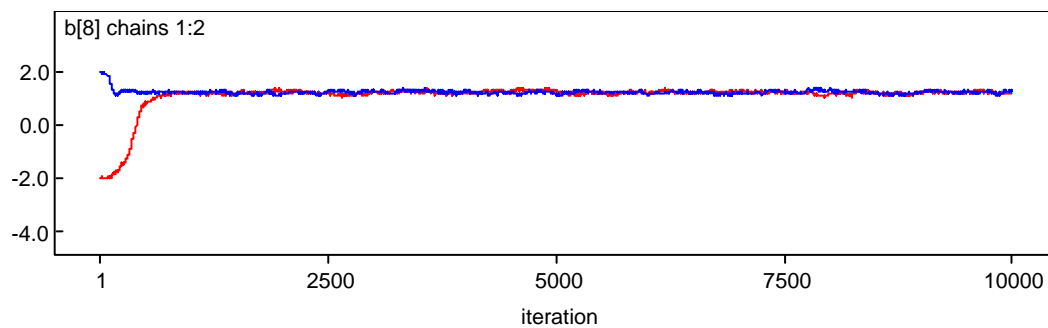
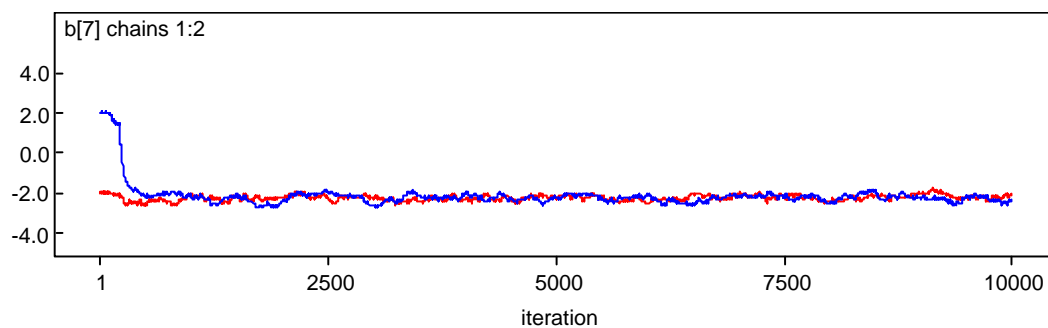
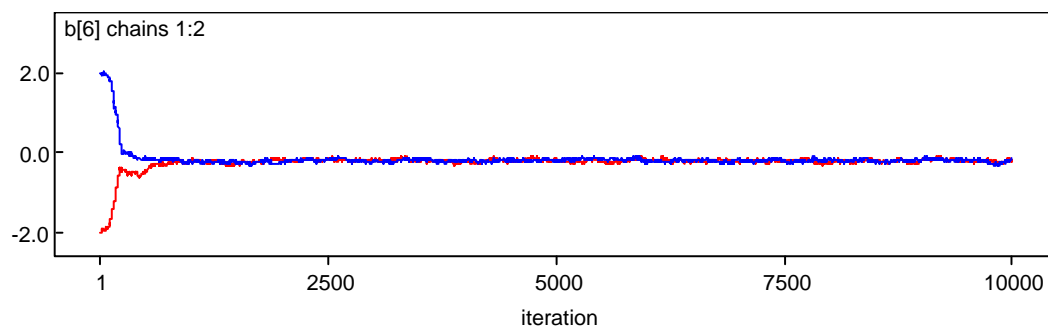
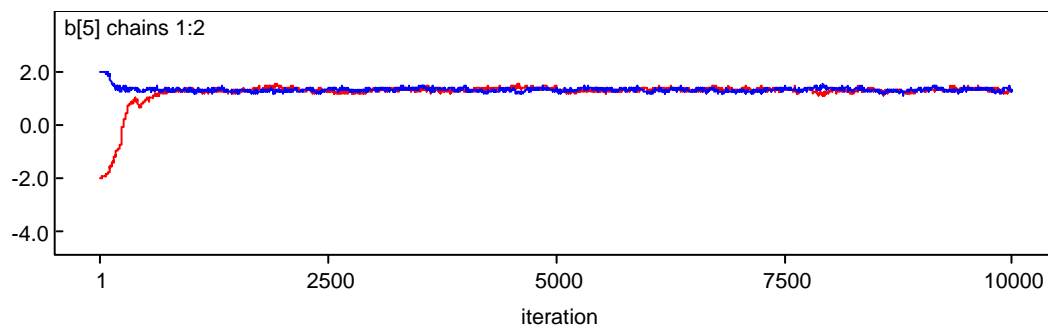
In summary, the GGUM-RM can provide appropriate and adequate estimates of change in unidimensional constructs over time when the binary or graded *agree-disagree* responses are collected from repeated measures designs using Likert or Thurstone attitude questionnaires, and the model may be very useful in many applications.

