

**A CONJOINT UNFOLDING IRT MODEL FRAMEWORK FOR THE
ANALYSIS OF PREFERENCE AND RESPONSE TIME DATA**

A Dissertation
Presented to
The Academic Faculty

by

Jordan Lea Sparks

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology
December 2023

**A CONJOINT UNFOLDING IRT MODEL FRAMEWORK FOR THE
ANALYSIS OF PREFERENCE AND RESPONSE TIME DATA**

Approved by:

Dr. James S. Roberts, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Susan Embretson
School of Psychology
Georgia Institute of Technology

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Dr. Daniel Spieler
School of Psychology
Georgia Institute of Technology

Dr. John Donoghue
ETS

Date Approved: October 4, 2023

TABLE OF CONTENTS

LIST OF TABLES	v
LIST OF FIGURES	vi
SUMMARY	vii
CHAPTER 1. Introduction	1
1.1 Background (Unfolding Item Response Theory)	1
1.2 Generalized Graded Unfolding Model (GGUM)	3
1.2.1 Estimation of GGUM Parameters	4
1.2.2 Collateral Information for Parameter Estimation	11
1.3 Analysis of Response Time Data in an IRT Framework	11
1.4 Research Hypothesis	16
CHAPTER 2. Method	18
2.1 Modifications to the GGUM	18
2.2 Parameter Recovery Simulation	20
2.2.1 Experimental Design	20
2.2.2 Data Generation	20
2.2.3 Parameter Estimation	22
2.3 Real Data Application	29
Chapter 3. Results	31
3.1 GGUM Parameter Recovery Simulation	31
3.1.1 Accuracy of the Item Parameter Estimates	35
3.1.2 Accuracy of the Person Parameter Estimates	39
3.2 GGUM-RT Parameter Recovery Simulation	40
3.2.1 Accuracy of the Item Parameter Estimates	44
3.2.2 Accuracy of the Person Parameter Estimates	47
3.2.3 Accuracy of Parameter Estimates in the Response Time Regression Model	47
3.2.4 Effect Sizes from the Response Time Regression Model	48
3.3.1 Estimation Accuracy Between the GGUM and the GGUM-RT with RMSD Outcomes	52
3.3.2 Estimation Accuracy Between the GGUM and the GGUM-RT with SE Outcomes	54
3.4 Empirical Data Application	60
3.4.1 Analysis	65
Chapter 4. Discussion	68
4.1 Final Thoughts on the Comparison between the GGUM-RT and GGUM	68
4.2 Pros and Cons	68
4.3 Scientific Importance of the Current Study	69
4.4 Future Directions	70

APPENDIX A. GUN CONTROL ATTITUDE STATEMENTS	72
APPENDIX B. ITEM CHARACTERISTIC CURVES (ICCs)	75
REFERENCES	84

LIST OF TABLES

Table 1	Mean RMSD of Parameter Estimates	32
Table 2	η^2 Values for ANOVA Effects	32
Table 3	RMSD Components of GGUM Parameter Estimates from the GGUM	33
Table 4	η^2 Values for ANOVA Effects with RMSD Components of GGUM Parameter Outcomes	34
Table 5	Mean RMSD of Parameter Estimates	41
Table 6	η^2 Values for ANOVA Effects	41
Table 7	RMSD Components of GGUM Parameter Estimates from the GGUM-RT	42
Table 8	η^2 Values for ANOVA Effects with RMSD Components of GGUM-RT Parameter Outcomes	43
Table 9	Mean R^2 Across Simulation Conditions	49
Table 10	Count of Smaller RMSD Values by Model Type	50
Table 11	η^2 Within-Family Values for Within-Subjects ANOVA Effects with RMSD Outcomes	51
Table 12	Mean SE of Parameter Estimates	55
Table 13	η^2 Within-Family Values for Within-Subjects ANOVA effects with SE Outcomes	55
Table 14	GGUM-RT Item Parameter Estimates of Gun Control Attitude Statements	61
Table 15	GGUM-RT Polynomial Regression Parameter Estimates	63
Table 16	Mean SE of Real Data Parameter Estimates	64
Table 17	η^2 Values for Real Data Effects	64

LIST OF FIGURES

Figure 1	Expected relationship of response time with GGUM parameters.	19
Figure 2	Interaction plot of the effect of test length and the number of response categories on item location parameter accuracy.	37
Figure 3	Interaction plot of the effect of test length and the number of response categories on item threshold parameter accuracy.	39
Figure 4	Interaction plot of the effect of model type and test length on person parameter accuracy.	45
Figure 5	Interaction plot of the effect of model type and test length on person parameter accuracy.	53
Figure 6	Interaction plot of the effect of model type and test length on item discrimination parameter accuracy.	54
Figure 7	Interaction plot of the effect of model type and test length on average item discrimination parameter standard error.	57
Figure 8	Interaction plot of the effect of model type and response categories on average item location parameter standard error.	58
Figure 9	Interaction plot of the effect of model type and response categories on average person parameter standard error.	59
Figure 10	Interaction plot of the effect of model type and sample size on average person parameter standard error.	60
Figure 11	Estimated item locations from the GGUM- RT.	65
Figure 12	Relationship of response time and the difference between estimated person and item location parameters.	66

SUMMARY

The Generalized Graded Unfolding Model (GGUM) is an unfolding item response theory model that produces single-peaked, nonmonotonic item characteristic curves consistent with a proximity-based response process. The model can be applied to binary or graded item responses, or a mixture of the two. This paper proposes a modification to the GGUM estimation procedure, referred to as the Generalized Graded Unfolding Model with Response Times (GGUM-RT), which includes response time as collateral information in estimating model parameters. The inclusion of response time data illustrates how (a proxy for) cognitive processing relates to the latitude of acceptance construct from social judgment theory. It is also demonstrated to improve the precision of model parameter estimates obtained from the standard GGUM.

INTRODUCTION

1.1 Background (Unfolding Item Response Theory)

The principle of an ideal point response process operates under the assumption that a person will prefer a stimulus (i.e., an item) to the extent that it is located near the individual's position on a latent continuum (or more generally, in a latent space). As the distance between a person and an item location increases, the probability that the person prefers the stimulus decreases. This principle is rooted in the measurement of attitudes, and its origin can be traced back to Thurstone's (1928) approach to collecting agree-disagree responses to statements to estimate a person's attitude. In this context, an individual is expected to agree with a statement on an attitude questionnaire to the extent that it is located near the individual's position on an underlying latent attitude continuum. As the distance between the statement and the individual increases, the probability of the individual agreeing with the statement declines. While it was not formally stated in Thurstone's (1928) work, his method of analyzing responses operates by assuming that responses to attitude statements result from an ideal point process (Coombs, 1964). This unique form of response data is appropriately analyzed using unfolding models, where the probability of agreement with an item is calculated as a function of the distance between a person and an item on the underlying continuum.

The unfolding model proposed by Coombs (1964), also referred to as a "distance model" for preferences, operates as if a person arrives at a response to an item by

comparing the location of the item in a latent space to their own location (i.e., their ideal point) in the same space:

$$\delta_{ij} = f(d_{ij}) = f \left(\left[\sum_k (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}} \right) \quad (1)$$

where

δ_{ij} = the strength of person j 's preference for item i ,

d_{ij} = the distance between points i and j ,

x_{ik} = the location of item i along dimension k ,

x_{jk} = the ideal point coordinate for person j along dimension k ,

and $f(d_{ij})$ is a monotonic transformation of distance.

Unfolding response processes are most famously illustrated by Carroll's (1972) tepid tea example. In his example, he discusses an experiment requiring preference judgments for several cups of tea. Each cup of tea was described along two dimensions: the temperature of the tea (ice cold, cold, lukewarm, hot, steaming hot) and the amount of sugar in the tea (no sugar, half a teaspoon, one teaspoon). Subjects presumably rated the degree to which they preferred each cup of tea using their own personal preference (their "ideal point" along the two dimensions). For example, a subject with a preference for half a teaspoon of sugar in their tea might rate the cups of tea with no sugar or one teaspoon of sugar poorly—in one instance, the tea would not be sweet enough for their liking, and in the other, it would be too sweet. Responses also varied depending on the temperature of the tea: a cup of tea with an ideal amount of sugar with a less than ideal temperature would be rated lower than a cup of tea closer to a subject's ideal along both dimensions. This

example demonstrates that preferences can be modeled by considering an individual's ideal location in a latent space relative to the locations of stimuli.

Over the last 50 years, unfolding models have been applied to a number of different fields in psychology, including an extension to multidimensional scaling known as multidimensional unfolding (Kruskal & Carroll, 1969; Carroll, 1972; Davison, 1983), and unfolding item response theory (Van Schuur, 1984; Cliff, Collins, Zatzkin, Gallipeau, & McCormick, 1988; Bartholomew, 1992; Post, 1992; Andrich & Luo, 1993; Hoijtink & Molenaar, 1994; Roberts, 1995; Andrich, 1996; Roberts & Laughlin, 1996; Roberts, Donoghue, & Laughlin, 2000). These models follow similar principles posited by Thurstone (1928) and Coombs (1964), and vary both in terms of their (non)parametric form as well as the types of response data they are designed to analyze. The most popular unfolding model in present day is the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), a parametric item response model that allows for the analysis of either binary or graded responses. It is the subject of the next section.

1.2 Generalized Graded Unfolding Model (GGUM)

The generalized graded unfolding model (GGUM; Roberts et al., 2000) produces single-peaked, nonmonotonic item characteristic curves consistent with a proximity-based (i.e., ideal point) response process. The GGUM has been applied to the measurement of a number of different topics, including smoking cessation, attitudes, emotions, physical attraction, and personality (Noel, 1999; Roberts et al., 2000; Roberts & Sparks, 2015; Roberts, Barrett, & King, 2016; Crowe, Sleep, Carter, Campbell, & Miller, 2018). Additionally, the GGUM has also been applied to personality measurement and the

measurement of developmental processes (DeMars & Erwin, 2003; de la Torre, Stark, & Chernyshenko, 2006; Carter et al., 2014).

The GGUM is defined by its category probability function, which is defined as:

$$P(Z_i = z | \theta_j) = \frac{\exp(\alpha_i [z(\theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik}]) + \exp(\alpha_i [(M - z)(\theta_j - \delta_i) - \sum_{k=0}^z \tau_{ik}])}{\sum_{w=0}^C [\exp(\alpha_i [w(\theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik}]) + \exp(\alpha_i [(M - w)(\theta_j - \delta_i) - \sum_{k=0}^w \tau_{ik}])]} \quad (2)$$

where

Z_i = an observable response to the i^{th} item,

$z = 0, 1, 2, \dots, C$; $z = 0$ corresponds to the strongest level of disagreement and $z = C$ refers to the strongest level of agreement,

C = the number of observable response categories minus 1,

$M = 2C + 1$ = the number of subjective response category (SRC) thresholds,

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

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

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

τ_{ik} = the k^{th} subject response category (SRC) threshold for the i^{th} item.

The value of τ_{i0} is defined as zero, and the remaining SRC thresholds are assumed to be symmetric about the item points, $\theta_j - \delta_i = 0$. These thresholds are not forced to be ordered and are not constrained to be constant across items.

1.2.1 Estimation of GGUM Parameters

Estimation procedures for item response theory model parameters can be categorized in three ways: procedures that estimate person parameters, procedures that estimate item parameters, and procedures that simultaneously estimate both. The most widely used estimation procedures for item response theory model item parameters are marginal maximum likelihood (MML; Bock & Aitkin, 1981) and the Bayesian estimation procedure marginal maximum a posteriori (MMAP) estimation (Mislevy, 1986). These item estimation procedures are commonly paired alongside expected a posteriori (EAP) estimation (Bock & Mislevy, 1982) to estimate person parameters. Other procedures that jointly estimate both person and item parameters include joint maximum likelihood (JML; Lord, 1974, 1980) and Markov-chain Monte Carlo (MCMC) methods (de la Torre et al., 2006). These procedures, along with several others, have been used to obtain estimates for GGUM parameters over the past 25 years. The first published version of the GGUM, a constrained version known as the graded unfolding model (GUM; Roberts, 1995; Roberts & Laughlin, 1996), used a joint maximum likelihood procedure (Lord, 1974, 1980) combined with a grid search algorithm to estimate both the item and person parameters. Since that time, GGUM parameters have been estimated using MML/EAP (Roberts, Donoghue, & Laughlin, 2002), MMAP/EAP (Thompson, 2014), genetic algorithms (GA)/EAP (Williams, 2017), Metropolis-Hastings Robbins-Monro (MHRM)/EAP (King, 2017) and MCMC (Roberts & Shim, 2010).

1.2.1.1 Joint Maximum Likelihood Estimation

Joint maximum likelihood (JML) is an iterative procedure that estimates item and person parameters simultaneously (Lord, 1974, 1980). The algorithm proceeds in two alternating stages. In the first stage, the item parameters are treated as fixed and known,

and then maximum likelihood estimates for the person parameters are found. Then, in the second stage, the item parameters are estimated by using the estimated person parameters as fixed and known values. Starting values for item parameters must be supplied on the very first iteration, and these are generally derived from the item responses without the use of an elaborate method. The process of estimating person parameters while treating item parameters as fixed and then estimating item parameters while treating person parameters as fixed continues until there is little change in item parameters from one iteration to the next.

The JML procedure limited the applications of the GUM, as simulation studies demonstrated that parameter estimates typically did not converge to the true values as the sample size increased (Roberts & Laughlin, 1996). This finding was in line with JML item parameter estimate inconsistencies seen in other IRT models (Andersen, 1973; Wright & Douglas, 1977; Swaminathan & Gifford, 1983). The algorithm also frequently struggled with converging on “local maxima” in the likelihood function – scenarios where the likelihood function has not truly been maximized, as the estimator has gotten “stuck” on a smaller peak rather than the absolute peak. This implied that it was less likely the parameters estimated were close to truth, leading to inaccurate and misleading results.

1.2.1.2 Marginal Maximum Likelihood/Expected A Priori Estimation

The full GGUM was estimated by Roberts et al. (2000) using an MML approach to estimate item parameters and an EAP procedure to estimate the person parameters. Marginal maximum likelihood integrates the person parameters out of the likelihood function using the prior probability density function and solves for item parameters in the

resulting margins of the likelihood function (Bock & Aitkin, 1981). The general marginal likelihood function takes the following form:

$$L_i = \prod_{j=1}^J \left[\int (P(Z_i = x_{ij} | \theta_j) g(\theta_j | \mu, \sigma^2)) d\theta_j \right] \quad (3)$$

where

x_{ij} is an observable response to the i th item for the j th individual,

$P(Z_i = x_{ij} | \theta_j)$ indicates the model probability of a particular response to item i ,

$g(\theta_j | \mu, \sigma^2)$ is an arbitrary prior distribution for θ_j with the population mean μ and the population variance σ^2 .

This method does not typically suffer from the same problem with inconsistent item parameter estimates associated with JML procedures, and perhaps most importantly, it greatly makes the local maxima in the likelihood function with respect to person parameters moot during the item parameter estimation. With the item parameter estimate determined, the estimates of the person parameters are then obtained by using EAP estimation. This method is derived from Bayesian statistical principles and refers to the expected value of the posterior probability distributions of person parameters. The posterior probability distributions are obtained with the equation:

$$P(\theta | \mathbf{X}_j) = \frac{L(\mathbf{X}_j | \theta) g(\theta | \mu, \sigma^2)}{P(\mathbf{X}_j)} \quad (4)$$

where $L(\mathbf{X}_j | \theta)$ is the likelihood of the response vector for individual j . This method requires the specification of prior distributions for person parameters which is usually the

same as that used in the MML procedure but is sometimes derived empirically. The mean of the posterior distribution is the EAP estimate for a given person, and thus, this technique makes the presence of any local maxima in the posterior of a particular theta parameter moot with regard to estimation.

The advantage of applying this MML/EAP approach to the GGUM over JML parameter estimation was more accurate results, particularly when large sample sizes (e.g., 750) were available for datasets consisting of approximately twenty 6-point items (Roberts et al., 2002). Further, parameter estimates were found to be accurate with several different prior distributions specified for the person parameters, regardless of whether the true distribution matched the prior distribution.

1.2.1.3 Marginal Maximum A Posteriori/Expected A Priori Estimation

The parameters for the GGUM have also been estimated with a marginal maximum a posteriori (MMAP) and EAP approach (Roberts & Thompson, 2011). The Bayesian MMAP procedure, developed by Mislevy (1986), has conceptual similarities to MML in that the person parameters are integrated out of the likelihood function. Prior distributions are specified for each item parameter and these distributions provide a second source of information to produce a posterior distribution. This approach yielded item parameter estimates that were typically more accurate than estimates produced by the MML procedure. The superior accuracy of MMAP as compared to MML is noticeable only when the number of response categories is 4 or less (Roberts & Thompson, 2011).

1.2.1.4 Markov chain Monte Carlo estimation

Another popular approach to GGUM parameter estimation is the Markov chain Monte Carlo (MCMC) approach (de la Torre et al., 2006; Roberts & Shim, 2010; Roberts & Thompson, 2011). MCMC procedures are a class of fully Bayesian methods used to estimate model parameters through repeated sampling from a joint posterior distribution (Patz & Junker, 1999). Although this joint distribution may be extremely complex, some MCMC algorithms can sample from far more simple distributions and systematically reject some draws in order to ultimately reflect a sample from the joint posterior distribution. The flexibility and accessibility that the MCMC approach offers for estimation has brought increasing attention to utilizing this method to estimate models across the psychometric literature (Fox, 2010).

Within the context of IRT parameter estimation, the general MCMC estimation procedure begins with the specification of prior distributions and starting values for the model parameters. Then, computational software constructs a Markov chain from an MCMC algorithm like the Metropolis-Hastings sampler (Hastings, 1970), the Gibbs sampler (Geman & Geman, 1984), or the Metropolis Hastings within Gibbs sampler (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). Parameter values are then repeatedly sampled (or drawn) from a distribution that eventually converges to the joint posterior distribution of model parameters after a large number of repetitions. It is common for researchers to discard a number of the initial draws of the procedure prior to convergence, referred to as “burn-in” states. Once convergence to this stationary distribution is achieved, the means of further generated samples for each parameter can be used as parameter estimates (i.e., expected a posteriori estimates).

MCMC as an estimation approach has grown more prevalent in psychometrics literature due to the accessibility of several open-source estimation programs, including OpenBUGS, JAGS, and STAN (Lunn, Thomas, Best, & Spiegelhalter, 2000; Plummer, 2003; Carpenter et al., 2017). However, despite MCMC estimation's increasing popularity, the approach can be time consuming: this estimation approach can take anywhere from minutes, to days, to upwards of a week to converge on a solution depending on the complexity of the model specified, the number of items, the number of response categories for each item, and the number of respondents (i.e., the total number of parameters estimated) in addition to the computer software and hardware choices that are implemented.

de la Torre et al. (2006) found that an MCMC approach produced estimates and standard errors that were typically more accurate than MML in cases of extreme GGUM items with few response categories, though a degradation in the accuracy of parameter estimates was still present. In a more recent study, Roberts and Thompson (2011) also found that MCMC produced results that were more accurate than MML estimation, particularly when the number of response categories was small (i.e., two or three); however, they posited that MMAP/EAP was a superior approach, as they found that the MMAP procedure generally produced more accurate estimates than MCMC. The MMAP procedure took seconds to converge relative to the hours typically necessary to reach a solution using GGUM item responses with OpenBUGS. Despite the substantial differences in convergence speed, MCMC is still commonly used.

1.2.2 Collateral Information for Parameter Estimation

Traditionally, item response theory models, including both cumulative and unfolding models, are estimated using information obtained from item responses, starting values and prior distributions for model parameters (if required by the estimation procedure). Collateral sources of information, including demographic information and response time data (Kuncel, 1973; Mislevy & Sheehan, 1989; Roskam, 1997; van der Linden, Entink, & Fox, 2010) have been proposed to assist in inferences made about model parameters and to improve the accuracy of parameter estimation in these models. This additional information has been modelled both separately and simultaneously alongside IRT parameters using hierarchical processes (van der Linden et al., 2010), simultaneous series of equations (Kuncel, 1973), and mixed IRT models (Smit, Kelderman, & van der Flier, 1999). While many types of external data may be of interest to researchers for inclusion in their item response theory models, response time data is an intuitive source of collateral information for psychometricians to leverage in many cases as it is typically an easily obtainable variable when data collection is performed via computers.