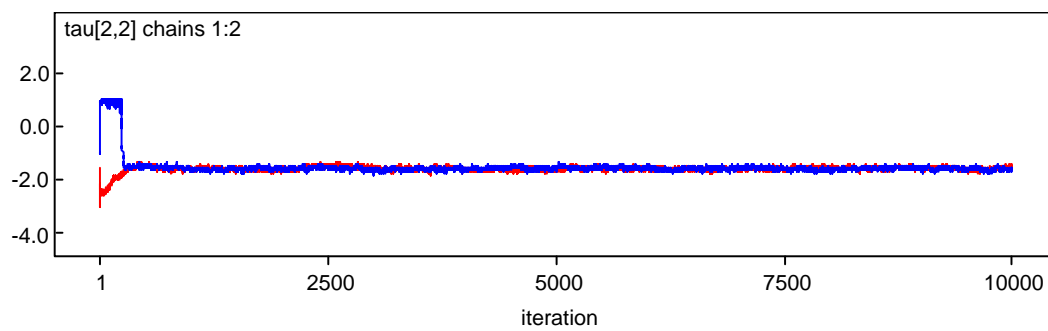
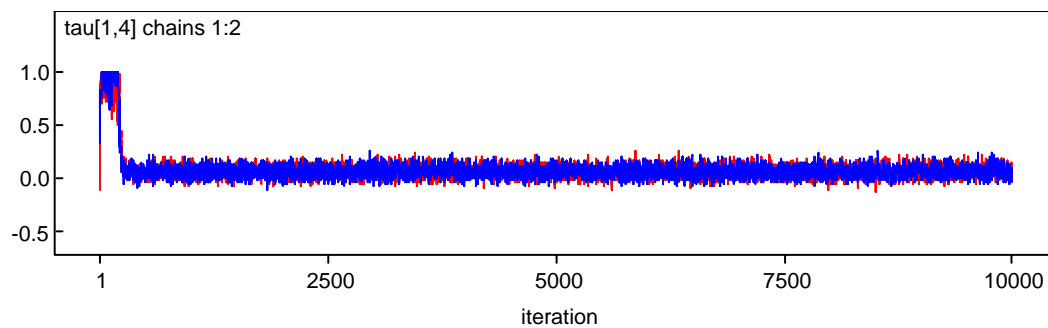
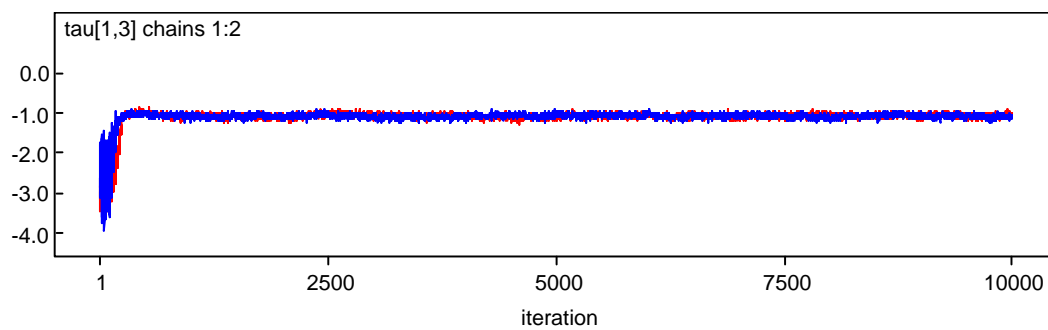
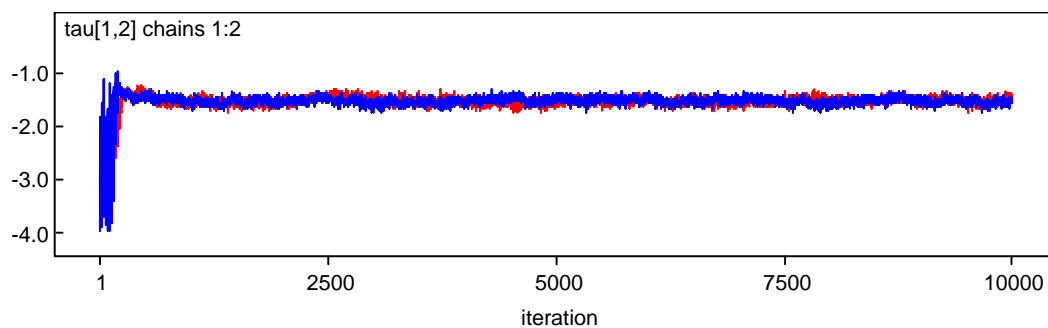
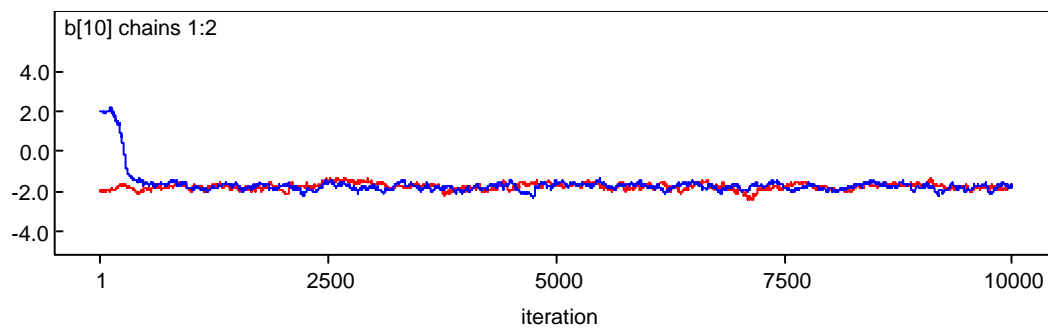
Appendix A: Time-series Plots for Alpha, Delta and Tau for Two Chains of 10000 Iterations for 10 Items, and Thetas for First Five Persons for Two Chains of 10000 Iterations

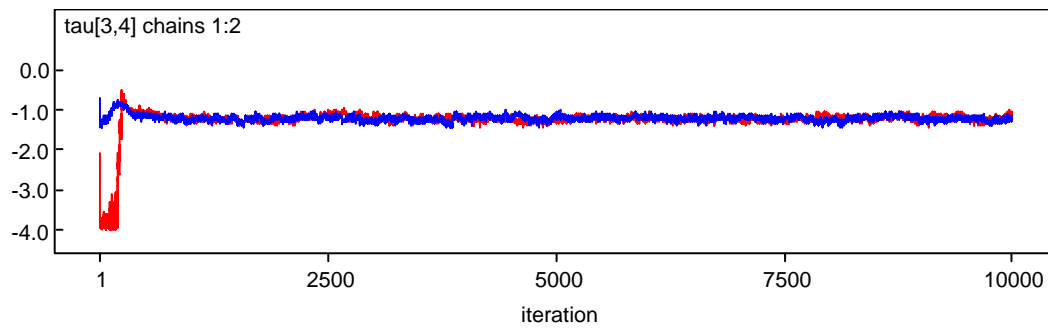
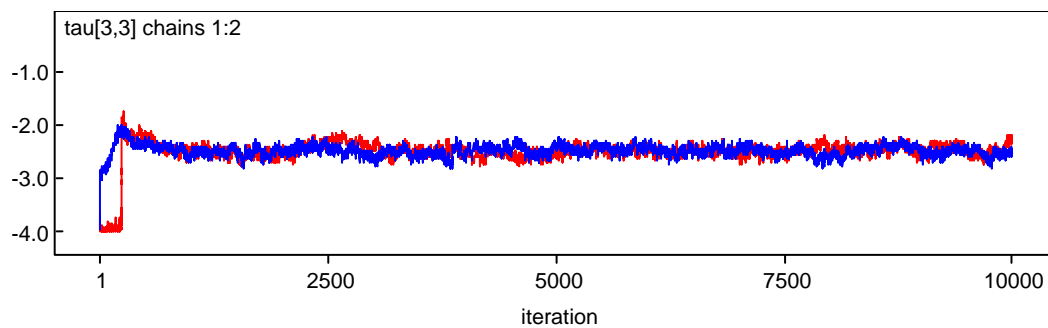
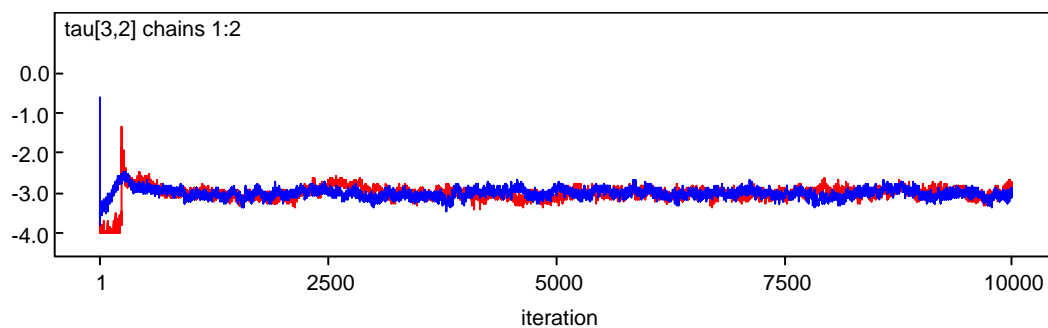
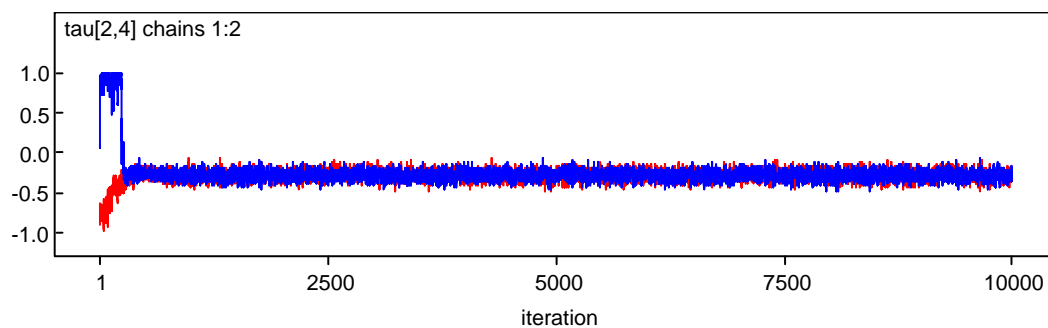
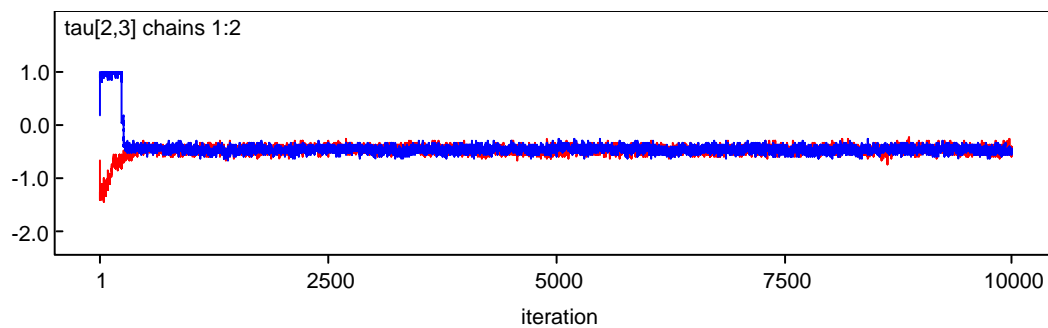


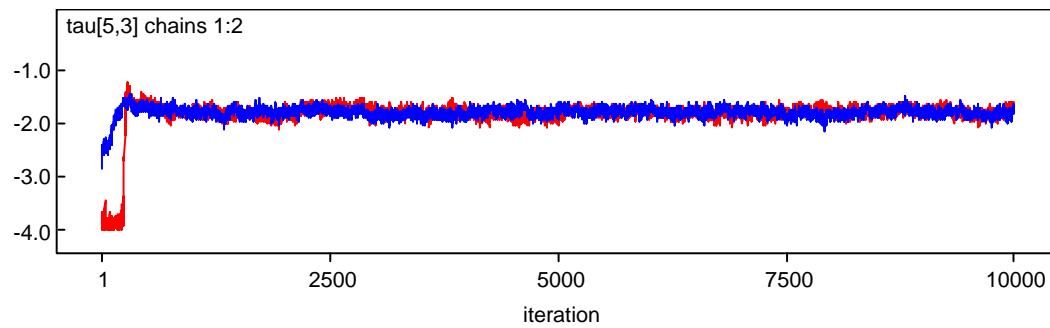
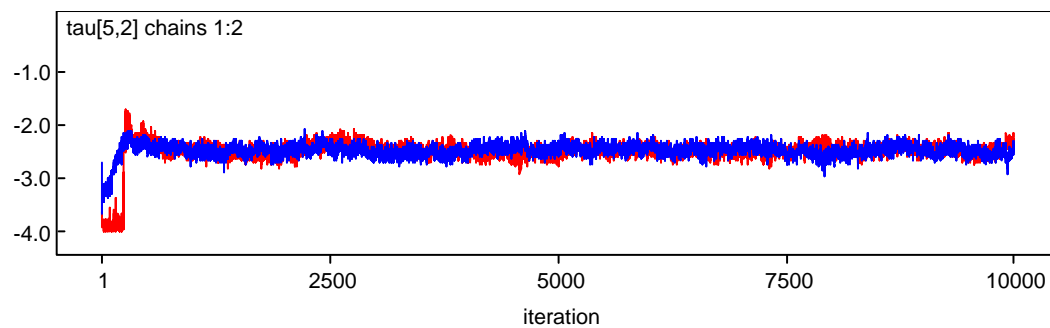
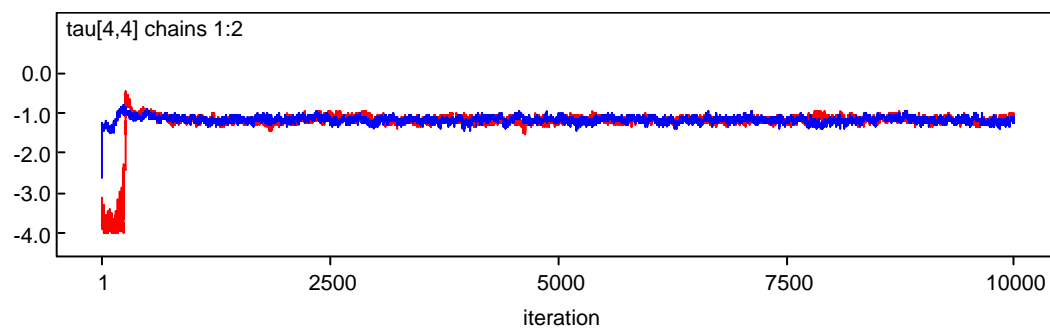
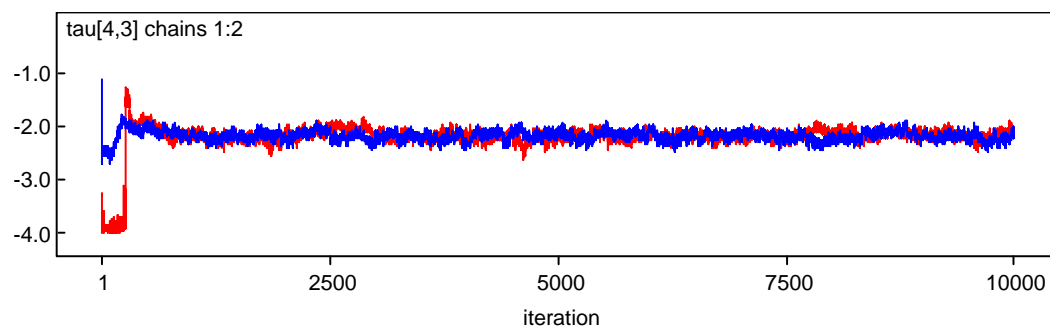
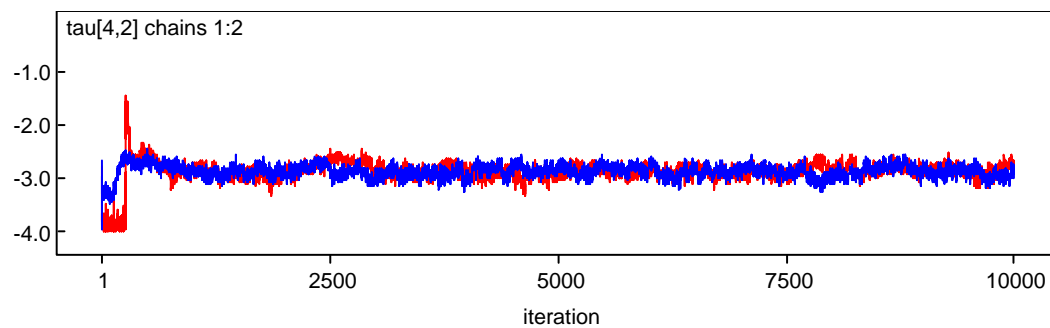