1.3 Analysis of Response Time Data in an IRT Framework

The inclusion of response time data in item response theory models is a relatively new concept in the field of psychometrics. These two forms of information were initially considered to be two different constructs that were typically modelled separately (Scheiblechner, 1979). To this end, there is a lot of discrepancy in the literature around how to best incorporate response time in item response theory models. The relationship between IRT parameters and response time parameters has been measured using several

different joint modeling methods (De Boeck & Jeon, 2019). Many of these methods have treated IRT parameters as separate from, but related to, response time parameters. For example, van der Linden et al. (2010) proposed a hierarchical approach, where a multivariate distribution for response time parameters and response parameters are assumed to be related, but not identical. Other approaches have included diffusion models (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011) for cognitive test taking and race models: process models that capture cognitive processes “racing” toward a finish line (Rouder, Province, Morey, Gomez, & Heathkote, 2015). In a different arm of the quickly growing literature on this topic, other researchers have followed the assumption that the same underlying construct (e.g., the signed distance between a person and an item along a latent continuum) drives the change in both response time parameters and response parameters (Ferrando & Lorenzo-Sava, 2007).

The presumed distribution of response times has also varied across the literature according to the type of data used and the IRT modeling approach implemented. Response times are typically modeled log-normally, although other distributions such as the exponential and gamma distributions have been proposed to model response times (Scheiblechner, 1979; Maris, 1993). Some researchers have suggested that the lognormal distribution provides a better fit to response times than other distributions proposed (Storms & Delbeke, 1992; van der Linden, Scrams, & Schnipke, 1999). In contrast, however, it has also been demonstrated that response times modelled with other distributions have typically yielded parameter estimates that are nearly the same as estimates obtained using the lognormal distribution (Storms & Delbeke, 1992).

The definition of the principles underlying response time differ across psychological domains, but similarities can intuitively be found between them to drive future research. In the measurement of personality, response time has been defined according to the “nearness hypothesis” coined by Kuncel (1973) which suggests that a latent driver of response time defined as “response difficulty” increases as the distance between item and person locations on a latent continuum decreases. This hypothesis has intrinsic similarities to an accepted psychophysics finding suggesting that the uncertainty of detecting a stimulus is greatest when the item is near a person’s absolute threshold (Guilford, 1954). The closest proxy to measuring this uncertainty, or response difficulty, in a latent space is response time. These findings can both be joined to modern psychometrics by linking response time to response variability. In a 1-PLM or 2-PLM IRT model, the variability of a binary item response is greatest when the person-item distance is equal to zero. In other words, when a person absolutely does or does not know the answer to an item, the variability decreases, much like the relationships found between response difficulty and person-item distance in personality measurement, and uncertainty and detection thresholds in psychophysics. Studies performed on personality data have found a consistent (although weak) negative relationship between response time and person-item distance (Kuncel, 1973, 1977; Ferrando & Lorenzo-Sava, 2007; Molenaar, Rózsa, & Kõ, 2021). These studies used the following IRT parameters based on the 2-PLM (or a more constrained version of the 2-PLM where a_i is set to unity) to construct a weighted person-item distance:

$$\delta_{ij} = \sqrt{a_i^2(\theta_j - b_i)^2} \quad (5)$$

where

- δ_{ij} = A measure of person-distance for person j to item i ,
- a_i = the item discrimination parameter for the i^{th} item,
- θ_j = the location of the j^{th} individual on the latent continuum,
- b_i = the location of the i^{th} item on the latent continuum.

In the model proposed by Ferrando & Lorenzo-Seva (2007), the estimated person-distance measure is used in a regression equation to reflect the inverse relationship of person-item distance and response time:

$$\ln(t_{ij}) = \mu + \omega_j + \gamma_i + \beta\delta_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \sim N(0, \sigma^2) \quad (6)$$

where

- β = a regression parameter reflecting the inverse relationship of person-item distance and response time,
- μ = a numerical constant indicating general response time,
- ω_j = an effect parameter for the slowness of the j^{th} respondent,
- γ_i = an effect parameter for the response time required for the i^{th} item.

This modeling approach was applied to binary personality items following a dominance response process and was estimated with two separate procedures, first with the MMAP/EAP procedure and again with MML using the expectation-maximization (EM) algorithm. These procedures shared two common points: First, item parameters of the 2-PLM were estimated from item responses using the MML/MMAP estimation procedure and were treated as known values. Then, the regression parameters and the person parameters were estimated in two separate stages, the calibration stage and the scoring stage. In the calibration stage, only the parameters used in describing response time were estimated. In the scoring stage, person parameters were estimated from item responses and

response times. These steps were repeated until the estimates of μ , ω , γ , σ , and β stabilized. The approach yielded modest improvements in the precision of model parameters over traditional item response theory methods.

van der Linden, Entink, and Fox (2010) proposed a hierarchical approach to IRT model building using response times and dominance-based responses to test items. In contrast to the model proposed by Ferrando and Lorenzo-Sava (2007), their approach assumed that responses and response times followed separate models, each with a different set of item and person parameters. These separate models were defined as first level models and were combined for response and response times with second level models for the joint distributions of item and person parameters. The first-level model of response data used the three-parameter normal-ogive (3-PNO) model:

$$P(U_{ij} = 1; \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\Phi(a_i(\theta_j - b_i)) \quad (7)$$

where

$\Phi(\cdot)$ = the normal distribution function,

θ_j = the ability parameter for the j^{th} test taker,

a_i = a discrimination parameter for the i^{th} item,

b_i = a difficulty parameter for the i^{th} item,

c_i = a guessing parameter for the i^{th} item.

Response times were modelled as lognormal with individual and item parameters that were different from those from the 3-PNO for scored item responses. Specifically, their lognormal response time model was given as:

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2\right\} \quad (8)$$

where

t_{ij} = the RT of the j^{th} test taker,

τ_j = a speed parameter for the j^{th} test taker,

β_i = a time intensity (i.e., demand) parameter for the i^{th} item,

α_i = a discrimination parameter for the i^{th} item.

Note that the parameters defined in the response time model are similar to those for the scored data IRT model, but not identical in function or interpretation. The population model specifies the joint distribution of the person and item parameters, where analogous parameters across response and response time models can covary. In other words, the person parameters from the two models can covary, as can the item parameters. It is at this second step of the procedure that the method can link information between models and allow one set of estimates to inform the other. The results to test items from van der Linden et al. (2010) mirrored that of results from work done on personality data (Ferrando & Lorenzo-Sava, 2007), indicating that the inclusion of response time can improve the precision of model parameter estimates.

1.4 Research Hypothesis

Item response theory models offer researchers a wealth of information about people and test items; however, none of this information can be relied upon if the accuracy of estimates is poor or the model fails to adequately portray the data. To that end, researchers should strive to use an IRT model that matches the psychological process which ultimately

leads to an item response. Additionally, they should continue to investigate ways to improve parameter estimation in IRT models. These two goals provide the thematic structure for the study outlined below.

This paper proposes a novel way to estimate item response theory parameters using the GGUM with response time and response data simultaneously. It is hypothesized that the inclusion of response time data may lead to more precise item and person parameter estimates, particularly in extreme regions where degradation in parameter estimates has previously been seen (Roberts & Thompson, 2011). Response times are, at best, underutilized by psychometricians using unfolding IRT models. However, this information is readily available whenever items are administered with computers and could potentially provide valuable information when estimating GGUM parameters. The proposed study will examine the accuracy of an MCMC parameter estimation procedure for GGUM parameters that includes item response time as an additional observed variable with one that does not include it. These research questions are formalized as follows:

R1: The amount of time taken to respond to attitude questionnaire items k is, at least partially, a function of the distance between the location of an individual and an item along a latent attitude continuum.

R2: Simultaneous incorporation of item response time and item responses within an unfolding IRT framework like the GGUM increases the precision of item and person estimates beyond that of an unfolding IRT model estimated solely using item responses.

METHOD

2.1 Modifications to the GGUM

It was hypothesized that the introduction of response time data can assist in estimation of GGUM parameters. Specifically, it was assumed that item response times are a function of the squared distance between a respondent's location (θ_j) and an item's location (δ_i) on a latent continuum. This function was expected to take the form of a polynomial regression equation, and can be expressed as follows:

$$f(t_{ij} | \theta_j, \delta_i) = \exp \left[b_0 + b_1(\theta_j - \delta_i)^2 + b_2(\theta_j - \delta_i)^{2^2} + \varepsilon_{ij} \right], \varepsilon_{ij} \sim N(0, \sigma^2) \quad (9)$$

where

t_{ij} = the response time of the j^{th} individual on the i^{th} item,

b_0, b_1, b_2 are polynomial regression parameters,

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

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

ε_{ij} = random error for the j^{th} individual and the i^{th} item.

This equation is consistent with the assumption that response times for item responses which follow from an ideal point process likely operate in a bimodal manner that deviates from the person-nearness hypothesis formed from dominance-response measures. In particular, it was anticipated that responses to items located in close proximity to an individual's own location on a latent continuum will result in short response times, items located slightly farther (in either direction along the continuum) will result in long response

times, and items located extremely far away will result in short response times, relatively speaking. This hypothesized relationship is illustrated in Figure 1. It should be noted that item response variability under the GGUM also exhibits a similar bimodal shape as a function of the same distance between the locations of an individual and an item. By squaring the distance between person parameters and item parameters, the relationship between response time, theta, and delta can be modelled using Equation (9).

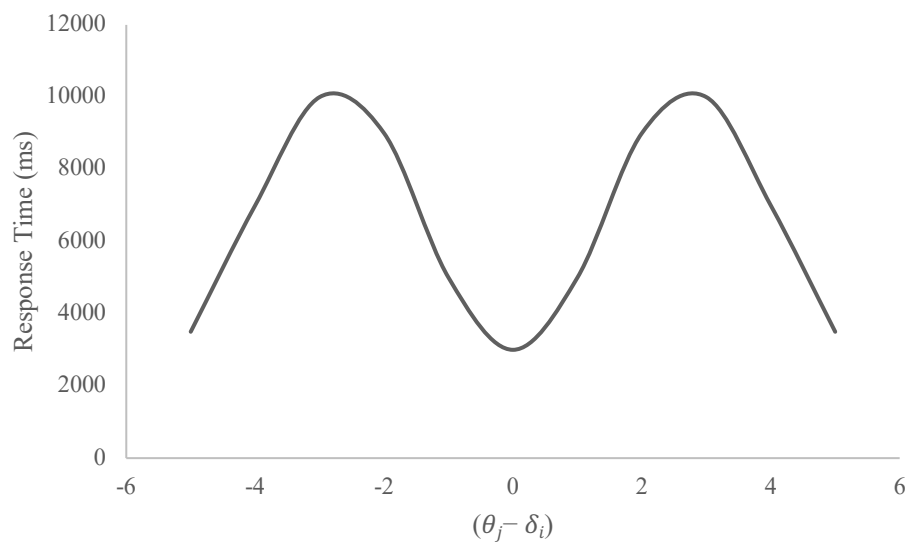


Figure 1 - Expected relationship of response time with GGUM parameters

This proposed addition into the MCMC estimation syntax for the GGUM allowed the person and item location estimates to use information from item response times simultaneously with item responses themselves. With the proper specification of the relationship shared between response time and the squared distance between person and item locations, it was posited that the estimates of both person and item location parameters would be more precise (i.e., have smaller standard errors) when this second source of item-level data is incorporated into the GGUM than when it is ignored. For the

remainder of this document, the resulting system of simultaneous equations will be referred to as the GGUM with response times (GGUM-RT) modeling process.

2.2 Parameter Recovery Simulation

2.2.1 Experimental Design

A simulation study was performed to assess the improvement in the accuracy of parameter recovery when using the proposed GGUM-RT modelling process. This investigation examined the effects of three factors on parameter recovery. These three factors, total sample size (500, 2000), number of items (10, 30), and number of response categories (2, 6), were fully crossed, resulting in a 2x2x2 between-subjects experimental design with 8 cells. Ten replications were conducted within each cell, resulting in a total of $2 \times 2 \times 2 \times 10 = 80$ observations. To compare the GGUM-RT modelling process's performance against the traditional GGUM, the simulated responses across all replications for this design were first estimated using the GGUM and then again with the GGUM-RT.

2.2.2 Data Generation

2.2.2.1 True Parameter Values

True person parameters, θ_j , were generated by randomly sampling from a standard normal distribution. Item parameters $(\alpha_i, \delta_i, \tau_{ik})$ were randomly sampled with replacement from a set of unidimensional GGUM item parameter estimates obtained from abortion attitude items (Roberts & Thompson, 2011). The unidimensional item parameter estimates described by Roberts and Thompson (2011) were divided into five intervals across the latent continuum. To ensure adequate information across the unidimensional continuum,

each of these intervals was then equally sampled (with replacement) to obtain true item parameter values for this simulation. By sampling from a set of previously defined unidimensional GGUM estimates, this generation method encouraged realism in the population of items and the correlations between item parameters. Each cell and replication within a cell was sampled independently.

2.2.2.2 Response Generation

The simulated theta and item parameter values were used along with the GGUM (Equation 2) to calculate the probability of a response in each of the item response categories. The calculated response category probabilities were used to form a multinomial distribution. Observed responses were obtained by sampling from the resulting multinomial distribution.

2.2.2.3 Response Time Generation

Response times were simulated (in milliseconds) to operate as a function of the distance between an individual and an item using Equation 9. For this design, b_0 was set to 4500, b_1 was set to 400, and b_2 was set to -20. This equation operates such that an individual would respond faster to an item when it is located very close or very far away from the individual's location on the latent continuum. However, there exists locations between these regions where an individual's response time is at its maximum, and the peakedness of the distribution of an individual's response times is assumed to be dependent on the width of an individual's latitude of noncommitment, or the range in which an individual perceives items to be neither clearly acceptable nor objectionable. This latitude is primarily a function of discrimination and SRCs associated with a given item (Roberts,

Rost, & Macready, 2010). The amount of noise (e.g., error) added to the model was determined from descriptive information obtained using the real data example described in section 2.3. Error was randomly sampled from a normal distribution with a mean of 0 and a standard deviation of 500ms. Response time was truncated at 200ms to remove unrealistic outliers.

2.2.3 *Parameter Estimation*

2.2.3.1 Starting Values

Starting values were obtained for GGUM parameters using methods previously implemented for the GGUM and the MGGUM (Thompson, 2014; Roberts & Sparks, 2015; King, 2017). Specifically, starting values for item location parameters ($\tilde{\delta}_i$) and person parameters ($\tilde{\theta}_j$) were obtained by performing a detrended correspondence analysis (DCA; Hill & Gauch, 1980) on the item responses and retaining the item and person locations on the first dimension as starting values. Starting values for item discrimination parameters ($\tilde{\alpha}_i$) were set to one for all items. The starting values for thresholds ($\tilde{\tau}_{ik}$) were obtained using a regression equation developed from previously documented estimation procedures for the GGUM (King, 2017). The values associated with the thresholds for item i , except for $\tilde{\tau}_{i0}$ which is set to zero, were a function of the origin constant O_i and the interthreshold distance constant Δ_i :

$$O_i = 1.002 + 0.449|\tilde{\delta}_i| - 0.093V_i \quad (10)$$

$$\Delta_i = 0.921 + 0.058|\tilde{\delta}_i| - 0.129V_i \quad (11)$$

where

V_i = the number of response categories for the i^{th} item,

$\tilde{\delta}_i$ = the starting value of the i^{th} item obtained from DCA.

The initial values for the k^{th} threshold of the i^{th} item were then calculated from the following equation:

$$\tilde{\tau}_{ik} = O_i + \Delta_i(C_i - k_i) \quad (12)$$

The starting value for the C^{th} threshold, $\tilde{\tau}_{iC}$ (the threshold closest to the point, $\theta_j - \delta_i = 0$), was the origin, O_i . The starting values for each following threshold were calculated by adding the interthreshold distance (Δ_i) to the starting value for the previous threshold. In instances where the value of $\tilde{\tau}_{ik}$ was less than 0.1, this value was arbitrarily set to 0.1 to ensure that the starting value was positive and not equal to zero. After the estimation of the traditional GGUM, the person and item results were used as the starting values for the person and item parameters for the GGUM-RT.

Given the novelty of this work, the starting values for the polynomial regression parameters for the GGUM-RT were obtained after the initial GGUM estimation was completed. A polynomial regression model was fit to the person and item parameters obtained from the GGUM and the resulting intercept and slope parameters were used as the starting parameters for the GGUM-RT. This was completed for each replication within a cell independently.

2.2.3.2 Prior Distributions

Regression parameters were estimated with an MCMC technique using uninformative prior distributions: $b_0 \sim N(0,10000)$, $b_1 \sim N(0,1000)$, $b_2 \sim N(0,100)$, $\varepsilon_{ij} \sim N(0,100)$. These priors were selected due to the novelty of the GGUM-RT. Person parameters, item location parameters, item discrimination parameters, and subjective response category parameters were estimated using the following prior distributions: $\theta_j \sim N(0,1)$, $\delta_i \sim N(0,4)$, $\alpha_i \sim N(0,0.25)$, and $\tau_{ik} \sim \text{lognormal}(\mu_{\tau_{ik}}, 1)$, where $\mu_{\tau_{ik}}$ is derived by King (2017) for the GGUM using the starting value for threshold $\tilde{\tau}_{ik}$:

$$\mu_{\tau_{ik}} = \ln \tilde{\tau}_{ik} - \frac{1}{2} \quad (13)$$

2.2.3.3 Software

A program written in R (R Core Team, 2020) was developed for constructing true item and person parameters, generating item responses, generating item response times, and calculating model start values. This dataset was read into JAGS (Plummer, 2003) for the MCMC estimation of GGUM and GGUM-RT parameters. Finally, scripts in R were written using the *psych* package (Revelle, 2019) and the multitude of packages subsumed by the *tidyverse* (Wickham et al., 2019) for calculating parameter recovery statistics and creating plots for this paper.

2.2.3.4 Analysis

The GGUM and GGUM-RT procedure were evaluated by assessing MCMC diagnostic information provided by JAGS to determine if the MCMC algorithm had converged to the target joint posterior distribution of all model parameters. The Gelman-

Rubin index (Gelman & Rubin, 1992), a convergence diagnostic that computes the potential scale reduction factor (PSRF) to indicate convergence, was used to determine if samples of a given parameter converged to a stationary level. Values between 1.0 and 1.2 are considered to be acceptable for convergence, with values closer to one indicating model convergence (Brooks & Gelman, 1997). It was anticipated that the MCMC procedure would need to sample from the Markov chain at least 10,000 times to achieve reasonable convergence; however, given the complexity of the proposed model, more iterations were sometimes required to complete the burn-in period. In instances where the Gelman-Rubin index was greater than 1.2, the MCMC procedure was run for another 5000 cycles and re-evaluated for convergence. Sample trace plots were also referenced to assess model convergence.

After the burn-in period was complete, the model was run a minimum of 5000 more iterations to obtain parameter estimates. Model parameter estimates were calculated using the mean of the post-burn-in sample values for each IRT parameter (i.e., EAP estimates). In order to address the influence that prior distributions had on the scale of parameter estimates, the MCMC estimates were then transformed to the same metric as generated theta parameters. This transformation was completed using a linear regression, where the generated theta values were regressed on the simulated theta estimates. The intercept parameter from these results was used to adjust the locations of the θ and δ estimates on the unidimensional continuum and the slope parameter was used to adjust the range, or scale, of the θ , δ , τ , α , b_1 , and b_2 estimates.

Accuracy of model parameter estimates was assessed based on the average discrepancy between estimated and true generated parameter values using the root mean square deviation (*RMSD*) statistic. *RMSD* is defined in the GGUM as:

$$RMSD = \left(\frac{\sum_{t=1}^T [\hat{\gamma}_t - \gamma_t]^2}{T} \right)^{1/2} \quad (14)$$

where

$\hat{\gamma}_t$ = estimated value of the t^{th} parameter of a given type,

γ_t = true value of the t^{th} parameter of a given type,

T = total number of parameters of a given type in any one replication.