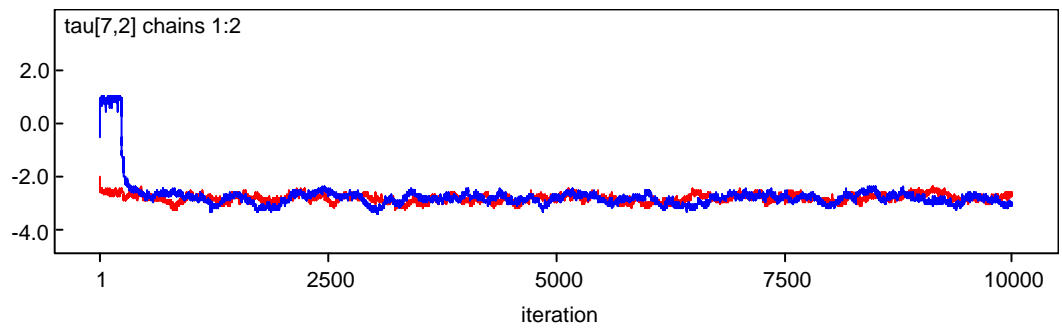
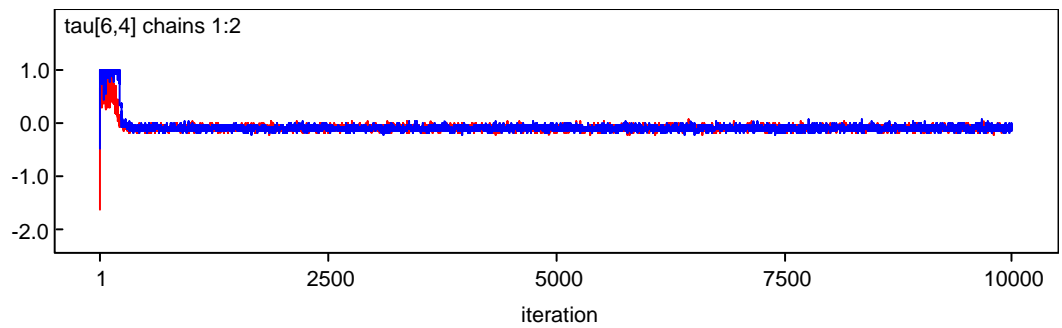
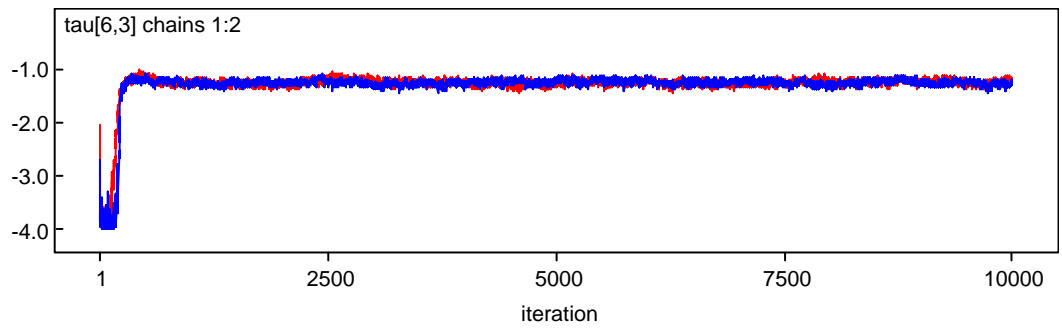
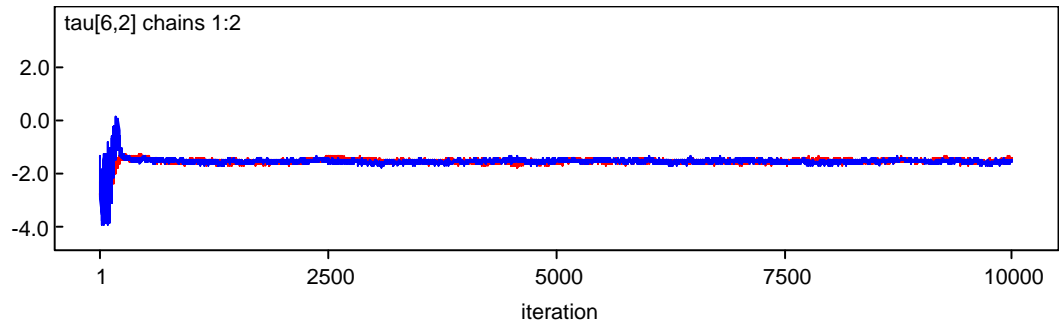
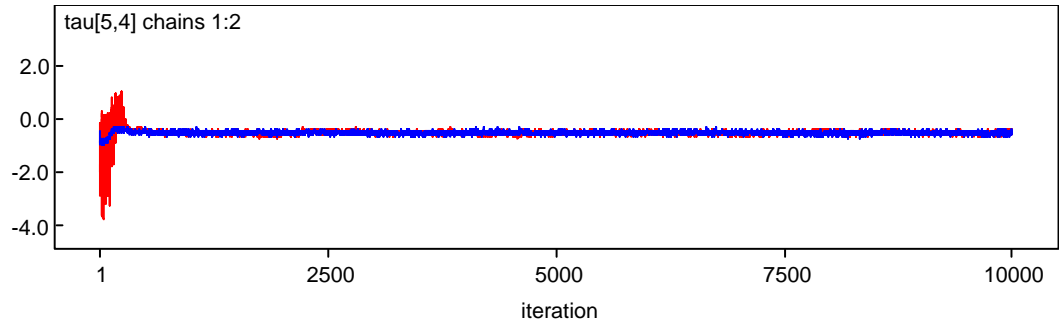


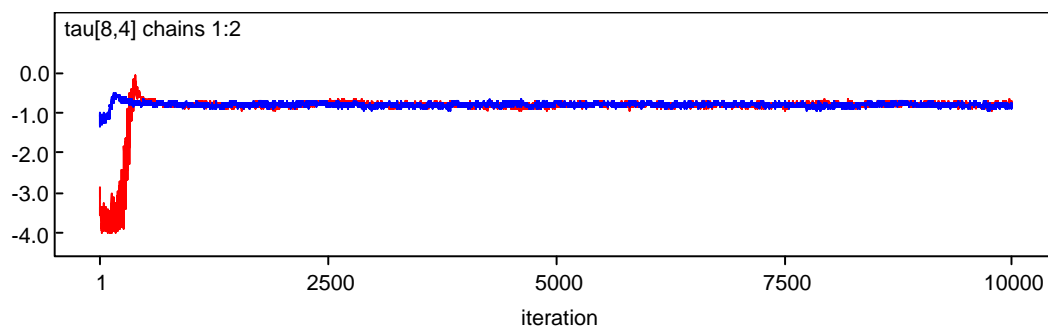
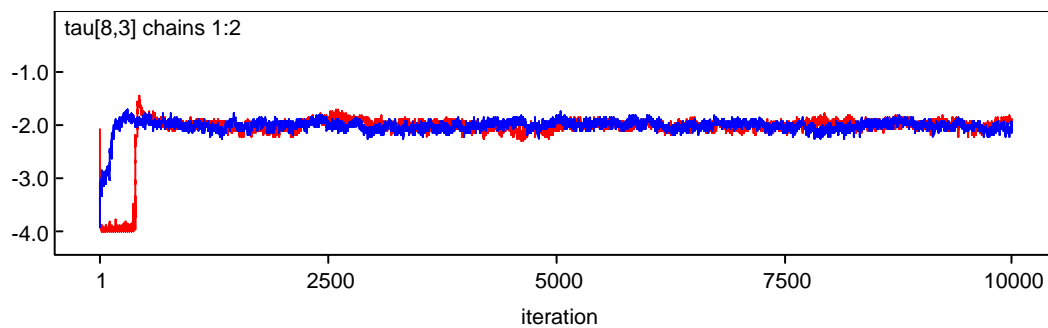
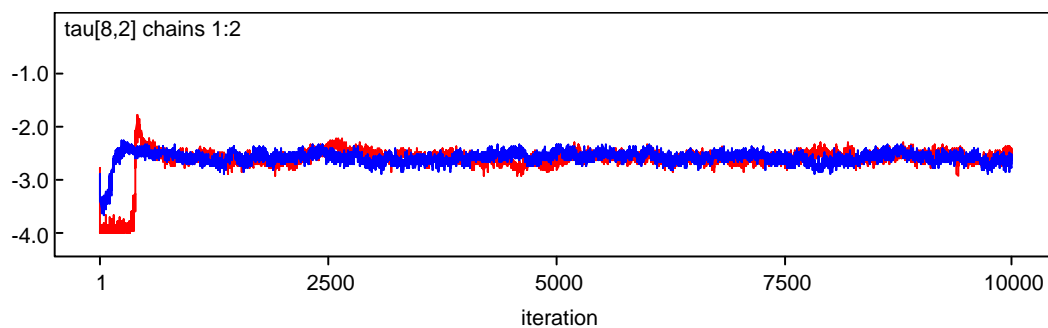
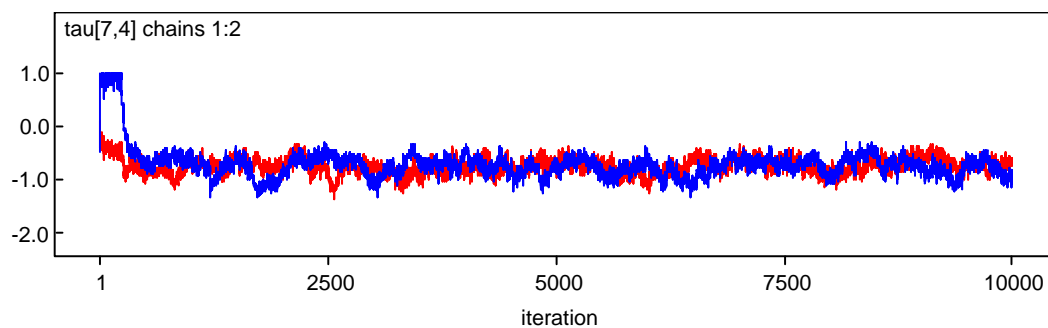
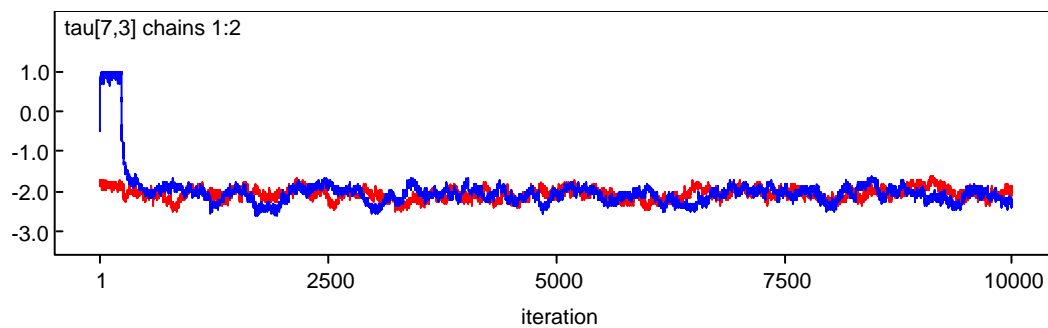