RMSD values were calculated for each item, response time, and person parameter type within a given replication. As *RMSD* has been documented to being sensitive to variance in the difference score between estimated and true values and the degree of squared bias of the estimate (Roberts & Laughlin, 1996), the components of the *RMSD* equation (broken out algebraically in equation 15) were also evaluated. These components simplified to different analytical forms depending on which sets of parameters were being evaluated. Specifically, for the GGUM parameters, the variance of the difference between the estimated parameters and the true parameters was calculated, and the squared difference between average estimated and true parameter was examined. However, for the polynomial regression parameters specific to the GGUM-RT, the *RMSD* equation was further simplified and had fewer components. The equation became a function only of the squared bias for a given replication as there only existed one true value and estimated value for these parameters across replications and they had no

variance. As a result, no further analysis of the *RMSD* values for these parameters was needed to explain the driving component behind statistically significant effects.

$$RMSD = \sqrt{S_{\hat{\gamma}}^2 + S_{\gamma}^2 - 2(S_{\hat{\gamma}\gamma}) + (\bar{X}_{\hat{\gamma}} - \bar{X}_{\gamma})^2} \quad (15)$$

where:

$S_{\hat{\gamma}}^2$ = the sample variance of the N estimated values for parameter γ ,

S_{γ}^2 = the sample variance of the N true values for parameter γ ,

$S_{\hat{\gamma}\gamma}$ = the sample covariance between the estimated and nominal values of γ ,

$\bar{X}_{\hat{\gamma}}$ = the average of the N estimated values for parameter γ ,

and

\bar{X}_{γ} = the average of the N nominal values for parameter γ .

A total of four sets of ANOVAs were conducted to examine GGUM parameters, GGUM-RT parameters, and the comparison of the two. First, four between-replications factorial ANOVAs were conducted to examine the four GGUM parameter types (α_i , δ_i , τ_{ik} , and θ_j) obtained by the GGUM with respect to their *RMSD*. Then, seven between-replications factorial ANOVAs were conducted to examine the four GGUM parameter types (α_i , δ_i , τ_{ik} , and θ_j) and the three regression parameters in Equation 9 (b_0 , b_1 , b_2) obtained by the GGUM-RT. As these sets of ANOVAs use results from two distinct models, first from the GGUM and then from the GGUM-RT, the Type I error rate was controlled within each set of analyses. For the first set of analyses examining the results of the GGUM, the Type I error rate was set to $\alpha = 0.05/4 = 0.0125$. For the second set of analyses examining the results of the GGUM-RT, the Type I error rate for the

seven ANOVA models was adjusted to $\alpha = 0.05/7 = 0.007$. Follow-up analyses examining the individual components of *RMSD* controlled Type I error rates similarly.

A third set of analyses was conducted to compare the *RMSD* for a given parameter (α_i , δ_i , τ_{ik} , and θ_j) estimated with either the GGUM-RT or GGUM. Each of the four ANOVAs in the set used a Type I error adjustment equal to $\alpha = 0.05/4 = 0.0125$. These four split-plot ANOVAs were run a second time using the standard error (SE) of the parameter estimates obtained from the MCMC as the outcome instead of *RMSD*. These split-plot ANOVAs constituted the fourth set, and again, a Type I error rate adjustment equal to $\alpha = 0.05/4 = 0.0125$ was used for them as well. Unlike *RMSD*, SE values are obtainable from the IRT estimation of real data. As a result, it was of interest to demonstrate the change in SEs between the GGUM and GGUM-RT models for both the simulation and the real data analysis.

Because the power to detect effects in GGUM simulations with 10 replications is typically high, the effect size estimate η_w^2 (Roberts & Thompson, 2011) was also used to determine the largest effects in each model. The effect size η_w^2 estimate is defined as:

$$\eta_w^2 = \frac{SS_{effect}}{SS_{family}} \quad (16)$$

This estimate indicates the proportion of sums of squares within a family (within-replications or between-replications families) tested by a given error term which is attributed to an effect of interest. Particular effects were considered interpretable when the calculated effect size (η_w^2) was greater than or equal to 0.10 and the associated p-value was lower than the adjusted alpha cut-off value.

2.3 Real Data Application

The data source to be used for a real data example was an attitude measurement study where individuals responded to a series of items developed to measure individual attitudes towards gun control. Data for this study was collected from 2009 to 2014. For this application, responses and item response times from 1229 respondents to 63 statements were analyzed. Responses used in this analysis were limited to those that passed a test-retest reliability check where individual responses were correlated across 10 repeated statements ($r > 0.7$). These statements are presented in Appendix A. This dataset was collected using a six-point response scale with *strongly disagree*, *disagree*, *slightly disagree*, *slightly agree*, *agree*, and *strongly agree* as possible responses. To promote (although not required) SRCs that were nicely ordered along the latent continuum, the *agree* and *strongly agree* responses were collapsed.

Prior to analyzing responses to the 63 gun control items, a principal components analysis was first performed on the data to determine the dimensionality of the dataset by assessing the first and second eigenvalues of the Pearson interitem correlation matrix. Any items that were found to have a communality less than 0.3 were discarded to ensure the variance in the final item set was appropriately explained by the underlying latent dimension. The remaining items were then analyzed with an MML calibration step using the GGUM2004 software. Any items that did not fit the GGUM according to the infit and outfit chi-square values reported by GGUM2004 were discarded. The remaining items in

the dataset were analyzed using the MCMC approach to the traditional GGUM as well as the modified GGUM procedure including response times.

A preliminary analysis of the data was conducted to identify any outlying cases or unusual RT behavior that persisted after the test-retest reliability check of the response data. There were no extreme outlying cases (e.g., cases likely due to the participant leaving the experiment room or due to a program error) to remove from the dataset. In addition, a number of transformations to the distribution of the empirical RT data were considered; however, none yielded a notable improvement over using the untransformed data. Model parameter estimates were then obtained using JAGS.

RESULTS

3.1 GGUM Parameter Recovery Simulation

Following the estimation of both the GGUM and GGUM-RT on the simulated datasets, the four sets of ANOVAs described in Chapter 2 were conducted to examine GGUM parameters, GGUM-RT parameters, and the comparison of the two. It should be noted that prior to calculating the *RMSD* values to run these analyses it was necessary to match the simulated signs corresponding to a particular end of the unidimensional latent trait continuum. Neither the GGUM nor DCA constrains the poles of the latent continuum to represent a particular ideology, so this was done prior to calculating *RMSD* values. Fortunately, reversing the continuum in this manner does not affect the GGUM likelihood.

Four between-replications factorial ANOVAs were conducted to examine the effect of sample size, the number of items, and the number of response categories for each of the four *RMSD* values for the different parameter types in the traditional GGUM (α_i , δ_i , τ_{ik} , and θ_j). To reduce the probability of Type I error due to repeated testing of the same model for each of the dependent measures, the Type I error rate was adjusted to $\alpha = 0.05/4 = 0.0125$. Statistically significant effects were only retained for interpretation if they also had an associated η^2 value of 0.10 or more. Post hoc paired comparisons of interpretable main effects and interactions were not needed as no effect was associated with more than a single degree of freedom. The results from the analysis are shown in

Tables 1 and 2. Average *RMSD* components and statistical findings on these components are shown in Tables 3 and 4.

TABLE 1 – MEAN RMSD OF PARAMETER ESTIMATES

<i>Effect</i>	<i>RMSD</i>			
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$
Sample Size				
500	0.414	0.243	0.266	0.266
2000	0.248	0.192	0.193	0.270
Test Length				
10	0.419	0.255	0.272	0.342
30	0.243	0.180	0.188	0.194
Response Categories				
2	0.543	0.269	0.234	0.318
6	0.119	0.166	0.225	0.219
Overall	0.366	0.218	0.230	0.268

Note. $\hat{\alpha}$ = item discrimination estimate; $\hat{\delta}$ = item location estimate; $\hat{\tau}$ = item threshold estimate; $\hat{\theta}$ = person location estimate.

TABLE 2 – η^2 VALUES FOR ANOVA EFFECTS

<i>Effect</i>	<i>RMSD</i>			
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$
Sample Size	<i>0.100</i>	<i>0.064</i>	<i>0.218</i>	<0.001
Test Length	<i>0.112</i>	<i>0.142</i>	<i>0.286</i>	<i>0.627</i>
Response Categories	<i>0.657</i>	<i>0.270</i>	0.003	<i>0.281</i>
Sample Size x Test Length	<i>0.003</i>	0.002	0.002	<0.001
Sample Size x Response Category	<i>0.040</i>	0.016	<i>0.055</i>	<0.001
Test Length x Response Category	<i>0.059</i>	<i>0.193</i>	<i>0.201</i>	<i>0.035</i>
Sample Size x Test Length x Response Category	0.002	0.003	0.001	0.001

Note. Values in italics were statistically significant effects at the $p < 0.0125$ level; values in bold were effects with η^2 greater than or equal to .10.

TABLE 3 - RMSD COMPONENTS OF GGUM PARAMETER ESTIMATES FROM THE GGUM

<i>Effect</i>	$\hat{\alpha}$		$\hat{\delta}$		$\hat{\tau}$		$\hat{\theta}$		
	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	
Sample Size									
500	0.073	0.192	0.047	0.023	0.068	0.009	0.069	0.002	
2000	0.020	0.077	0.029	0.021	0.034	0.010	0.069	0.002	
Test Length									
10	0.062	0.212	0.044	0.040	0.065	0.016	0.104	0.003	
30	0.032	0.057	0.031	0.004	0.036	0.003	0.035	0.001	
Response Categories									
2	0.084	0.262	0.046	0.041	0.050	0.016	0.092	0.003	
6	0.009	0.008	0.029	0.004	0.052	0.003	0.046	0.001	
Overall	0.047	0.135	0.038	0.022	0.051	0.009	0.069	0.002	

TABLE 4 – η^2 VALUES FOR ANOVA EFFECTS WITH RMSD COMPONENTS OF GGUM

PARAMETER OUTCOMES

<i>Effect</i>	$\hat{\alpha}$		$\hat{\delta}$		$\hat{\tau}$		$\hat{\theta}$	
	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias
Sample Size	<i>0.182</i>	<i>0.087</i>	<i>0.101</i>	0.001	<i>0.295</i>	0.002	0.000	0.000
Test Length	<i>0.059</i>	<i>0.159</i>	0.053	<i>0.253</i>	<i>0.207</i>	<i>0.264</i>	<i>0.611</i>	<i>0.461</i>
Response Categories	<i>0.367</i>	<i>0.428</i>	<i>0.093</i>	<i>0.278</i>	0.001	<i>0.206</i>	<i>0.268</i>	<i>0.227</i>
Sample Size x Test Length	<i>0.019</i>	<i>0.023</i>	0.001	0.000	0.003	0.012	0.001	0.001
Sample Size x Response Category	<i>0.119</i>	<i>0.078</i>	0.028	0.000	<i>0.061</i>	0.005	0.000	0.000
Test Length x Response Category	<i>0.048</i>	<i>0.139</i>	0.059	<i>0.237</i>	<i>0.134</i>	<i>0.191</i>	<i>0.073</i>	<i>0.132</i>
Sample Size x Test Length x Response Category	0.016	<i>0.020</i>	0.021	0.001	0.000	0.000	0.000	0.001

Note. Values in italics were statistically significant effects at the $p < 0.0125$ level; values in bold were effects with η^2 greater than or equal to .10

3.1.1 Accuracy of the Item Parameter Estimates

The average *RMSD* of item discrimination parameter estimates across all cell conditions was 0.366. Sample size had a moderate effect on the accuracy of the estimates, $\eta^2 = 0.100$, with a larger sample size resulting in more accurate estimation of item discrimination parameters (*RMSD* = 0.414 & 0.248 for sample sizes of 500 & 2000, respectively).

Test length had a moderate effect on the accuracy of the item discrimination parameter estimates, $\eta^2 = 0.112$. Longer test lengths resulted in more accurate estimation of item discrimination parameters (*RMSD* = 0.243) than shorter test lengths (*RMSD* = 0.419).

The number of response categories in the model also had a strong effect on the accuracy of the item discrimination parameter estimates, $\eta^2 = 0.657$. Item discrimination parameters were more accurately estimated with six response categories (*RMSD* = 0.119) than two response categories (*RMSD* = 0.543).

The main effect of sample size had a meaningful effect on the variance of the difference component, the main effect of test length had a meaningful effect on the item bias component of the *RMSD* calculation, and the main effect of the number of response categories had a meaningful effect on both components. This indicated that there was not a single component driving the results found in the overall *RMSD* results.

The average *RMSD* of item location parameter estimates across all cell conditions was 0.218. Test length had a moderate effect on the accuracy of the estimates, $\eta^2 = 0.142$,

with longer test lengths resulting in more accurate estimation of item location parameters ($RMSD = 0.255$ & 0.180 for test lengths of 10 and 30, respectively). However, this main effect had meaningful influence on only the item bias component of the $RMSD$ calculation.

The number of response categories also had a strong effect on the accuracy of the item location parameter estimates, $\eta^2 = 0.270$. Item location parameters estimated were more accurately estimated with six response categories ($RMSD = 0.166$) than with two response categories ($RMSD = 0.269$). Only the item bias component of $RMSD$ was meaningfully impacted by changes in the number of response categories.

The interaction of test length and the number of response categories in the model had a strong effect on the accuracy of the item location parameter estimates, $\eta^2 = 0.193$. Specifically, although the values of $RMSD$ remained low, on average, for item location parameters estimated for designs with six response categories regardless of test length, there was a significant decrease in average $RMSD$ values for item location parameters estimated in designs with two response categories as test length was increased from 10 to 30. The item bias component of $RMSD$ meaningfully impacted this interaction effect. This effect is shown in Figure 2.

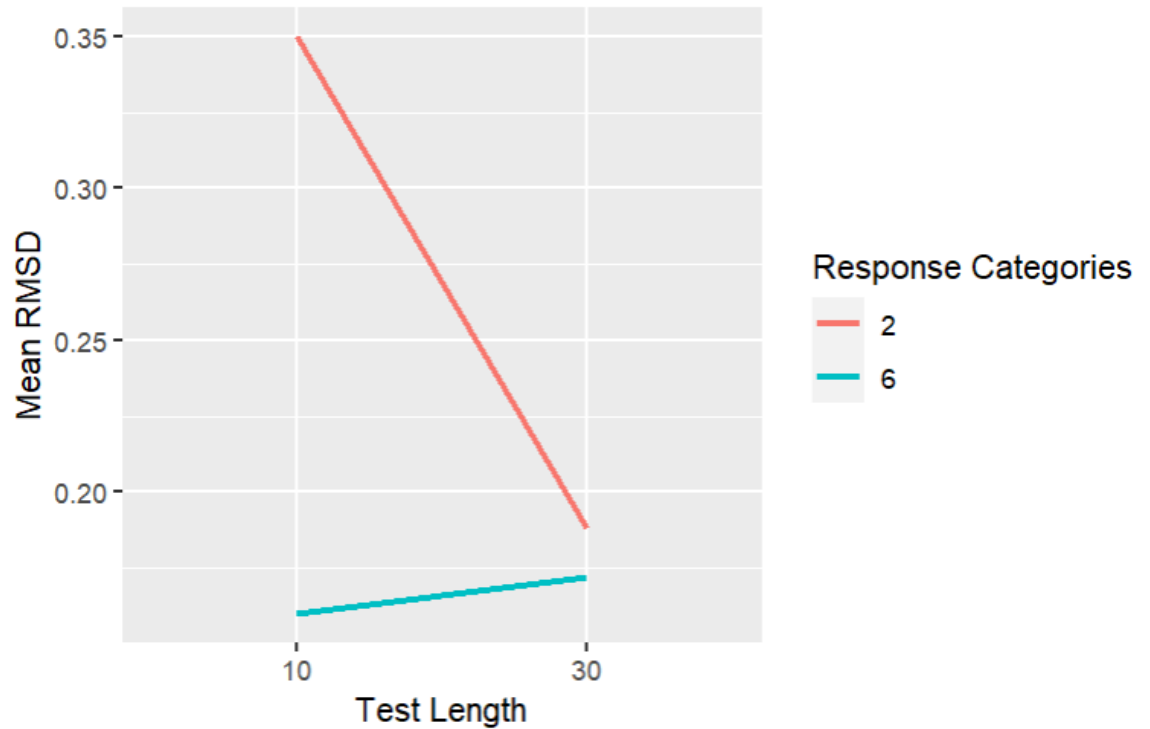


Figure 2 - Interaction plot of the effect of test length and the number of response categories on item location parameter accuracy.

The average *RMSD* of subjective response category threshold parameter estimates across all cell conditions was 0.230. Sample size had a strong effect on the accuracy of the estimates, $\eta^2 = 0.218$, with larger sample sizes resulting in more accurate estimation of subjective response category threshold parameters (*RMSD* = 0.266 & 0.193 for sample sizes of 500 & 2000, respectively). This was primarily due to the effect of sample size on the variance of the difference between estimated and true values.

Test length also had a strong effect on the accuracy of the item threshold parameter estimates, $\eta^2 = 0.286$. Item threshold parameters estimated were more accurately estimated with longer test lengths (*RMSD* = 0.188) than with shorter test lengths (*RMSD* = 0.272). It

was found that both the variance of the difference component and the item bias component of *RMSD* were meaningfully affected by changes in test length.

The interaction of test length and the number of response categories in the model also had a strong effect on the accuracy of the item threshold parameter estimates, $\eta^2 = 0.201$. Specifically, although the values of *RMSD* remained reasonably low, on average, and stable for item threshold parameters estimated for designs with six response categories regardless of test length, there was a significant and substantial decrease in average *RMSD* values for item threshold parameters estimated in designs with two response categories as test length was increased from 10 to 30. Both the variance of the difference component and the item bias component of *RMSD* meaningfully impacted this significant interaction effect. This effect is shown in Figure 3.

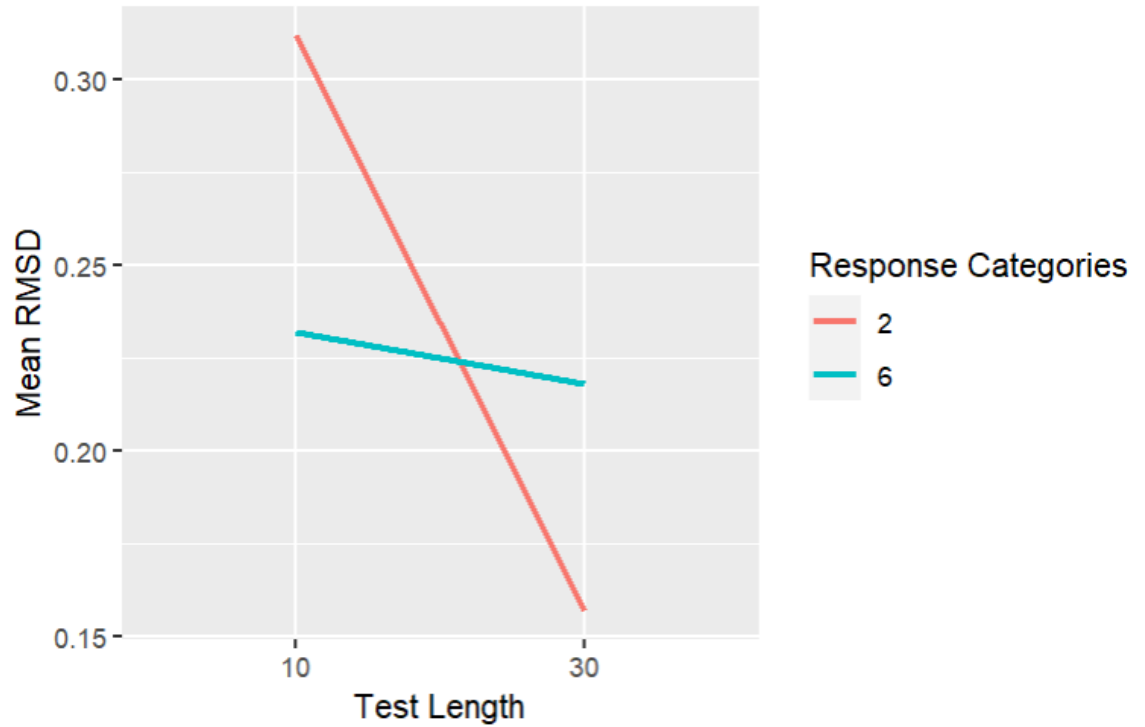


Figure 3 - Interaction plot of the effect of test length and the number of response categories on item threshold parameter accuracy.