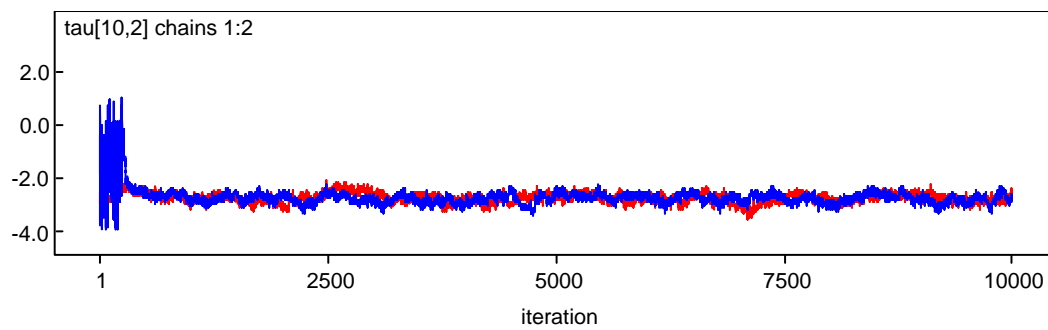
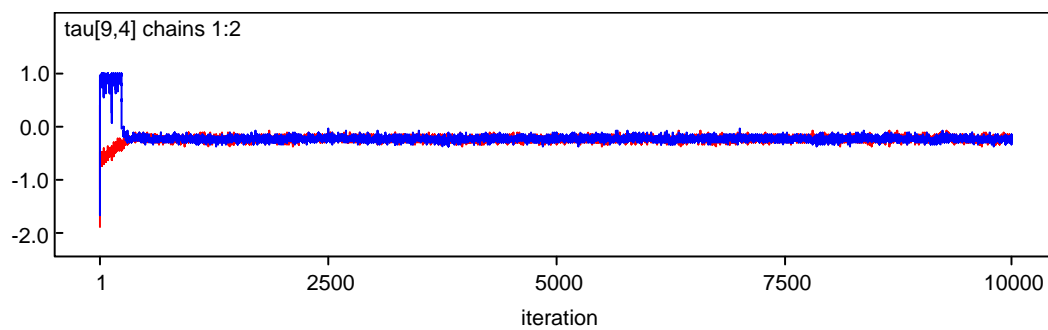
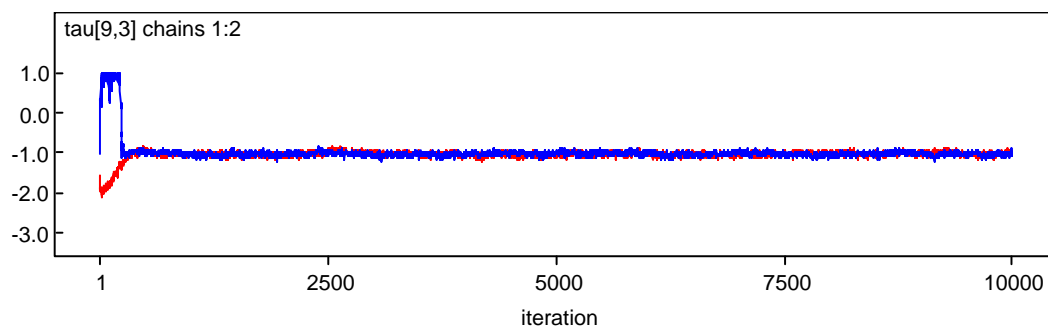
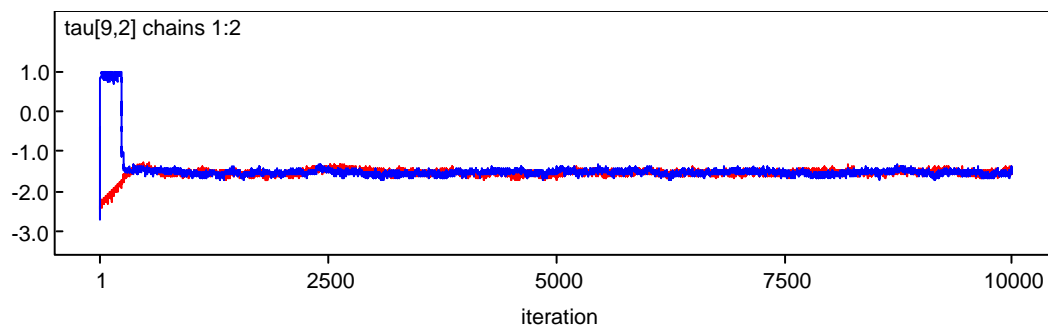


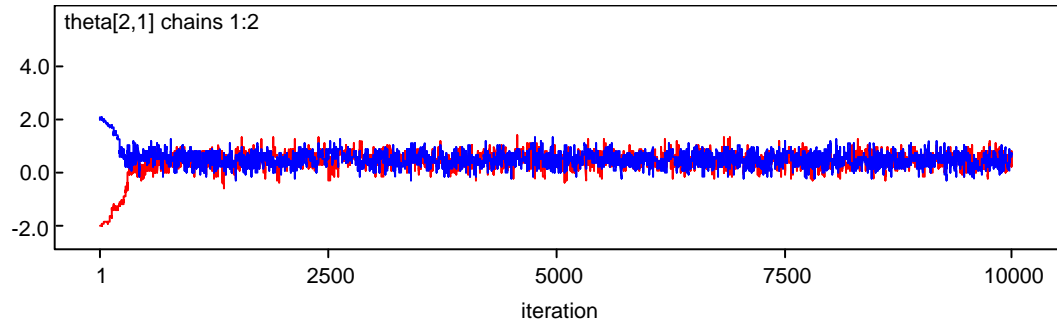
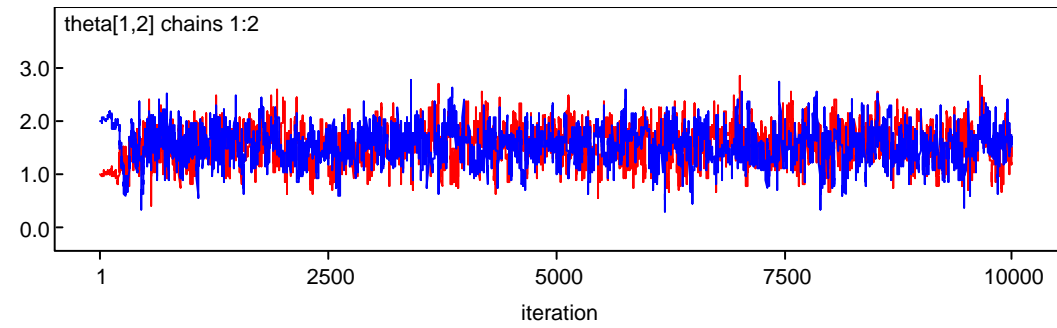
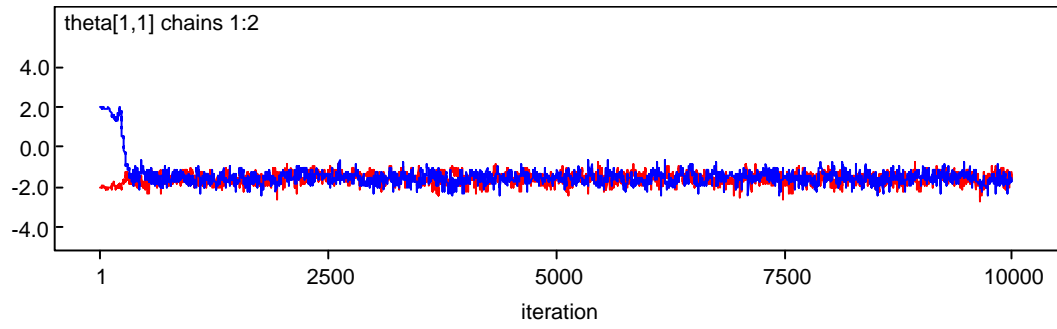
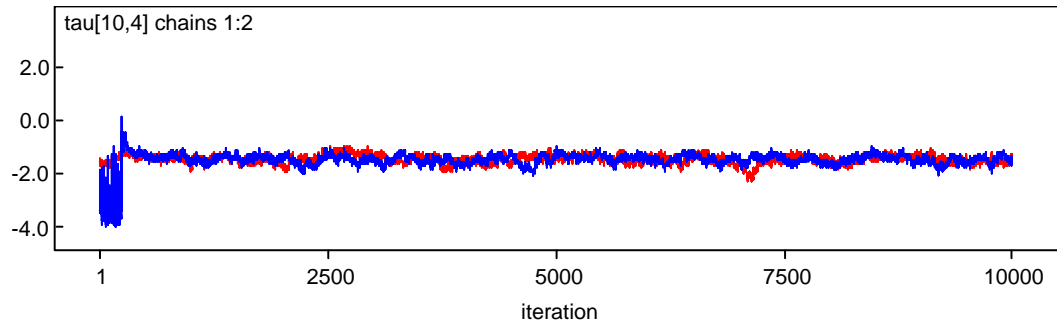
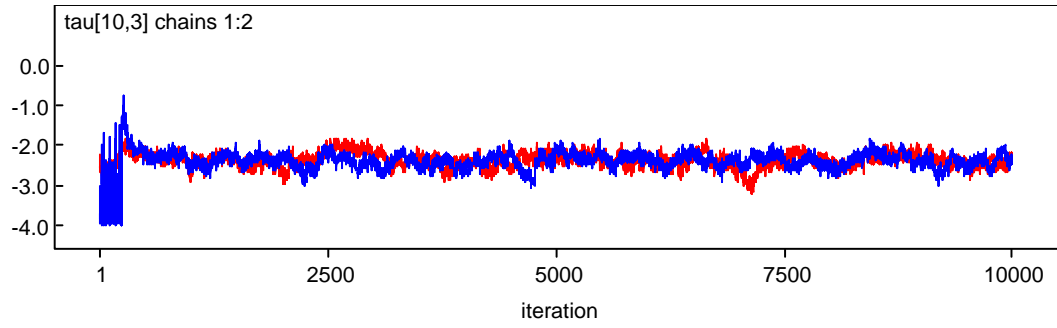


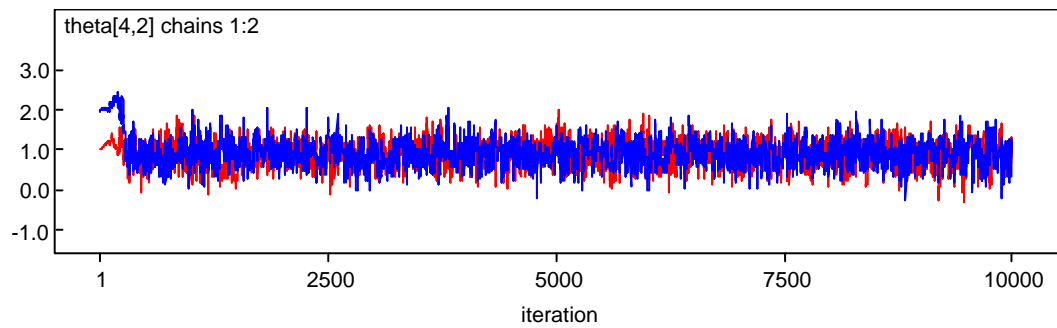
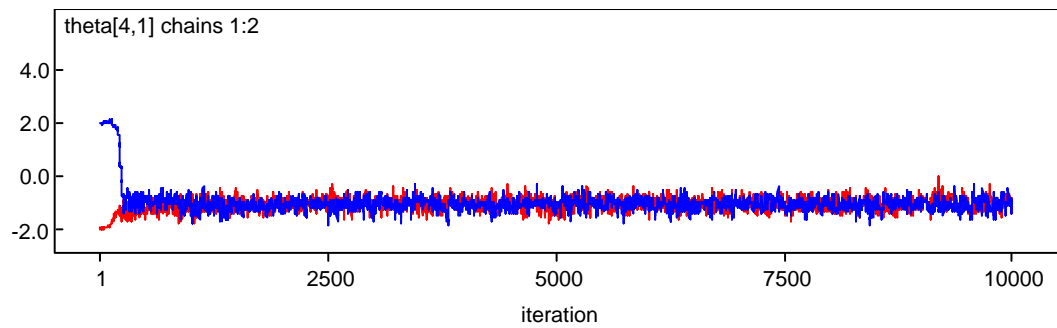
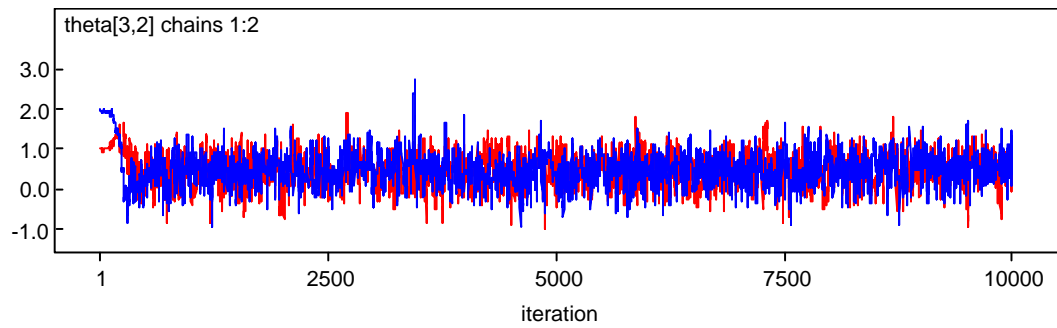
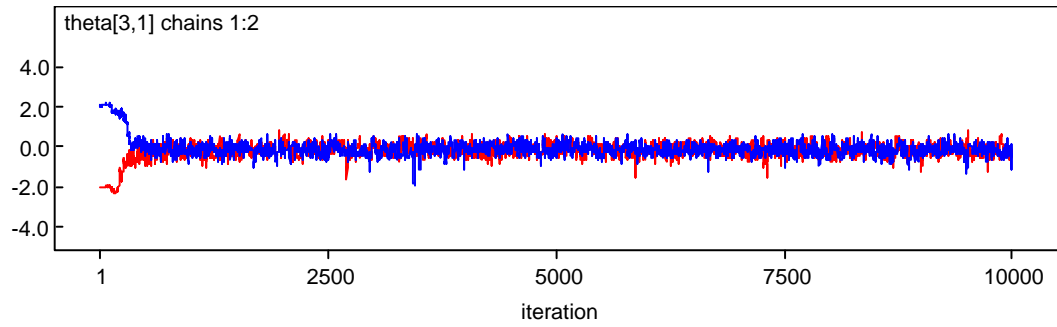
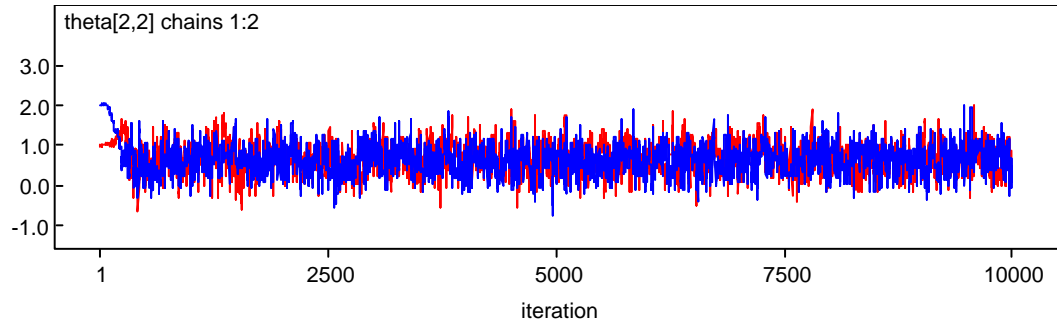