3.1.2 Accuracy of the Person Parameter Estimates

The average *RMSD* of person estimates across all cell conditions was 0.268. Test length had a strong effect on the accuracy of the estimates, $\eta^2 = 0.627$, with longer test lengths resulting in more accurate estimation of person parameters (*RMSD* = 0.342 & 0.194 for test lengths of 10 & 30, respectively).

The number of response categories had a strong effect on the accuracy of the person parameter estimates, $\eta^2 = 0.281$. Person parameters estimated were more accurately estimated with six response categories (*RMSD* = 0.219) than with two response categories (*RMSD* = 0.318).

The main effect of sample size and the main effect of the number of response categories both had meaningful effects on the variance of the difference component of the *RMSD* calculation and the item bias component. This indicated that there was not a single component driving the change in values of *RMSD*. Both the variance of the difference component and the item bias component were improved with larger sample sizes and more response categories, when each of these effects was analyzed individually.

3.2 GGUM-RT Parameter Recovery Simulation

Using results from the GGUM-RT hybrid model, seven between-replications factorial ANOVAs were conducted to examine the effect of sample size, test length, and the number of response categories for each of the four *RMSD* values for the different parameter types in the traditional GGUM and the three regression parameters in Equation 9 (i.e., α_i , δ_i , τ_{ik} , θ_j , b_0 , b_1 , and b_2). The Type I error rate for each of these seven ANOVA models was adjusted to $\alpha = 0.05/7 = 0.007$. It was found that all effects with η^2 values of 0.10 or more were also statistically significant, which made effect size the only criterion of interpretation in a practical sense. The results from the analysis are shown in Tables 5 and 6. The components of the calculated *RMSD* values and statistical findings on these components are shown in Tables 7 and 8.

TABLE 5 – MEAN RMSD OF PARAMETER ESTIMATES

<i>Effect</i>	<i>RMSD</i>						
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$	\hat{b}_0	\hat{b}_1	\hat{b}_2
Sample Size							
500	0.357	0.083	0.159	0.191	76.132	19.956	1.711
2000	0.136	0.047	0.077	0.144	10.242	19.152	1.857
Test Length							
10	0.235	0.077	0.120	0.177	34.511	25.414	2.645
30	0.258	0.053	0.116	0.157	51.862	13.693	0.924
Response Categories							
2	0.397	0.077	0.085	0.195	54.879	23.832	2.099
6	0.096	0.052	0.151	0.140	31.494	15.276	1.470
Overall	0.247	0.065	0.118	0.167	43.187	19.554	1.784

Note. $\hat{\alpha}$ = item discrimination estimate; $\hat{\delta}$ = item location estimate; $\hat{\tau}$ = item threshold estimate; $\hat{\theta}$ = person location estimate; \hat{b}_0 = polynomial regression parameter estimate; \hat{b}_1 = polynomial regression parameter estimate; \hat{b}_2 = polynomial regression parameter estimate.

TABLE 6 – η^2 VALUES FOR ANOVA EFFECTS

<i>Effect</i>	<i>RMSD</i>						
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$	\hat{b}_0	\hat{b}_1	\hat{b}_2
Sample Size	0.226	0.217	0.418	0.078	0.284	0.001	0.005
Test Length	0.002	0.089	0.001	0.014	0.020	0.252	0.705
Response Categories	0.420	0.103	0.273	0.105	0.036	0.134	0.094
Sample Size x Test Length	0.011	0.044	0.022	0.095	0.050	0.086	0.004
Sample Size x Response Category	0.107	0.028	0.018	0.032	0.033	0.007	0.000
Test Length x Response Category	0.004	0.001	0.000	0.022	0.028	0.002	0.018
Sample Size x Test Length x Response Category	0.009	0.011	0.007	0.040	0.025	0.009	0.000

Note. Values in italics were statistically significant effects at the $p < 0.007$ level; values in bold were effects with η^2 greater than or equal to .10.

TABLE 7 – RMSD COMPONENTS OF GGUM PARAMETER ESTIMATES FROM THE GGUM-RT

	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$					
<i>Effect</i>	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	
Sample Size									
500	0.056	0.152	0.005	0.004	0.028	0.001	0.043	0.001	
2000	0.010	0.016	0.001	0.002	0.006	0.001	0.020	<0.001	
Test Length									
10	0.030	0.060	0.002	0.005	0.016	0.002	0.029	<0.001	
30	0.036	0.109	0.004	0.001	0.018	0.001	0.034	0.001	
Response Categories									
2	0.059	0.164	0.004	0.004	0.009	0.002	0.043	0.001	
6	0.007	0.004	0.002	0.002	0.025	0.001	0.020	<0.001	
Overall	0.033	0.084	0.003	0.003	0.017	0.001	0.031	0.001	

TABLE 8 – η^2 VALUES FOR ANOVA EFFECTS WITH RMSD COMPONENTS OF GGUM-RT

PARAMETER OUTCOMES

<i>Effect</i>	$\hat{\alpha}$		$\hat{\delta}$		$\hat{\tau}$		$\hat{\theta}$	
	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias	Variance of Difference	Squared Bias
Sample Size	0.239	0.144	0.117	0.088	0.393	0.020	0.086	0.064
Test Length	0.005	0.018	0.045	0.322	0.004	0.091	0.004	0.041
Response Categories	0.306	0.196	0.042	0.100	0.223	0.049	0.085	0.048
Sample Size x Test Length	0.006	0.027	0.096	0.024	0.014	0.044	0.097	0.059
Sample Size x Response Category	0.160	0.130	0.032	0.024	0.068	0.060	0.044	0.043
Test Length x Response Category	0.006	0.018	0.031	0.035	0.003	0.015	0.034	0.042
Sample Size x Test Length x Response Category	0.006	0.026	0.041	0.006	0.004	0.009	0.048	0.045

Note. Values in italics were statistically significant effects at the $p < 0.0125$ level; values in bold were effects with η^2 greater than or equal to .10.

3.2.1 Accuracy of the Item Parameter Estimates

The average *RMSD* of item discrimination parameter estimates across all conditions was 0.247. Sample size had a strong effect on the accuracy of the estimates, $\eta^2 = 0.226$, with a larger sample size resulting in more accurate estimation of item discrimination parameters (*RMSD* = 0.357 & 0.136 for sample sizes of 500 & 2000, respectively).

The number of response categories in the model also had a strong effect on the accuracy of the item discrimination parameter estimates, $\eta^2 = 0.420$. Item discrimination parameters were more accurately estimated with six response categories (*RMSD* = 0.096) than two response categories (*RMSD* = 0.397).

There was also a moderate effect found for the interaction of sample size and the number of response categories, $\eta^2 = 0.107$. Specifically, although both response category conditions benefited from increases in sample size, it was found that samples with two response categories had a much greater increase in item discrimination parameter accuracy with larger sample sizes compared to samples with 6 response categories. This effect is shown in Figure 4.

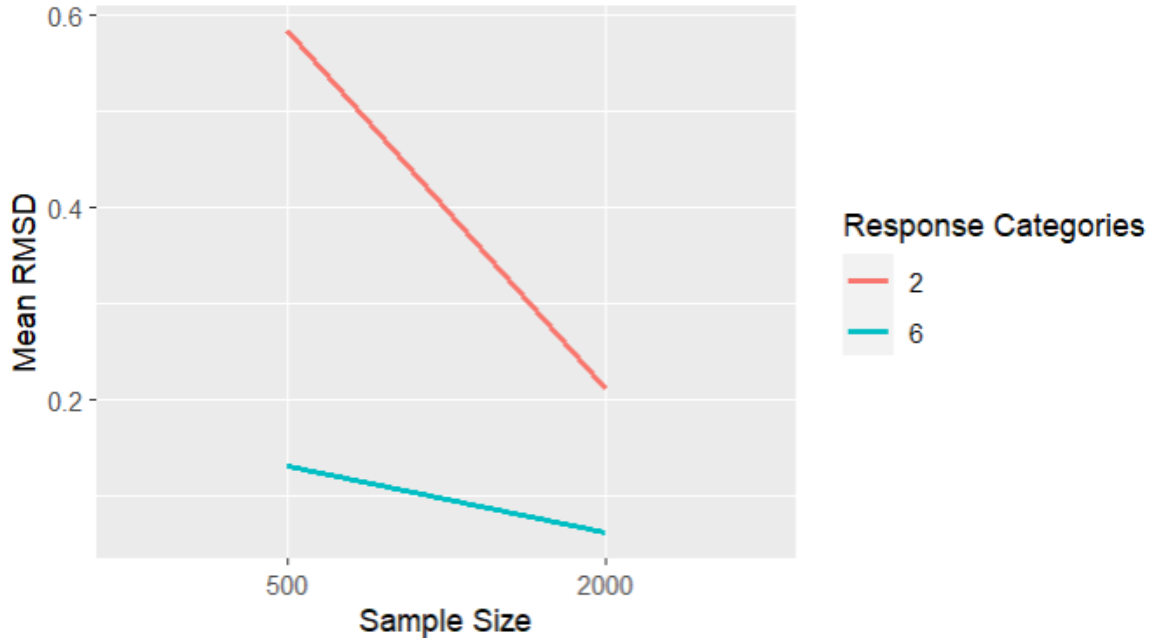


Figure 4 - Interaction plot of the effect of response categories and sample size on item discrimination parameter accuracy

It was discovered that the main effect of sample size, the main effect of the number of response categories, and the interaction of the two had meaningful effects on the variance of the difference between estimated and true discrimination parameters component of the *RMSD* calculation and the squared item bias component. This indicated that there was not a single component driving the results found in the overall *RMSD* results. As expected, both of these component measures decreased with increasing sample size.

The average *RMSD* of item location parameter estimates across all conditions was 0.065. Sample size had a strong effect on the accuracy of the estimates, $\eta^2 = 0.217$, with a larger sample size resulting in more accurate estimation of item location parameters (*RMSD* = 0.083 & 0.047 for sample sizes of 500 & 2000, respectively). It was found that the variance of the difference between estimated and true item parameters component of

the *RMSD* calculation meaningfully differed across sample sizes in the expected direction, indicating that this component drove the significant *RMSD* effect.

The number of response categories also had a moderate effect on the accuracy of the item location parameter estimates, $\eta^2 = 0.103$. Item location parameters estimated were more accurately estimated with six response categories (*RMSD* = 0.052) than with two response categories (*RMSD* = 0.077). An analysis of the individual components of the *RMSD* calculation revealed that this effect was largely driven by the variance of the difference between estimated and true item parameters component, suggesting this component decreased as the number of response categories increased.

The average *RMSD* of subjective response category threshold parameter estimates across all cell conditions was 0.118. Sample size had a strong effect on the accuracy of the estimates, $\eta^2 = 0.418$, with larger sample sizes resulting in more accurate estimation of subjective response category threshold parameters (*RMSD* = 0.159 & 0.077 for sample sizes of 500 & 2000, respectively). In addition, sample size had a meaningful effect on the value of the variance of the difference between estimated and true threshold parameters component of the *RMSD* calculation.

The number of response categories also had a strong effect on the accuracy of the category threshold parameter estimates, $\eta^2 = 0.273$. Item threshold parameters estimated were more accurately estimated with two response categories (*RMSD* = 0.085) than with six response categories (*RMSD* = 0.151). It was found that the variance of the difference component of *RMSD* meaningfully impacted changes by the number of response categories.

3.2.2 Accuracy of the Person Parameter Estimates

The average *RMSD* of person estimates across all cell conditions was 0.167. The number of response categories had a moderate effect on the accuracy of the person parameter estimates, $\eta^2 = 0.105$. Person parameters were more accurately estimated with six response categories (*RMSD* = 0.140) than with two response categories (*RMSD* = 0.195).

The main effect of the number of response categories had significant effects on the variance of the difference component of the *RMSD* calculation and the item bias component, but neither had effect sizes large enough to meet the recommended interpretation criteria.

3.2.3 Accuracy of Parameter Estimates in the Response Time Regression Model

The average *RMSD* for the b_0 response time parameter estimates in the response time regression model across all cell conditions was 43.187. Sample size had a strong effect on the accuracy of the estimates, $\eta^2 = 0.284$, with a larger sample size resulting in more accurate estimation of this response time parameter (*RMSD* = 76.132 & 10.242 for sample sizes of 500 & 2000, respectively).

The average *RMSD* of b_1 response time parameter estimates across all cell conditions was 19.554. Test length had a strong effect on the accuracy of the estimates, $\eta^2 = 0.252$, with longer test lengths resulting in more accurate estimation of this response time parameter (*RMSD* = 25.414 & 13.693 for test lengths of 10 & 30, respectively).

The number of response categories also had a moderate effect on the accuracy of the b_1 parameter estimates, $\eta^2 = 0.134$. These quantities were more accurately estimated with six response categories ($RMSD = 15.276$) than with two response categories ($RMSD = 23.832$).

The average $RMSD$ of b_2 response time parameter estimates across all cell conditions was 1.784. Test length had a strong effect on the accuracy of the estimates, $\eta^2 = 0.705$, with longer test lengths resulting in more accurate estimation of this response time parameter ($RMSD = 2.099$ & 1.470 for test lengths of 10 & 30, respectively).

3.2.4 *Effect Sizes from the Response Time Regression Model*

As response time was modelled using a polynomial regression equation, it was possible to calculate the resulting R^2 values to determine the magnitude of the effect of the difference between a person's location and the location of an item and response time. The average R^2 values across conditions are in Table 9. Across all conditions, the average R^2 value was moderate-to-high, 0.621. An ANOVA was conducted to assist in the interpretation of the magnitude of the differences between the R^2 values across simulation conditions. The main effect of sample size and the main effect of test length were found to be statistically significant. Larger sample sizes resulted in significantly larger amounts of explained variance. The opposite was found for increased test length: longer tests resulted in significantly lower amounts of explained variance. No statistically significant effects were found for the number of response categories. Although the parameter estimates for b_1 and b_2 were found to be more precise in conditions with longer test lengths, the overall

effect size for the relationship between the IRT parameters and the response time parameters decreased.

TABLE 9 – MEAN R^2 ACROSS SIMULATION CONDITIONS

<i>Effect</i>	<i>Mean R^2</i>
Sample Size	
500	0.568
2000	0.674
Test Length	
10	0.680
30	0.561
Response Categories	
2	0.598
6	0.644
Overall	0.621

3.3 Comparison of the GGUM and GGUM-RT

Prior to running statistical analyses, descriptive statistics pertaining to the *RMSD* values for the GGUM parameter estimates obtained from the GGUM and the GGUM-RT model types were first examined. For each replication, the *RMSD* for each model was compared, and a tally for the number of times a given model produced a lower *RMSD* value was calculated for each parameter. This tally is shown as a frequency table in Table 10. Other *RMSD* descriptive statistics have been previously shown in Tables 1 and 5. These frequencies offered additional context to the statistical findings and suggest that there were instances where the GGUM-RT did offer improvements in *RMSD* above and beyond the GGUM.

TABLE 10 – COUNT OF SMALLER RMSD VALUES BY MODEL

TYPE	<i>RMSD</i>	
	GGUM	GGUM-RT
α	7	73
$\hat{\delta}$	1	79
$\hat{\tau}$	5	75
$\hat{\theta}$	7	73

Next, two sets of four split-plot ANOVAs were conducted to compare the accuracy of the GGUM-RT against the GGUM parameter estimates with a Type I error adjustment equal to $\alpha = 0.05/4 = 0.0125$. The first set of four used *RMSD* as the outcome variable and the second set used *SE* as the outcome variable. Model type (GGUM or GGUM-RT) was treated as a within-subjects factor as both models were applied to the same response data for each replication in the design. This analysis used the same criteria for interpretation as the previous analysis, where statistically significant effects were only retained for discussion if the effect size was greater than or equal to 0.10. However, only the within-subjects effects meeting this interpretation criteria were examined in detail as the focus of this analysis is on the comparison of the two model types. Between-subjects effects for *RMSD* values were described in the previous two sections.

One difference between the previous analysis and the present analysis was the calculation of the effect size. This analysis utilized η_w^2 , which uses the total sum of squares value for a family of effects as the denominator of the calculation. For within-

subjects effects, the total sums of squares value was calculated as the total sum of squares of the within-subjects effects, including the within-subjects error term. (For the between-subjects effects, the total sums of squares value was calculated similarly, with the exception of the sums of squares for the intercept which was not included in this sum. However, these calculations are not pertinent in the current analysis.) The statistical results from the *RMSD* analysis are shown below in Table 11.

TABLE 11 – η^2 WITHIN-FAMILY VALUES FOR WITHIN-SUBJECTS ANOVA EFFECTS WITH RMSD OUTCOMES

Effect	<i>RMSD</i>			
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$
Within-Subjects				
Model type	<i>0.173</i>	<i>0.732</i>	<i>0.647</i>	<i>0.463</i>
Response categories x model type	<i>0.090</i>	<i>0.048</i>	<i>0.073</i>	<i>0.021</i>
Test length x model type	0.238	<i>0.021</i>	<i>0.083</i>	0.185
Sample size x model type	<i>0.019</i>	0.002	0.001	<i>0.031</i>
Response categories x test length x model type	<i>0.148</i>	<i>0.062</i>	<i>0.065</i>	<i>0.042</i>
Response categories x sample size x model type	0.014	0.011	0.005	0.013
Test length x sample size x model type	<i>0.038</i>	0.000	0.002	<i>0.027</i>
Response categories x test length x sample size x model type	0.027	0.000	0.001	0.010

Note. Values in italics were statistically significant effects at the $p < 0.0125$ level; values in bold were effects with η^2 greater than or equal to .10.

3.3.1 Estimation Accuracy Between the GGUM and the GGUM-RT with RMSD Outcomes

The largest difference in estimation accuracy was observed for the item location parameters ($RMSD = 0.218$ and 0.065 for GGUM and GGUM-RT, respectively), with the main effect of model type accounting for 73.2% of the variation in the item location parameter $RMSDs$.

The second largest difference in estimation accuracy was observed for the item threshold parameters ($RMSD = 0.230$ and 0.118 for GGUM and GGUM-RT, respectively), with the main effect of model type accounting for 64.7% of the variation in the item threshold parameter $RMSDs$.

Statistically significant differences in estimation accuracy were also observed for the person parameters ($RMSD = 0.268$ and 0.167 for GGUM and GGUM-RT, respectively), with the interaction of test length and model type accounting for 18.5% of the variation in person parameter $RMSDs$ and the main effect of model type alone accounting for 46.3% of the variation in the person parameter $RMSDs$.

The significant interaction of test length and model type for the person parameters, shown in Figure 5, demonstrated that the GGUM yielded larger improvements to accuracy as test length increased than the GGUM-RT, which saw relatively low $RMSD$ values regardless of test length.

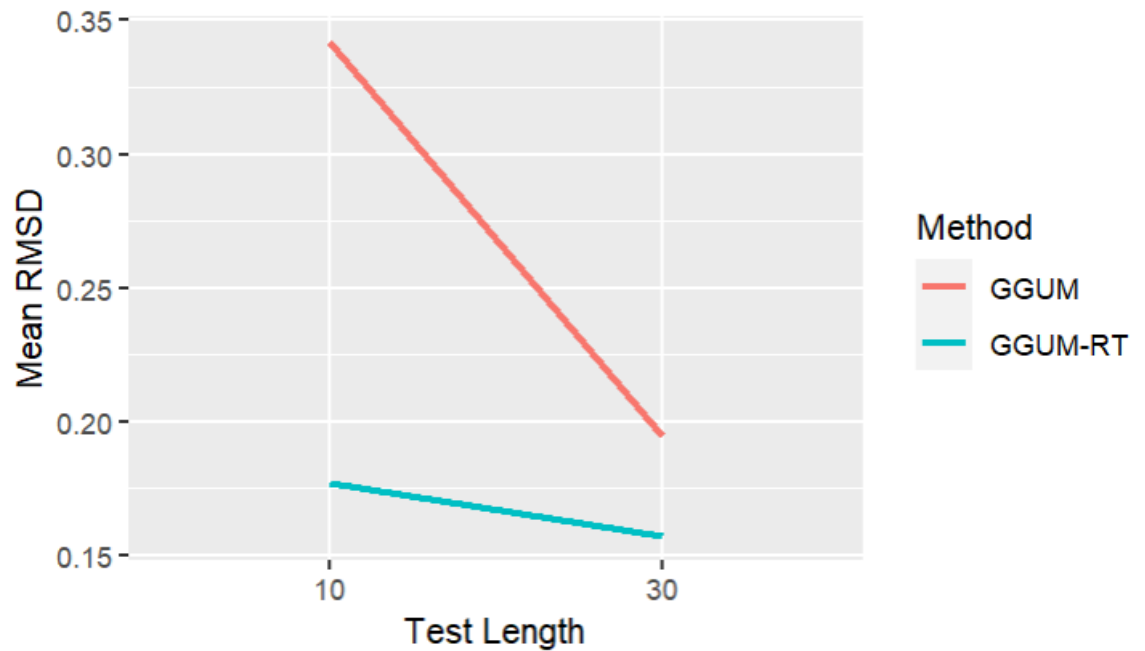


Figure 5 - Interaction plot of the effect of model type and test length on person parameter accuracy.