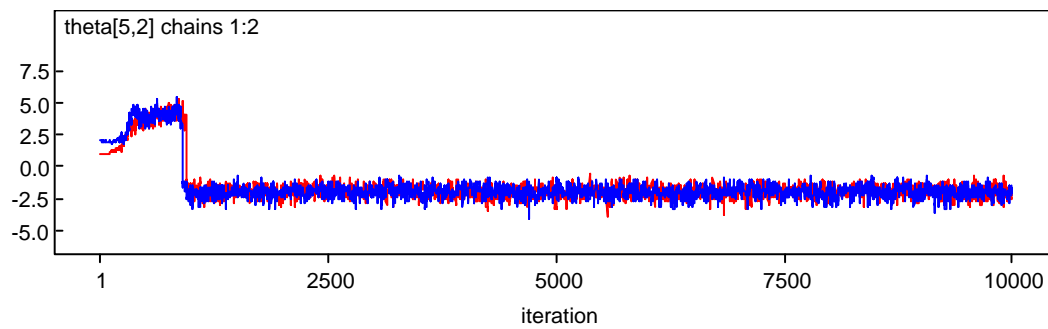
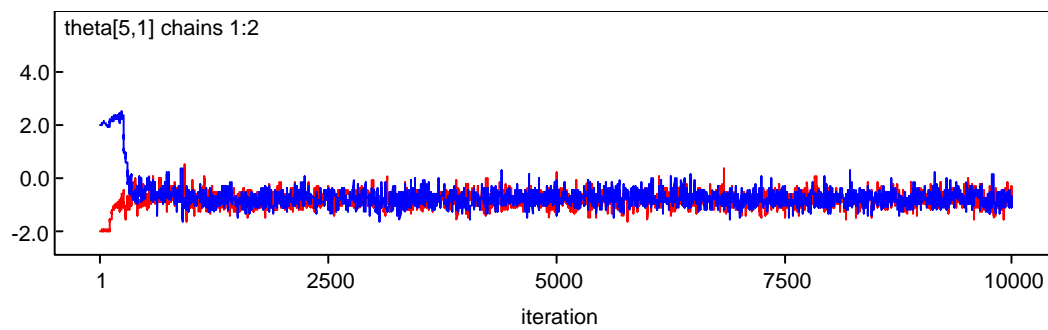












References:

- Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001), *The NAEP 1998 technical report*. Washington, DC: National Center for Education Statistics.
- Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, 50, 3-16.
- Andrich, D. (1988). The application of an unfolding model of the PIRT type to the measurement of attitude. *Applied Psychological Measurement*, 12, 33-51.
- Andrich, D. (1996). A general hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology*, 49, 347-365.
- Andrich, D., de Jong, J. H. A., & Sheridan, B. E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In J. Rost & R. Langeheine (Eds), *Application of latent trait and latent class models in the social science* (pp. 58-68). Munster, Germany: Waxmann Verlag.
- Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single stimulus responses. *Applied Psychological Measurement*, 17, 253-276.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (ED.), *Educational Measurement (2nd ed.)* Washington, DC: American Council on Education.
- Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. . In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison: University of Wisconsin Press.

- Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice*. San Francisco: Holden Day.
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change” – or should we? *Psychological Bulletin*, *74*, 68-80.
- Cui, W., Roberts, J. S., & Bao, H. (2004). *Data demands for the Generalized Graded Unfolding Model*, Poster presented at the annual conference of National Council on Measurement in Education, San Diego, CA, April, 2004.
- Dobois, B., & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*. *35*, 869-884.
- Donoghue, 1999. Establishing two important properties of two IRT-based models for unfolding data. Unpublished manuscript.
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495-515.
- Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, *27*, 3-26.
- Fischer, G. H., & Pazer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika*, *56*, 637-651.
- Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *57*, 177-192.
- Garside, R. F. (1956). The regression of gains upon initial scores. *Psychometrika*, *21*, 67-77.

- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*. CRC Press, Boca Raton, 2nd edition.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices*. Second edition. New York: Springer.
- Kim, S., Cohen, A. S., Baker, F. B., Subkoviak, K. J., & Leonard, T. (1994). An investigation of hierarchical Bayes procedures in item response theory. *Psychometrika*, 59,405-421.
- Likert, R. (1932). *A technique for measurement of attitudes*. New York: 1932.
- Lord, F. M. (1962). Test reliability – a correction. *Educational and Psychological Measurement*, 22, 511-512.
- Lord, F. M. (1963). Elementary models for measuring change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 21-38). Madison: University of Wisconsin Press.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
- Master, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159-176.
- Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. Chicago: Scientific Software International.
- Reckase, M. D., & Martineau, J. (2004). *The vertical scaling of science achievement tests*. Paper commissioned by the Committee on Test Design for K-12 Science

Achievement, Center for Education, National Research Council. Washington, DC:
National Research Council.

- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP Parameter Estimates in the Generalized Graded Unfolding Model. *Applied psychological Measurement, 26*, 192-207.
- Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.
- Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211-233.
- Roberts, J. S., Lin, Y., & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement, 25*, 177-196.
- Roberts, J. S., & Ma, Qianli (2006). IRT models for the assessment of change across repeated measurements. In R. Lissitz (Ed.), *Longitudinal and Value Added Modeling of Student Performance*. Maple Grove, MN: JAM Press.
- Rogosa, D., & Willett, J. B. (1983). Demonstrating the reliability the difference score in the measurement of change. *Journal of Educational Measurement, 20*, 335-343.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual*,

version 1.4. Cambridge: MRC Biostatistics Unit.

<http://www.mrc-bsu.cam.ac.uk/bugs>

te Marvelde, J.M., & Glas, C.A.W. (2006). Application of Multidimensional Item

Response Theory Models to Longitudinal Data. *Educational and Psychological Measurement, 66*, 5-34.

Thorndike, E. L. (1924). The influence of the chance imperfections of measures upon the relation of initial score to gain or loss. *Journal of Experimental Psychology, 13*, 7225-7232.

Thurston, L. L. (1927). The Method of Paired Comparisons for Social Values. *Journal of Abnormal and Social Psychology, 21*, 384-400.

Thurston, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology, 33*, 529-554.

van Schuur, W. H., & Kiers, H. A. L. (1994). Why factor analysis is often the incorrect model for analyzing bipolar concepts, and what model can be used instead. *Applied Psychological Measurement, 18*, 97-110.

Wang, W., & Chyi-In, W. (2004). Gain score in item response theory as an effect size measure. *Educational and Psychological Measurement, 5*, 758-780.

Wang, W., Wilson, M., & Adams, R. J. (1998). Measuring individual differences in change with multidimensional Rasch models. *Journal of Outcome Measurement, 240-265*

Wilder, J. (1967). *Stimulus and response: The Law of Initial Value*. Bristol: Wright.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain score obsolete? *Applied Psychological Measurement, 20*, 59-69.

- Williams, R. H., & Zimmerman, D. W. (1999). Nonindependence of parameters of the validity and reliability of gain scores. *Perceptual and Motor Skills, 88*, 679-681.
- Zimmerman, D. W., & Williams, R. H. (1982). Gain score in research can be highly reliable. *Journal of Educational Measurement, 19*, 149-154.
- Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology, 51*, 343-351.