Finally, statistically significant differences in estimation accuracy were observed for the item discrimination parameters ($RMSD = 0.331$ and 0.247 for GGUM and GGUM-RT, respectively), with the two-way interaction of test length and model type accounting for 23.8% of the variation in person parameter $RMSDs$ and the main effect of model type accounting for 17.3% of the variation in item discrimination $RMSDs$. Figure 6 visualizes the interaction effect. The GGUM yielded larger improvements to accuracy as test length increased than the GGUM-RT. The GGUM-RT yielded relatively low $RMSD$ values regardless of test length. For test lengths of 30 items, there were negligible differences between the two models.

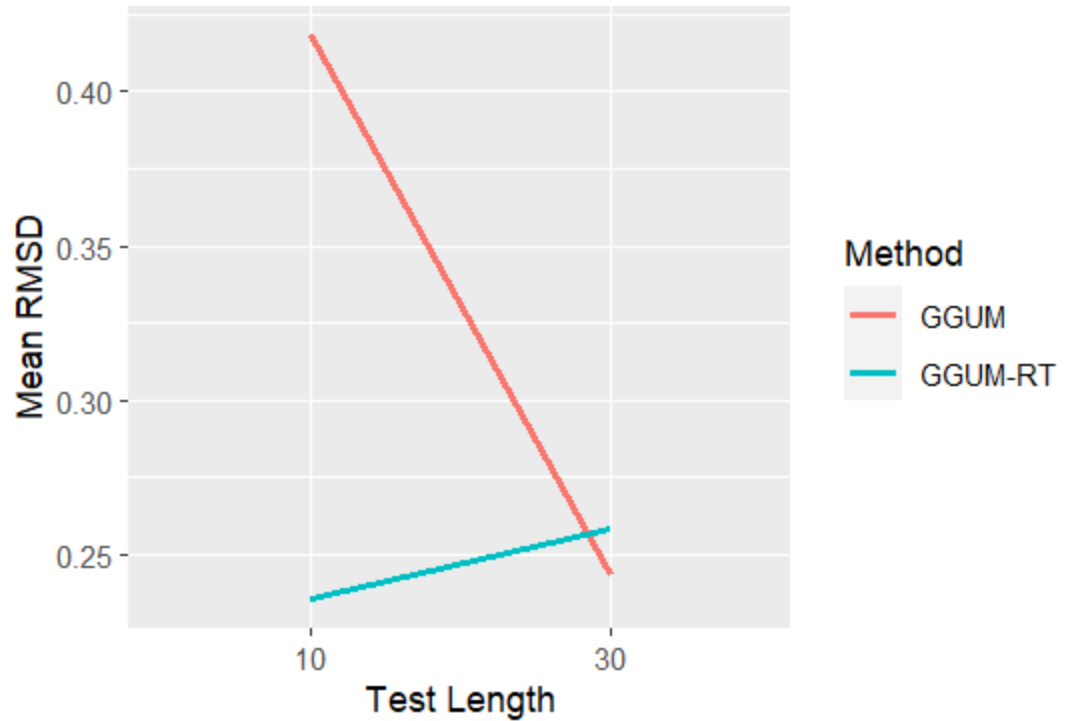


Figure 6 - Interaction plot of the effect of model type and test length on item discrimination parameter accuracy.

3.3.2 Estimation Accuracy Between the GGUM and the GGUM-RT with SE Outcomes

The second outcome that was used to compare the GGUM and the GGUM-RT was standard error of the parameter estimates. While *RMSD* is a measure of accuracy of estimated parameters compared to their true, generated values, the comparison of *SEs* gives researchers information pertaining to whether the precision of the estimates was improved from the estimation of the first model to the second. As they are innately different measures, the results derived with these two measures varied. Average standard errors of parameter estimates and statistical results can be found in Tables 12 and 13.

TABLE 12 – MEAN SE OF PARAMETER ESTIMATES

Effect	SE							
	$\hat{\alpha}$		$\hat{\delta}$		$\hat{\tau}$		$\hat{\theta}$	
	GGUM	GGUM -RT	GGUM	GGUM -RT	GGUM	GGUM -RT	GGUM	GGUM- RT
Sample Size								
500	0.003	0.002	0.017	0.006	0.010	0.003	0.004	0.004
2000	0.002	0.001	0.017	0.004	0.010	0.002	0.003	0.002
Test Length								
10	0.003	0.002	0.018	0.005	0.011	0.002	0.004	0.003
30	0.002	0.002	0.015	0.005	0.009	0.002	0.003	0.003
Response Categories								
2	0.003	0.002	0.023	0.005	0.012	0.002	0.004	0.003
6	0.002	0.001	0.01	0.005	0.008	0.003	0.003	0.003
Overall	0.003	0.002	0.017	0.005	0.010	0.002	0.004	0.003

TABLE 13 – η^2 WITHIN-FAMILY VALUES FOR WITHIN-SUBJECTS ANOVA EFFECTS WITH SE OUTCOMES

Effect	SE			
	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$
Within-Subjects				
Model type	<i>0.691</i>	<i>0.641</i>	<i>0.787</i>	<i>0.124</i>
Response categories	<i>0.061</i>	<i>0.214</i>	<i>0.082</i>	<i>0.219</i>
x model type				
Test length	<i>0.111</i>	<i>0.014</i>	0.005	<i>0.365</i>
x model type				
Sample size	<i>0.063</i>	0.004	0.004	<i>0.169</i>
x model type				
Response categories x test length	<i>0.035</i>	<i>0.010</i>	<i>0.010</i>	<i>0.035</i>
x model type				
Response categories x sample size	<0.001	0.001	0.004	0.000
x model type				
Test length x sample size	<i>0.004</i>	0.001	0.000	0.004
x model type				
Response categories x test length	0.002	0.001	0.001	0.003
x sample size				
x model type				

Note. Values in italics were statistically significant effects at the $p < 0.0125$ level; values in bold were effects with η^2 greater than or equal to .10.

A statistically significant interaction between test length and model type was found for the item discrimination parameters ($SE = 0.003$ and 0.002 for GGUM and GGUM-RT, respectively). The interaction effect accounted for 11.1% of the variation in item discrimination parameter SE s.

The significant interaction of test length and model type for the item discrimination parameters demonstrated that, although the GGUM-RT provided smaller standard errors than the GGUM regardless of the number of response categories, as the number of response categories was increased, the GGUM saw significantly more improvement in the size of its standard errors, on average, relative to the GGUM-RT. These results can be seen in Figure 7.

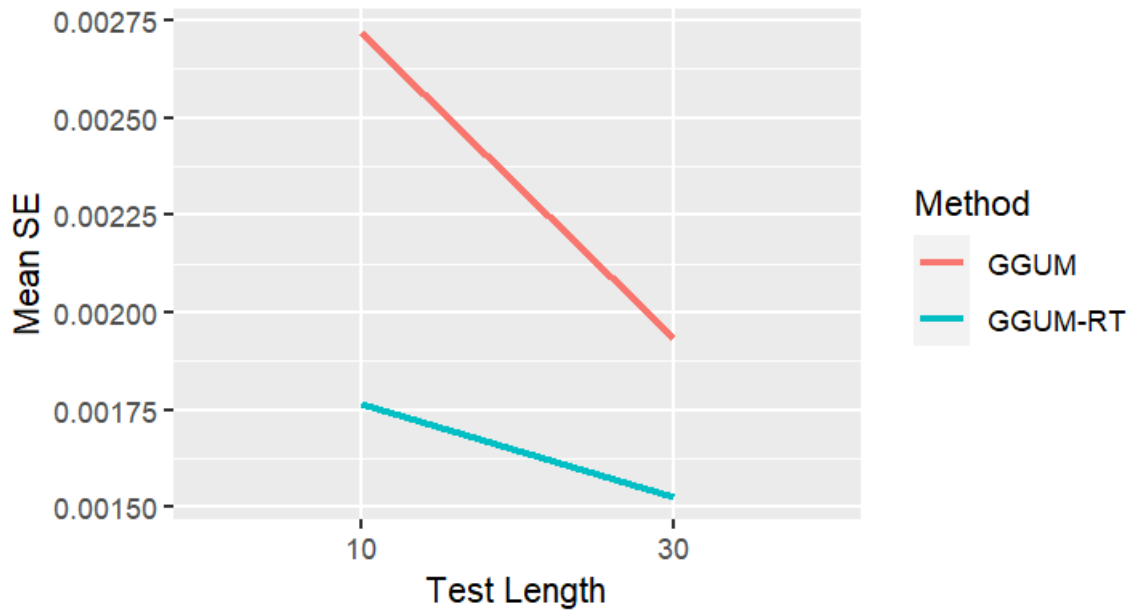


Figure 7 - Interaction plot of the effect of model type and test length on average item discrimination parameter standard error.

The main effect of model type was found to be statistically significant for the item discrimination parameters as well as the item threshold parameters. This effect accounted

for 69.1% of the variation in item discrimination parameter *SEs* and 78.7% of the variation in item threshold parameter *SEs*. The estimates from the GGUM-RT had significantly smaller average standard errors than the estimates from the GGUM.

The item location parameters had two strong statistically significant effects that met the effect size threshold for interpretation. These estimates had overall standard errors of 0.017 and 0.005 for the GGUM and GGUM-RT, respectively. The main effect of model type accounted for 64.1% of the variation in item location *SEs* and the interaction of the number of response categories and model type accounted for 21.4% of the variation in item location *SEs*.

The significant interaction of the number of response categories and model type for the item location parameters demonstrated that, although the GGUM-RT provided smaller standard errors than the GGUM regardless of the number of response categories, as the number of response categories was increased, the GGUM saw significantly more improvement in the size of its standard errors, on average, relative to the GGUM-RT. These results are shown in Figure 8.

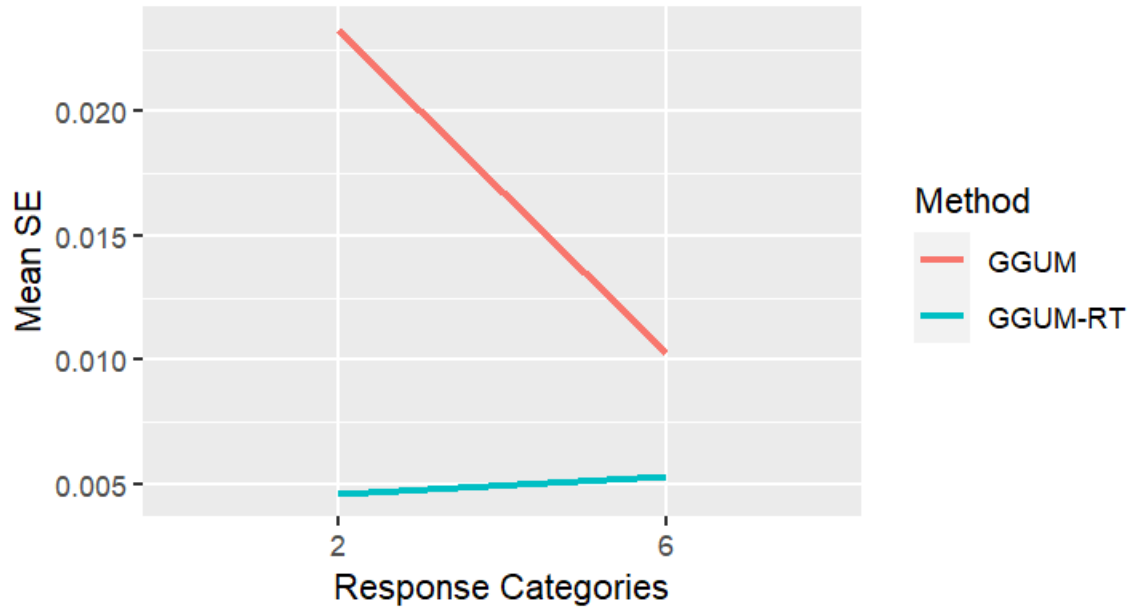


Figure 8 - Interaction plot of the effect of model type and response categories on average item location parameter standard error.

The person parameters had four meaningful and significant effects. The person parameters had overall standard errors of 0.004 and 0.003 for the GGUM and the GGUM-RT, respectively. The significant main effect of model type accounted for 12.4% of the variation in the person parameter SEs. The interaction effect of test length and model type accounted for 36.5% of this variation. The patterns of these effects mimicked the findings from the *RMSD* results for item locations above. Lastly, the interaction effect of the number of response categories and model type accounted for 21.9% of the variation in person parameter *SEs* and the interaction of sample size and model type accounted for 16.9%.

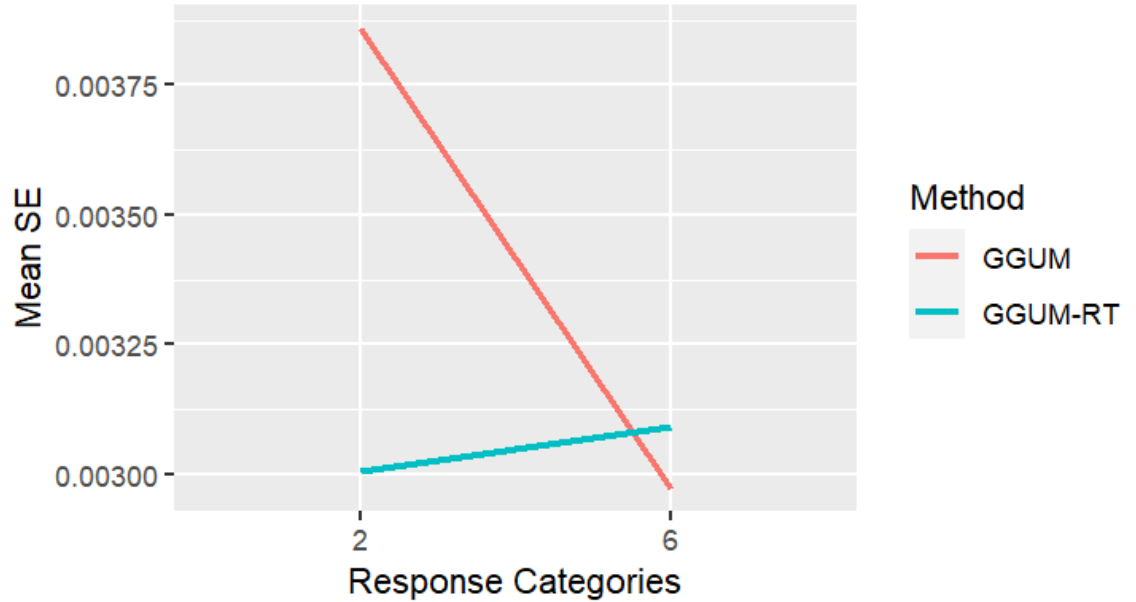


Figure 9 - Interaction plot of the effect of model type and response categories on average person parameter standard error.

The significant interaction of the number of response categories and model type for the person parameters, shown in Figure 9, demonstrated that although the GGUM-RT provided smaller standard errors than the GGUM for cells with fewer response categories, these differences attenuated as the number of response categories was increased. A simple main effects analysis of these results confirmed that, when the number of response categories was six, there were no statistically significant differences between the SEs obtained from the GGUM and the GGUM-RT for the person parameters.

The significant interaction of sample size and model type for the person parameters demonstrated that although the GGUM-RT and the GGUM had similar SEs for smaller sample sizes, the GGUM-RT saw a relatively larger reduction in average standard error as sample size was increased. This effect is shown in Figure 10. A simple main effects analysis of these results confirmed that, for conditions where the sample size was 500, there

was no statistically significant differences between the SEs obtained from the GGUM and the GGUM-RT for the person parameters.

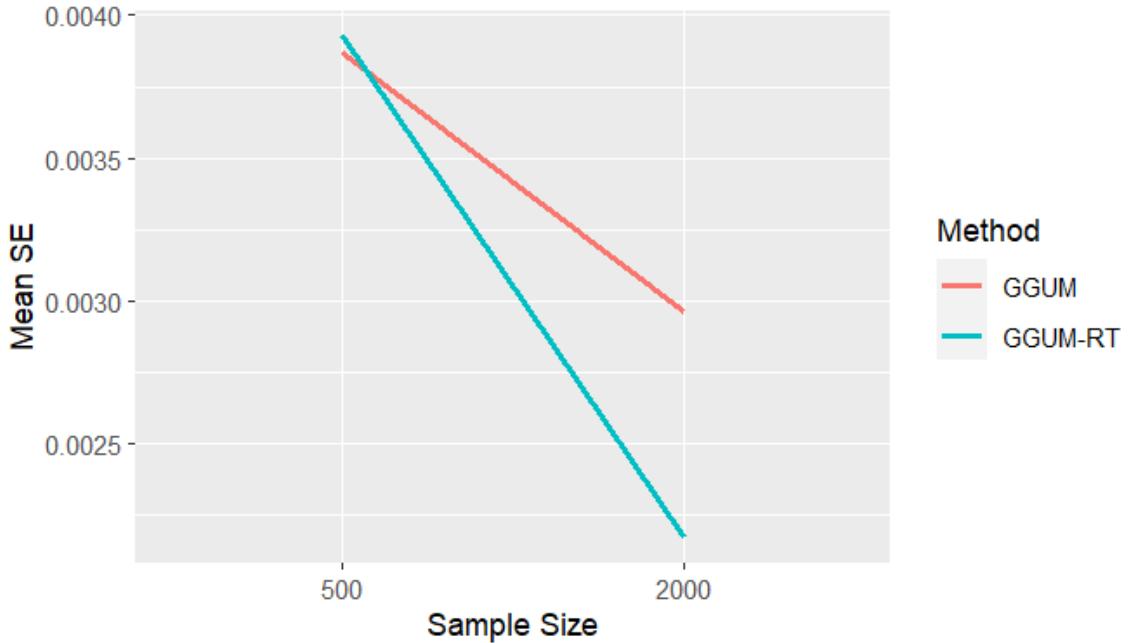


Figure 10 - Interaction plot of the effect of model type and sample size on average person parameter standard error.

3.4 Empirical Data Application

The results of the simulation study demonstrated that the addition of response times into the GGUM has the potential to improve the accuracy of GGUM parameter estimates and decrease the standard error of the estimates in a variety of conditions. In order to further examine the applicability of this extension of the GGUM, the GGUM-RT and the traditional GGUM were both used to model a set of real data from an attitude questionnaire where individuals responded to a series of items developed to measure attitude towards gun control.

Prior to applying the two IRT model types, the dataset was first prepared using the procedure described in Chapter 2. Although the original dataset contained 1229 respondents and 63 gun control attitude statements, the number of responses (and response times) in the cleaned dataset was 851 and the number of attitude statements was 53. Although several participant responses were removed on the basis of the reliability of their responses, the check of the response times of the remaining participants revealed that all response times were under two minutes. This suggested that all responses remaining in the dataset were 1) reliable on the basis of the correlation check and 2) did not exhibit massive outliers in response time due to external factors (computer errors, leaving the testing room, etc.). Therefore, no additional responses were removed based on the response time data. The item parameter estimates for the resulting solution using the GGUM-RT are given in Table 14. The parameter estimates for the polynomial regression terms are given in Table 15. The corresponding item characteristic curves (ICCs) are located in Appendix B.

TABLE 14 – GGUM-RT ITEM PARAMETER ESTIMATES OF GUN CONTROL ATTITUDE STATEMENTS

Item	$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}_1$	$\hat{\tau}_2$	$\hat{\tau}_3$	$\hat{\tau}_4$
Gq14	0.72	-2.73	3.27	2.89	2.48	0.91
Gq4	0.48	-2.53	2.93	2.07	0.51	0.28
Gq1	0.94	-2.53	3.40	2.52	1.54	2.12
Gq2	0.61	-2.47	4.86	3.70	5.07	3.15
Gq17	0.85	-2.38	2.95	2.52	1.22	1.92
Gq15	1.42	-2.34	3.82	3.21	3.08	2.32
Gq12	1.52	-2.25	3.60	2.85	2.68	1.94
Gq16	0.69	-2.19	2.33	2.50	1.47	2.31
Gq19	1.47	-2.19	3.40	3.11	2.81	2.18
Gq18	1.41	-2.18	3.67	2.94	3.19	2.11

TABLE 14 (continued)

Gq11	1.19	-2.16	2.96	2.57	2.37	1.88
Gq20	1.55	-2.13	3.41	2.91	2.96	2.24
Gq13	1.42	-1.91	3.19	2.61	2.09	1.91
Gq21	0.97	-1.56	3.43	2.55	2.49	1.81
Gq23	1.07	-1.46	2.72	2.15	1.81	1.21
Gq24	1.04	-1.46	3.05	2.31	2.39	1.48
Gq3	0.97	-1.39	3.03	1.65	1.88	1.23
Gq22	0.77	-1.29	2.15	1.34	1.22	0.46
Gq29	0.79	-1.28	2.17	1.42	1.24	0.28
Gq26	0.74	-1.13	2.80	1.79	1.76	0.51
Gq31	0.50	-1.04	2.79	1.23	0.95	0.24
Gq25	0.98	-1.03	2.66	1.53	1.75	0.81
Gq27	1.01	-0.94	3.16	2.43	2.56	1.74
Gq30	0.56	-0.94	2.64	0.88	0.73	0.15
Gq35	1.07	-0.23	2.93	2.42	2.62	1.87
Gq37	1.05	-0.21	2.55	2.35	2.95	1.36
Gq36	1.12	0.01	2.15	1.50	2.34	0.80
Gq6	0.85	0.12	1.35	1.05	0.37	0.06
Gq33	1.49	0.14	1.43	0.95	0.47	0.14
Gq39	0.50	0.15	0.70	0.70	0.30	0.25
Gq41	1.27	0.15	1.30	0.90	0.18	0.17
Gq38	0.73	0.20	2.74	1.70	2.59	1.15
Gq5	1.34	0.20	1.32	0.91	0.73	0.42
Gq32	1.20	0.25	1.52	0.77	1.08	0.29
Gq48	0.66	0.76	1.91	1.13	0.16	0.07
Gq47	0.45	0.93	2.14	0.97	1.02	0.60
Gq8	0.61	1.40	2.76	1.78	0.93	0.26
Gq44	0.85	1.43	2.12	1.18	0.78	0.43
Gq50	0.88	1.57	2.10	1.23	0.16	0.19
Gq52	0.78	1.78	2.40	1.34	0.39	0.19
Gq45	0.96	1.83	2.42	1.74	0.71	0.46
Gq59	0.86	1.88	3.38	2.74	3.19	1.93
Gq9	0.97	1.95	3.71	2.97	3.66	2.13

TABLE 14 (continued)

Gq49	1.13	2.05	2.77	2.04	0.99	1.01
Gq46	1.03	2.11	2.53	2.10	1.01	1.17
Gq55	0.59	2.14	3.19	3.08	4.70	2.98
Gq53	0.96	2.19	1.29	1.20	0.56	0.74
Gq54	0.91	2.20	4.01	3.30	3.81	2.80
Gq60	1.20	2.22	1.31	1.26	1.00	1.27
Gq58	1.06	2.24	2.34	1.92	0.78	1.28
Gq61	0.81	2.40	3.92	3.38	4.10	2.89
Gq62	0.74	2.64	4.25	3.27	4.63	3.46
Gq63	1.31	2.81	2.54	2.47	1.93	1.64

TABLE 15 – GGUM-RT POLYNOMIAL REGRESSION**PARAMETER ESTIMATES**

	Posterior mean	95% CI
b_0	8066.891	[7331.365; 8802.417]
b_1	-31.253	[-126.773; 64.267]
b_2	-0.711	[-11.577; 10.155]

The difference in the mean standard error values for the GGUM and the GGUM-RT were evaluated to assess whether there was statistically significant reduction in the standard errors for each parameter type due to the inclusion of response time. The average standard errors of the parameter estimates are presented in Table 16. These averages were calculated from the standard error values obtained from JAGs for each estimated parameter.

TABLE 16 – MEAN SE OF REAL DATA PARAMETER ESTIMATES

<i>SE</i>							
$\hat{\alpha}$		$\hat{\delta}$		$\hat{\tau}$		$\hat{\theta}$	
GGUM	GGUM-RT	GGUM	GGUM-RT	GGUM	GGUM-RT	GGUM	GGUM-RT
0.0024	0.0019	0.0254	0.0242	0.0246	0.0233	0.0049	0.0040

These difference in the standard errors were assessed with four different repeated measures tests, one for each GGUM parameter type ($\alpha_i, \delta_i, \tau_{ik}$, and θ_j). For these tests, each row of data contained the average standard error for a given parameter type for a given replication. As a result, there were 80 data points for each test. The Type I error rate for these four models was adjusted to $\alpha = 0.05/4 = 0.0125$. Statistically significant effects were only considered interpretable if they also had eta-squared values of 0.10 or more. These results are shown in Table 17.

TABLE 17 – η^2 VALUES FOR REAL DATA EFFECTS

<i>SE</i>			
$\hat{\alpha}$	$\hat{\delta}$	$\hat{\tau}$	$\hat{\theta}$
<i>0.125</i>	0.0002	0.0003	<i>0.105</i>

Note. Values in italics were statistically significant effects at the $p < 0.0125$ level; values in bold were effects with η^2 greater than or equal to .10.

In the current study, it was found that the gun control attitude statements fit relatively well on a unidimensional scale. Items located towards the negative end of the scale represented “pro-gun” attitudes, items located towards the center of the scale represented neutral gun control attitudes, and items located towards the positive end of the scale represented attitudes in favor of gun control. The spread of the item location parameter estimates for the GGUM-RT model is shown in Figure 11.

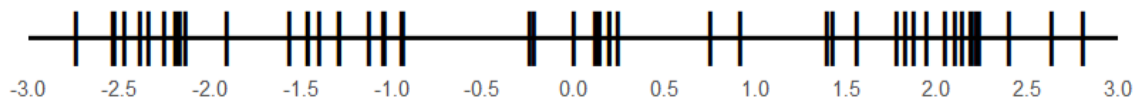


Figure 11 - Estimated item locations from the GGUM- RT

3.4.1 Analysis

Out of the four parameter types that were tested, the item discrimination parameters and the person parameters were the only two parameters that showed statistically significant differences with effect sizes that met the interpretation criteria. The item discrimination parameters had overall standard errors of 0.004 and 0.002 for the GGUM and GGUM-RT, respectively. There was a moderate effect, with model type accounting for 12.5% of the variation in item discrimination *SEs*. The person parameter estimates had overall standard errors of 0.0049 and 0.0040 for the GGUM and GGUM-RT, respectively. The effect of model type accounted for 10.5% of the variation in the person parameter *SEs*. These results were similar, but not as strong as the results found in the simulation analysis, where standard errors consistently dropped across all parameter types with the GGUM-RT model.

It was hypothesized that a bimodal distribution existed between response time and the difference between estimated person parameters and item location parameters, such that the distribution was centered around zero. This relationship is plotted using the real data results in Figure 12. The expected relationship appears visible amongst the outliers in the dataset (i.e., individuals who took longer to respond to the attitude statements). However, a deeper dive examining the individuals who responded more quickly to the statements yielded much noisier results. The hypothesized bimodal relationship was less prevalent amongst these respondents.

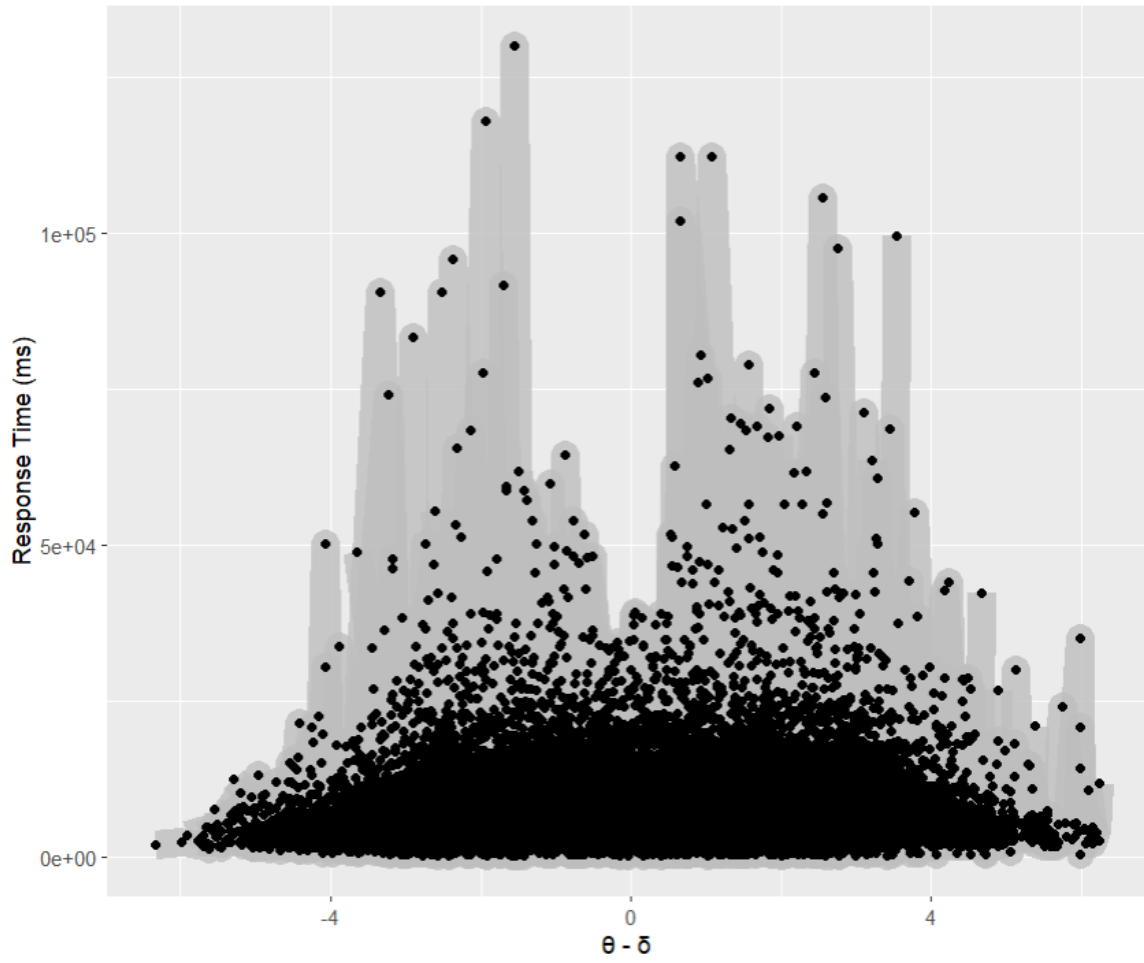


Figure 12 - Relationship of response time and the difference between estimated person and item location parameters

DISCUSSION

4.1 Final Thoughts on the Comparison Between the GGUM-RT and GGUM

The primary goal of this paper was to introduce a novel unfolding IRT model that conjointly estimated model parameters using responses and response times. The GGUM-RT successfully and consistently demonstrated its capability to increase the accuracy and reduce the standard error of the GGUM parameter estimates in the simulated study.

The GGUM-RT was found to reduce the standard error of the estimates for all GGUM parameters in the simulation, and the item discrimination and person parameters in the real data application. Although the current iteration of the GGUM-RT did not reduce enough error in the item location and item threshold parameters in the real data application to be meaningful, it did not increase the standard error of the estimates. This suggests that while the GGUM-RT did not statistically significantly support the shape of the hypothesized distribution, nor did it provide unanimous improvements to parts of the estimation process in the real data application, the inclusion of response times into the GGUM improved SEs of some estimates and did not negatively impact any of them. These promising simulation and real data results suggest that the GGUM-RT may indeed improve unfolding IRT measurement in practice, and the model certainly deserves further investigation.

4.2 Pros and Cons

The largest benefit of the inclusion of the response times in the GGUM is the potential to increase the accuracy of the person parameter estimates and the associated

standard errors. In many use cases, the ability to more accurately estimate the locations of an individual's ability or attitude on a latent continuum is extremely valuable. Further, the additional data provided in the proposed model is of no additional cost to the researcher as most research software records response time data by default in computer administered tests and surveys. To that end, a GGUM that includes response times is a valuable addition to the unfolding IRT literature.

One caveat in including response time data is that it may contain many sources of extraneous variation. The magnitude of the effect that response times can contribute towards the precision of item response theory model parameters is limited by the amount of uncontrollable noise in the response time data (e.g., individual response speed, distractions in the study room, etc.). The amount of noise in this type of data will vary by use case, and until more research is conducted across different types of unfolding data, preliminary analyses of response time data should first be conducted in order to ascertain its applicability to this model.

4.3 Scientific Importance of the Current Study

This study provided the opportunity to assess the degree to which typical unfolding IRT estimates could be improved via the inclusion of response time data. The addition of response time data to item response data, occasionally referred to as conjoint item response theory models, are somewhat recent in the psychometric literature and unexplored in the unfolding model context. Evaluation of the increase in accuracy and precision for unfolding model parameter estimates with the incorporation of response times is a novel and exciting contribution of this work.

The proposed procedure is a sensible extension of the GGUM that allows researchers to supplement response data without the need to develop complicated, external methods of data collection. Most data collection software has the capability to automatically record response times, affording the ability to collect this data simultaneously with item responses themselves. To that end, this project serves as a practical and valuable source of information for researchers interested in exploring external means to better estimate GGUM parameters.

4.4 Future Directions

Although the work in this document demonstrates strong evidence that response times can, and likely should, be included in the GGUM, there were clear next steps identified as possible future directions for the GGUM-RT. The most notable next step is a revision to the polynomial regression equation specified in Equation 9. Although in the real dataset presented in this document it was revealed that GGUM parameters could be estimated by specifying a polynomial equation with a single intercept and two slope parameters, there are intuitively individual differences in response times that should likely be modelled at the individual level. To that end, future extensions on the GGUM-RT might include regression terms that vary at the person level or, alternatively, the inclusion of either latent classes or a form of speed parameter as described by van der Linden (2007). Standardizing the regression weights would also simplify the starting values and prior distributions used in the model by allowing users to plug in standard normal priors instead of calculating sample-specific prior values for unstandardized coefficients. Other possible next steps would be to move away from the rigid polynomial

regression equation model altogether and instead develop a more flexible conjoint modelling approach.

Finally, different ways of transforming RT data were not explored in this document. Depending on the dataset and the area of research, response time data may be handled a variety of different ways, through lognormal transformations, square root transformations, and others. The unidimensional unfolding data accessible during this project did not benefit from these transformations and therefore, they were not implemented; however, depending on the type of data used and the sample of respondents surveyed, it may be of interest to explore how different response time datasets perform with this model and what adjustments can be made to accommodate those transformations without sacrificing the improvements offered by this approach.

APPENDIX A. GUN CONTROL ATTITUDE STATEMENTS

Table A.1. Gun control attitude statements

Every adult should have the right to bear arms.
The government has no right to limit a person's access to owning a gun.
A person should retain the right to own a gun based on his or her preferences.
No one should take away the right of the people to keep and bear arms.
In order to defend our country against terrorism, we should be allowed to own guns.
If it is against the law to own a gun, then law abiding citizens will be at a disadvantage to those who are willing to break the law.
Every law abiding adult should have the right to own a gun because each person has the right to protect oneself.
No person should be denied the right to bear arms.
Owning a gun is an individual matter and should not be controlled by anyone.
People should have the freedom to defend themselves with guns.
People should make their own decisions about owning a gun.
Owning a gun should be a personal choice.
Regardless of my own views on gun ownership laws, I believe others have the right to choose for themselves.
Anyone should have the right to own a gun unless they are under the age of 21.
Everyone should have the right to own a gun unless they have a criminal record.
Owning a gun should always be legal except when the person desiring a gun has been diagnosed as mentally unstable.
A person should be allowed to own a gun unless he or she has a previous criminal record where a gun was used.
Gun ownership should generally be legal except for assault weapons.
An individual should have the right to own a gun unless he or she has been diagnosed as a drug/alcohol abuser.
An individual should have the right to own a gun if he or she has received the appropriate safety training.
An individual should have the right to own a gun unless the person is a foreign national.
Gun ownership should not be regulated except in high crime areas.

Table A.1. (Continued)

Everyone should have the right to own a gun unless the gun uses armor piercing bullets.

Owning a gun should generally be legal except for people who are diagnosed as having below-normal intelligence.

Gun ownership should only be limited when purchasing semi-automatic or automatic guns.

My feelings about the right to own a gun are very mixed.

I cannot whole-heartedly support either side of the gun ownership debate.

My feelings about gun ownership are neither positive or negative.

The right to own a gun should be decided on the state level.

The right to own a gun has both positive and negative consequences.

There are aspects both for and against the legality of gun ownership that appeal to me.

There are both good and bad repercussions of laws that limit gun ownership.

There are certain instances where private gun ownership should be allowed, but there are other instances where it should be banned.

I am apathetic about the gun ownership debate.

The legality of gun ownership debate has no effect on my life.

I am indifferent about the gun ownership issue.

I am neither for nor against laws that seek to limit gun ownership.

No one should have the right to bear arms unless the gun is registered with the government.

There should be gun ownership limitations except in neighborhoods with high crime rates where individuals can protect themselves.

Owning a gun should be illegal, except in extreme cases, when one's life is being threatened.

No one should have the right to bear arms unless they receive formal firearms training.

Guns should not be accessible unless one is using the gun for an exhibition or demonstration.

Only members of the military and police officers should own guns.

No one should have the right to own a gun except for reasons of self-defense.

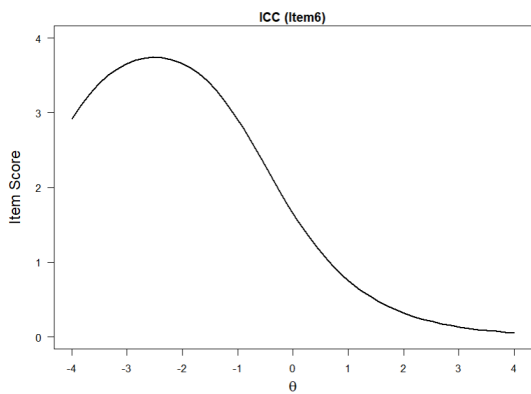
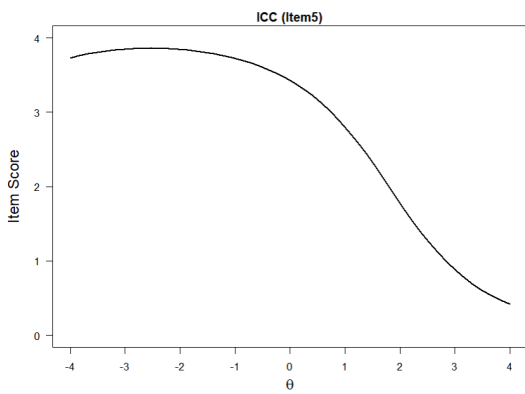
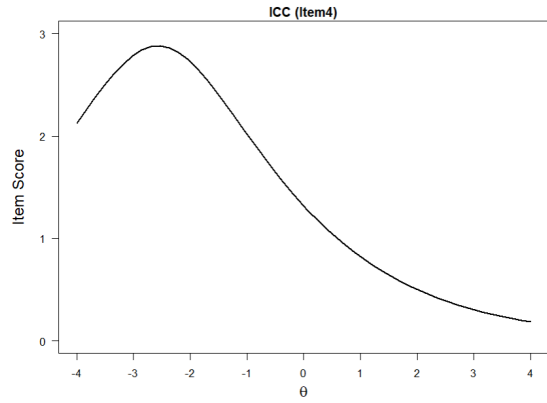
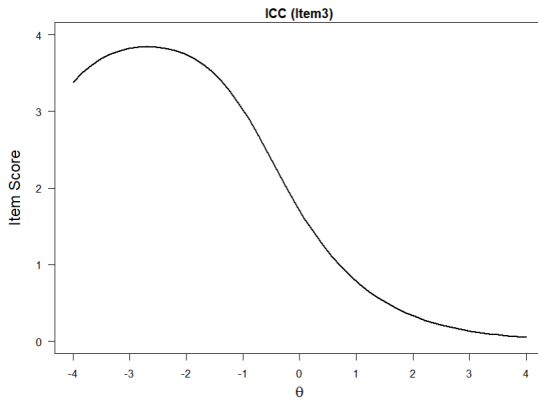
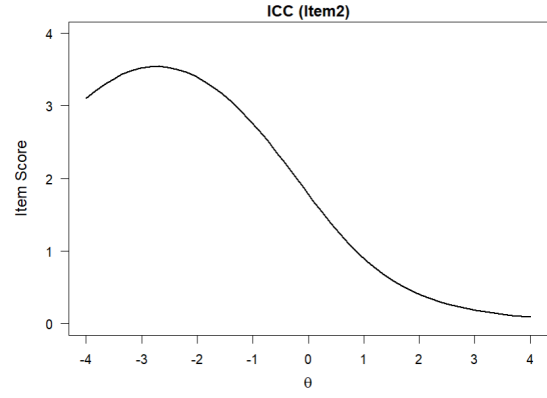
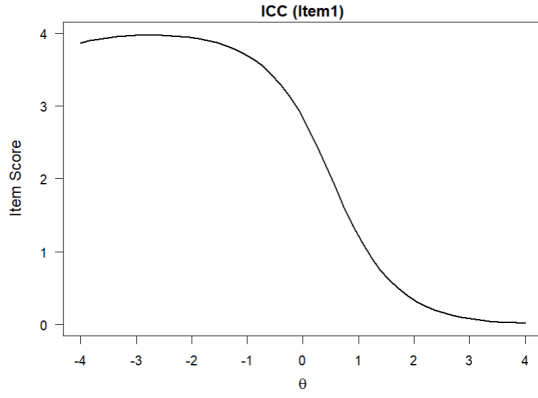
Gun ownership should generally be restricted to those cases where there are no children in the household.

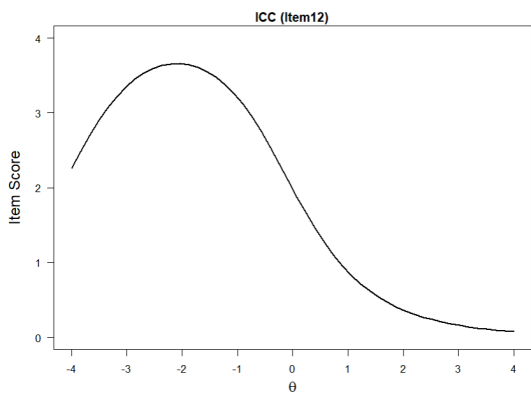
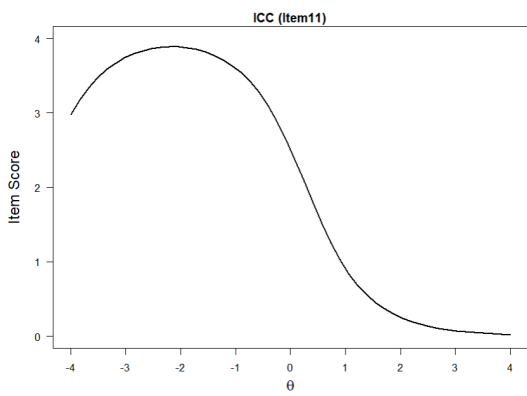
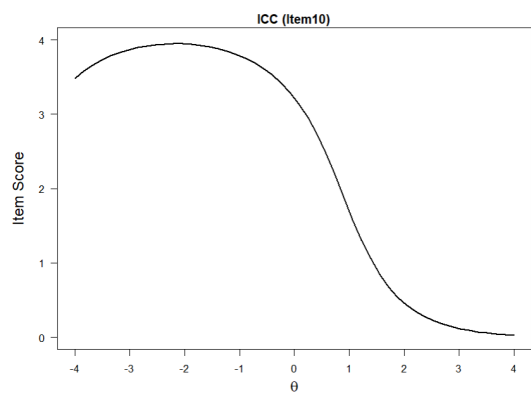
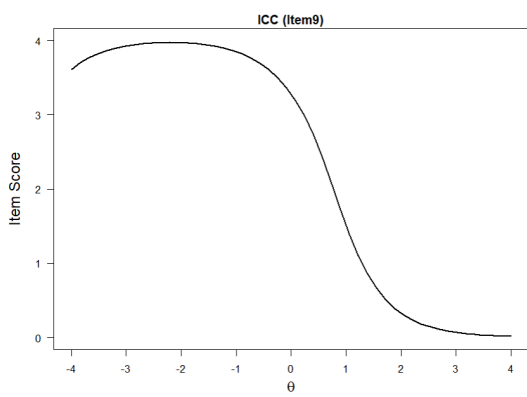
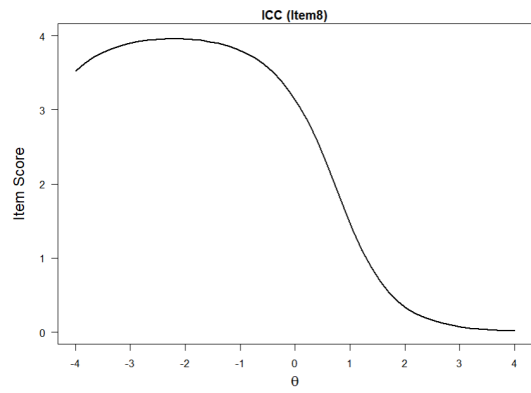
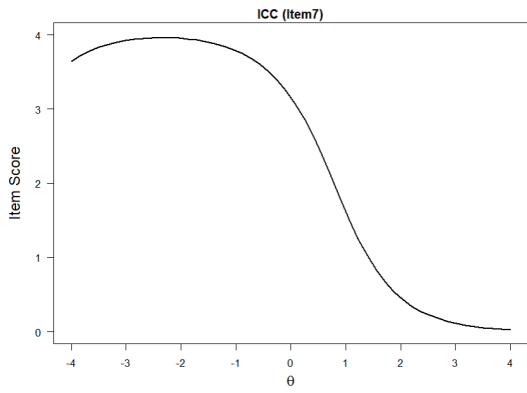
No one should have the right to own a gun except for occupational purposes.

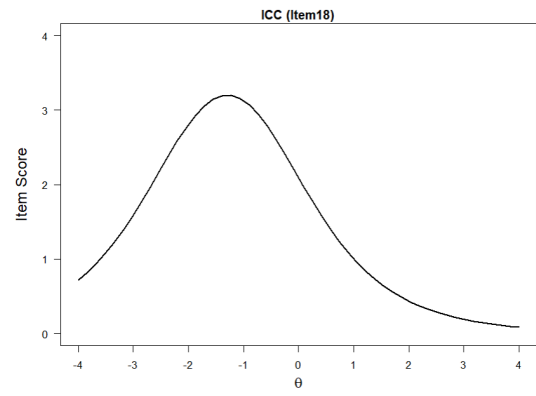
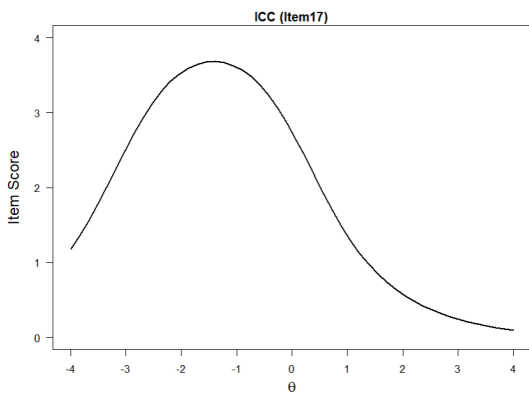
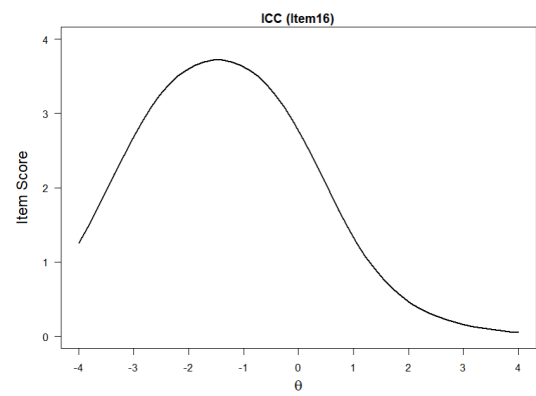
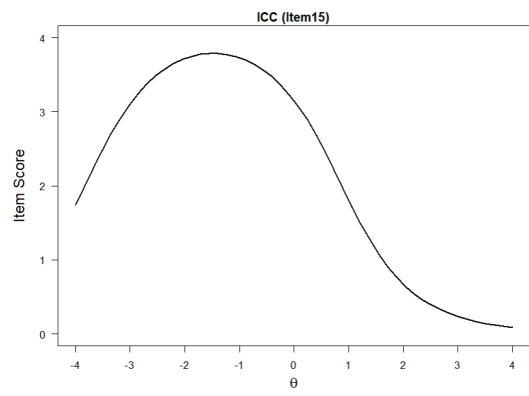
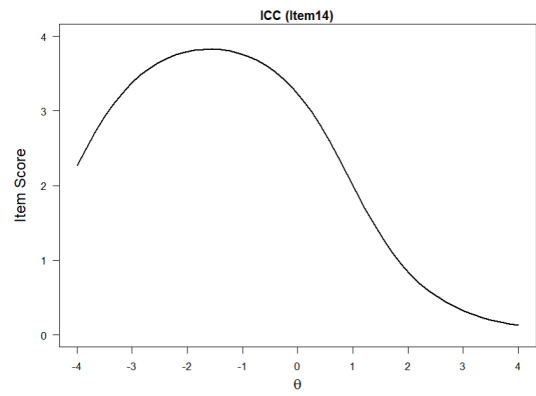
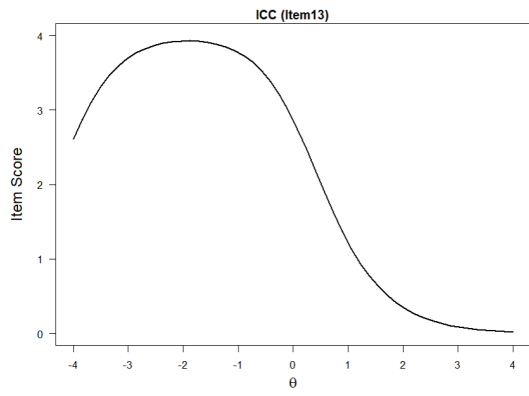
Table A.1. (Continued)

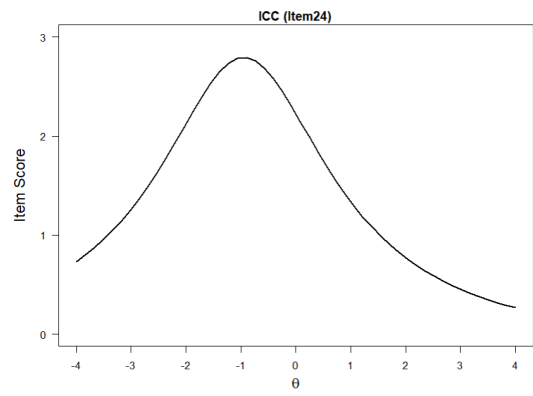
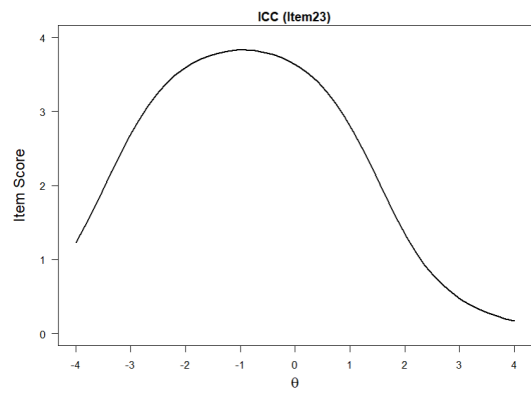
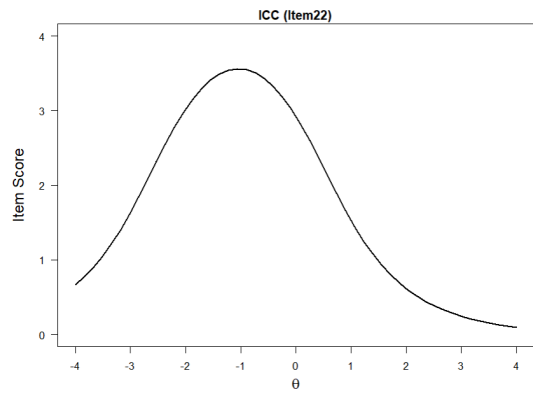
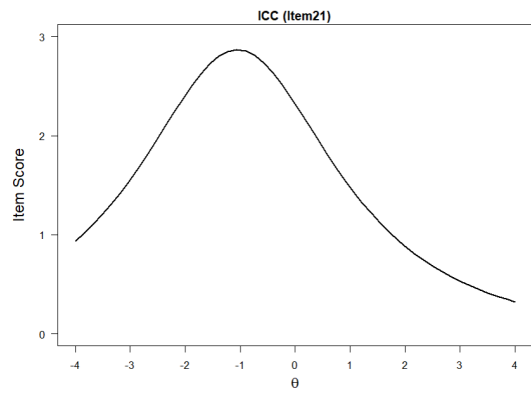
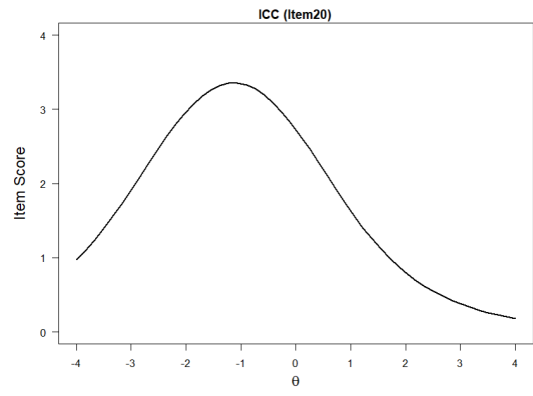
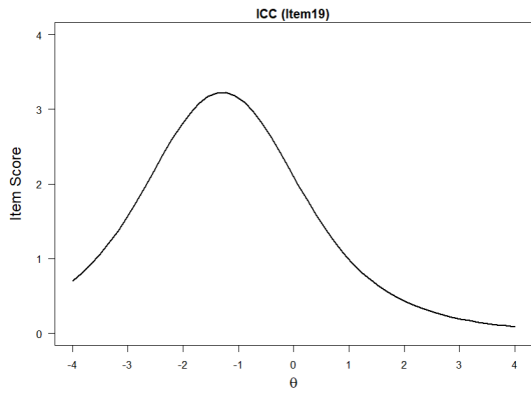
Gun ownership should be illegal unless one has military experience.
Privately-owned guns should be limited to hunting weapons.
No citizen should own a gun unless they are deemed mentally stable.
No citizen has the right to own a gun except when it is necessary to control the wild animal population in nearby areas.
Owning a gun is immoral.
The government has the right to regulate gun ownership to protect its citizens.
Owning a gun is dangerous.
Owning a gun is unacceptable under any circumstances.
Guns should not be made readily available to everyone.
No one should have the right to own a gun because guns are no longer necessary for survival.
The government should regulate gun ownership to protect innocent bystanders.
The government should have the right to restrict gun ownership so people do not take the law into their own hands.
All guns should be banned.
Citizens should not be allowed to own a gun for any reason.
The government has the right to impose gun ownership limits on its citizens.
Limits on gun ownership are absolutely necessary.
The constitution should be amended to eliminate private gun ownership.

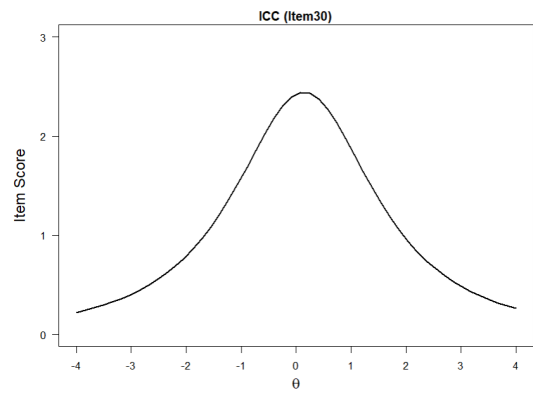
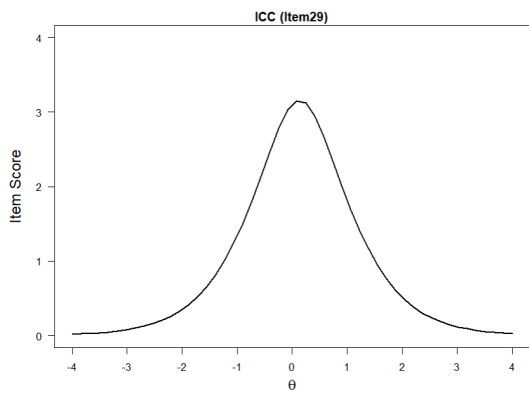
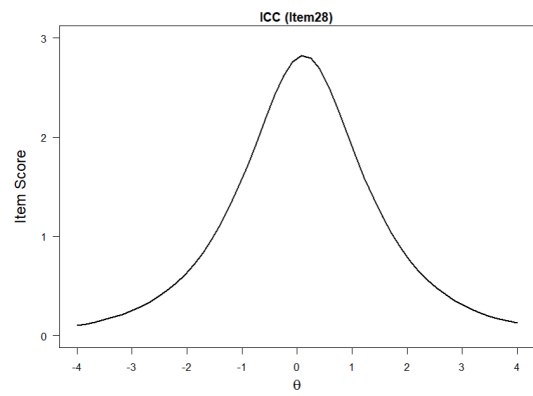
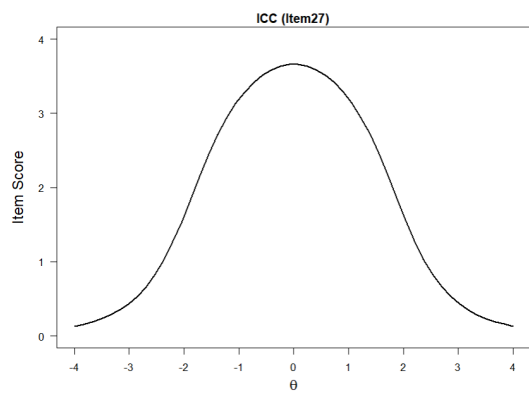
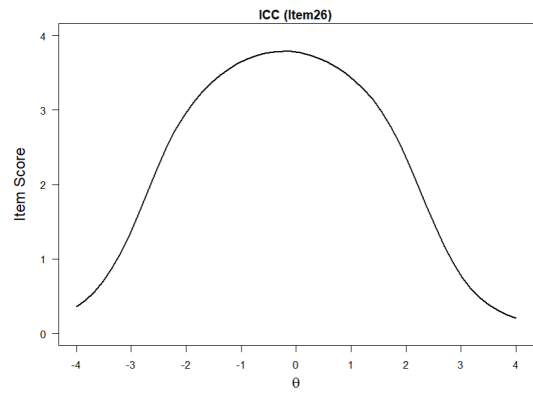
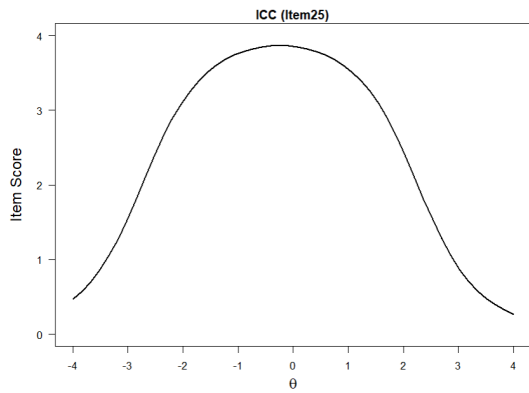
APPENDIX B. ITEM CHARACTERISTIC CURVES (ICCS)

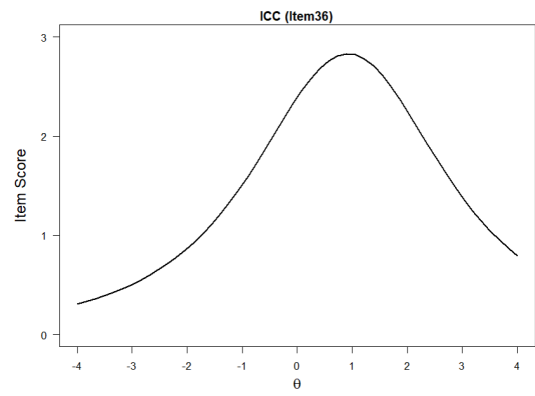
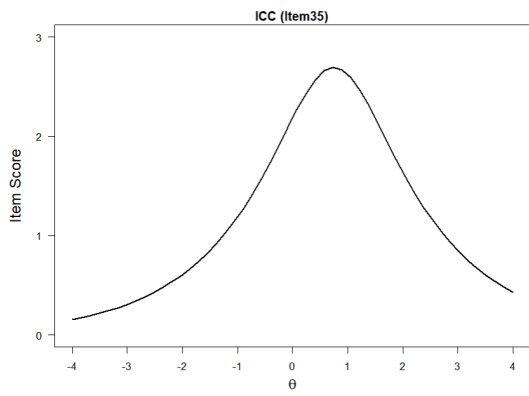
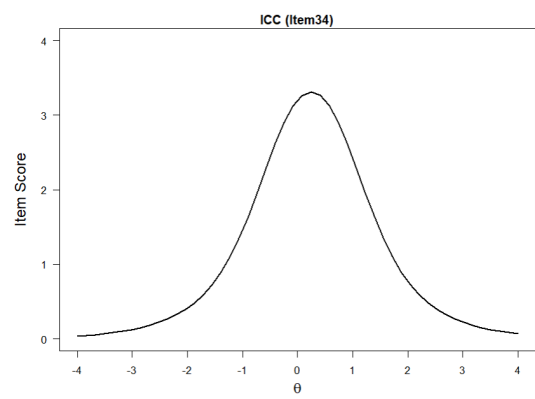
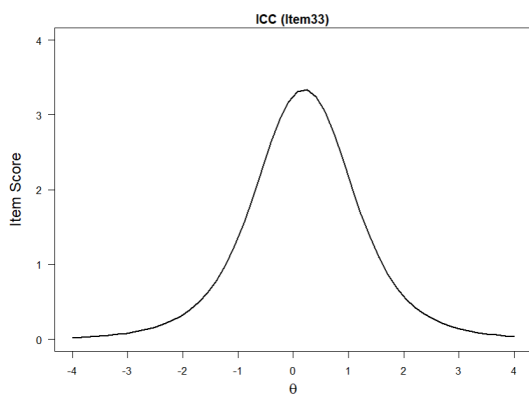
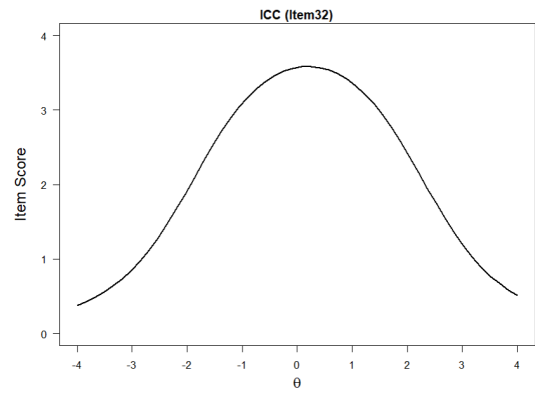
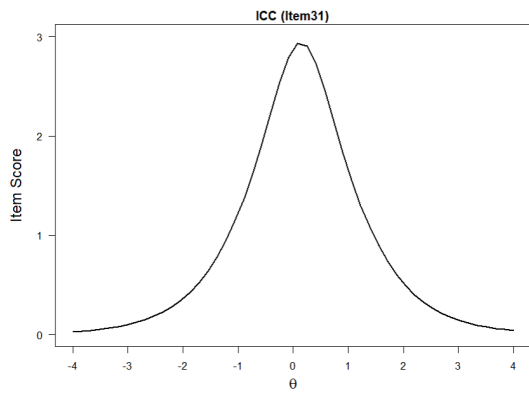


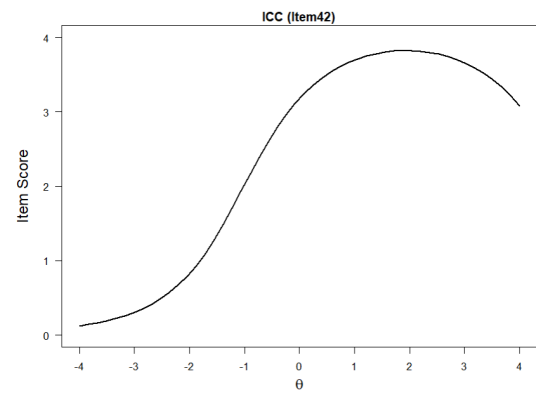
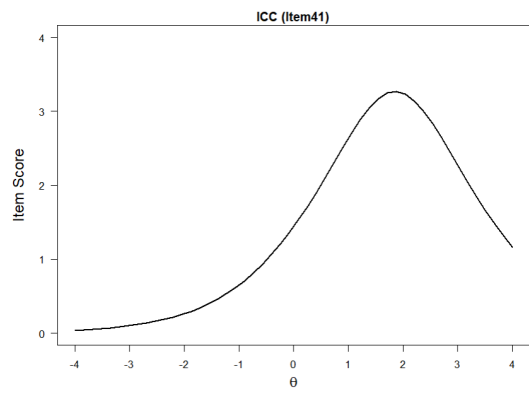
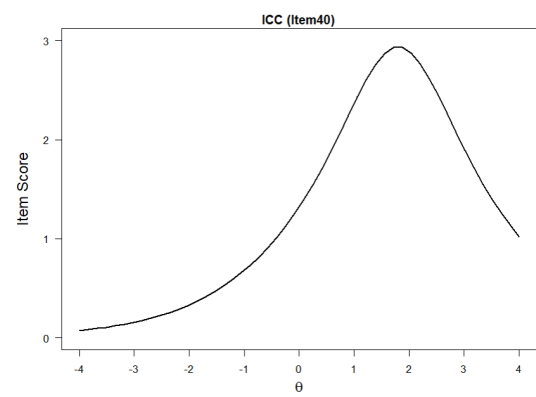
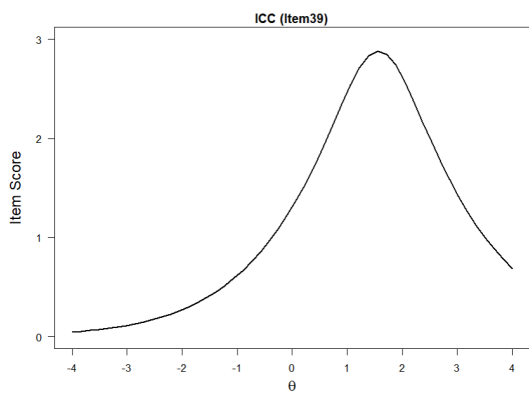
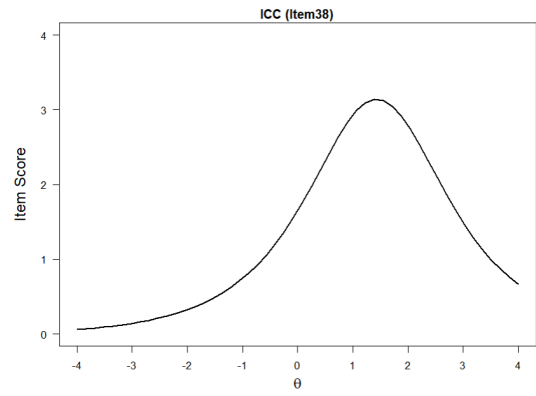
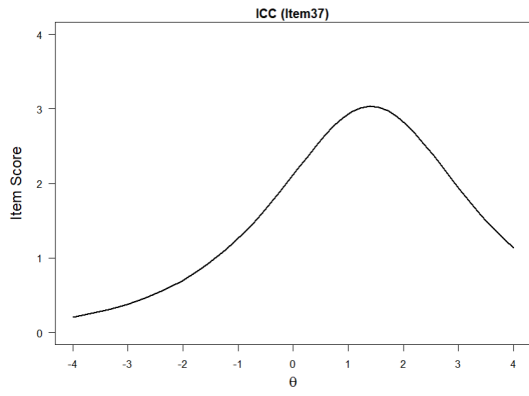


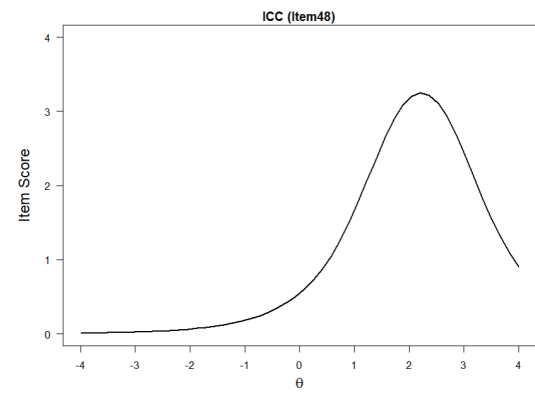
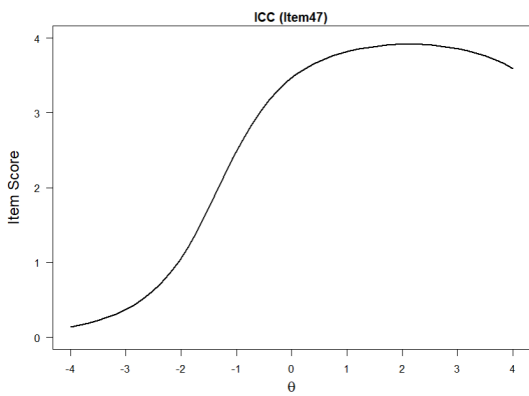
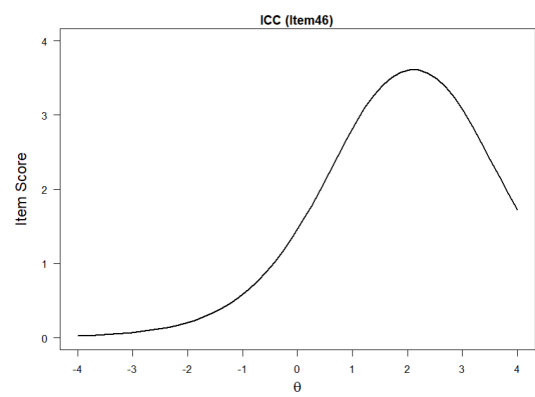
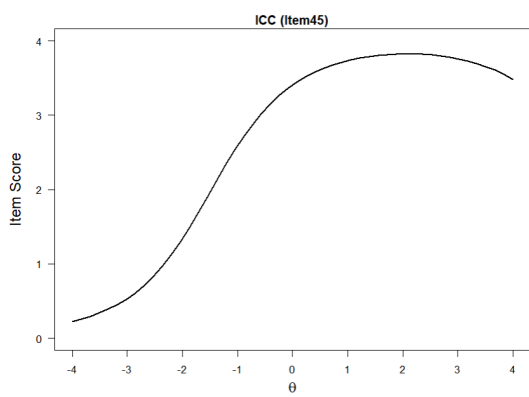
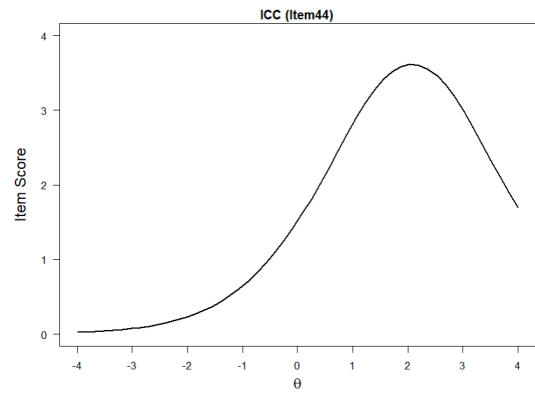
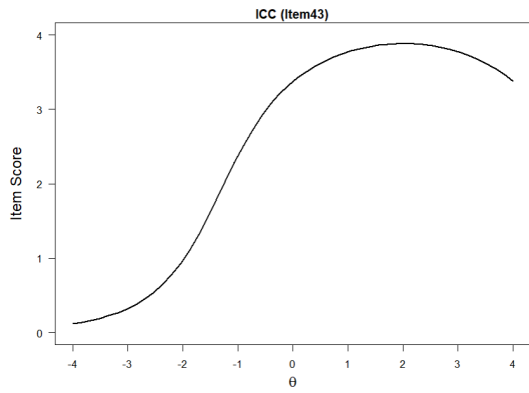


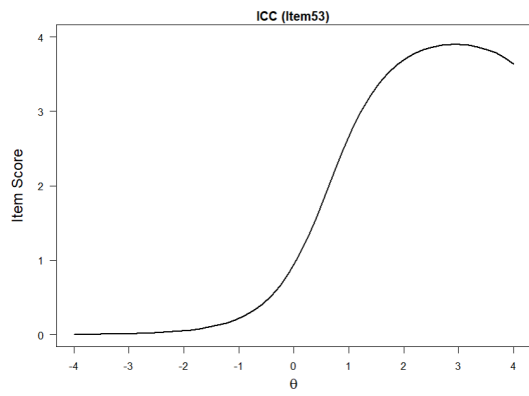
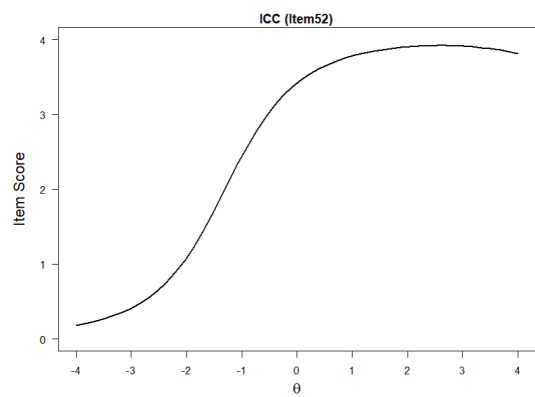
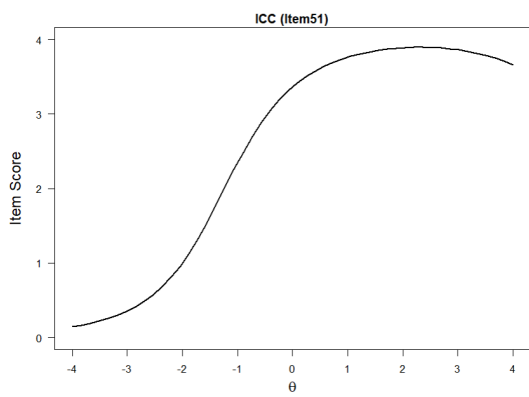
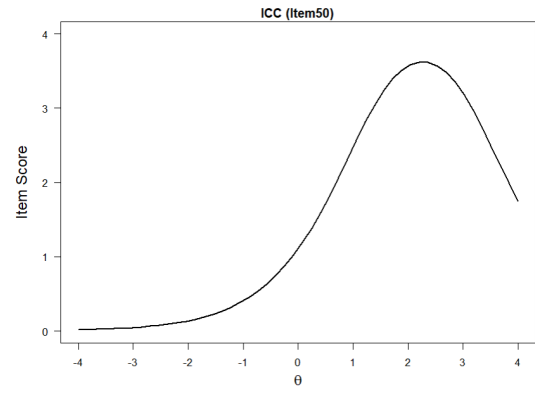
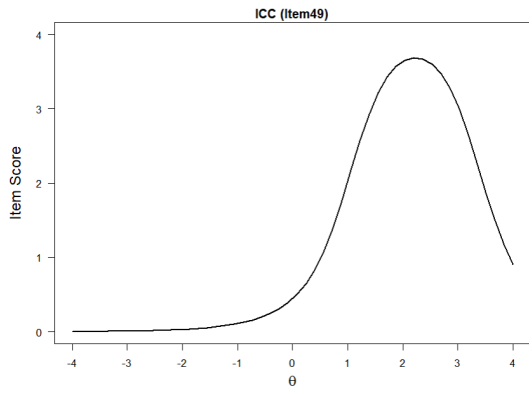












REFERENCES

- Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 26, 31-44.
- Andrich, D. (1996). A 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., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17, 253-276.
- Bartholomew, D. J. (1992). PARELLA: Measurement of latent traits by proximity items. *Journal of Classification*, 9, 161-163.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.
- Brooks, S. P., & Gelman, A. (1997). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Carroll, J. D. (1972). Individual Differences and Multidimensional Scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences* (Vol. 1). New York: Seminar Press.
- Carter, N. T., Dalal, D. K., Boyce, A. S., O'Connell, M. S., Kung, M. C., & Delgado, K. (2014). Uncovering curvilinear relationships between conscientiousness and job performance: How theoretically appropriate measurement makes an empirical difference. *Journal of Applied Psychology*, 99, 564-586.
- Cliff, N., Collins, L. M., Zarkin, J., Gallipeau, D., & McCormick, D. J. (1988). An ordinal scaling method for questionnaire and other ordinal data. *Applied Psychological Measurement*, 12, 83-97.
- Coombs, C. H. (1964). *A Theory of Data*. New York: Wiley.

- Crowe, M., Sleep, C., Carter, N., Campbell, W., & Miller, J. (2018). Self-esteem and narcissism: An item response theory analysis of curvilinearity. *Personality and Individual Differences, 128*, 16-20.
- Davison, M. L. (1983). *Multidimensional Scaling*. New York: Wiley.
- De Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in psychology, 10*, 102.
- de la Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement, 30*(3), 216-232.
- DeMars, C. E., & Erwin, T. D. (2003). Revising the scale of intellectual development: Application of an unfolding model. *Journal of College Student Development, 44*, 168-184.
- Ferrando, P. J., & Lorenzo-Sava, U. (2007). An Item Response Theory Model for Incorporating Response Time Data in Binary Personality Items. *Applied Psychological Measurement, 31*(6), 525-543.
- Fox, J. P. (2010). *Bayesian Item Response Modeling: Theory and Applications*: Springer New York.
- Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science, 7*, 457-511.
- Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE transactions on pattern analysis and machine intelligence, 6*, 721-741.
- Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw-Hill.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57*, 97-109.
- Hill, M. O., & Gauch, H. G. (1980). Detrended correspondence analysis: An improved ordination technique. *Vegetatio, 42*, 47-58.
- Hojtink, H., & Molenaar, I. (1994). An item response model with single peaked item characteristic curves: The PARELLA model. *Quality and Quantity, 28*, 99-116.
- King, D. R. (2017). *Stochastic approximation of the multidimensional generalized graded unfolding model with the Metropolis-Hastings Robbins-Monro algorithm*. (Doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.
- Kruskal, J. B., & Carroll, J. D. (1969). Geometric models and badness of fit functions. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (Vol. 2). New York: Academic.

- Kuncel, R. B. (1973). Response processes and relative location of subject and item. *Educational and Psychological Measurement, 33*, 545-563.
- Kuncel, R. B. (1977). The subject-item interaction in intermetric research. *Educational and Psychological Measurement, 37*, 665-678.
- Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika, 39*, 247-264.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325-337.
- Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58*, 445-469.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21*, 1087-1092.
- Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika, 51*, 177-195.
- Mislevy, R. J., & Sheehan, K. M. (1989). The role of collateral information about examinees in item parameter estimation. *Psychometrika, 54*, 661-679.
- Molenaar, D., Rózsa, S., & Kö, N. (2021). Modeling Asymmetry in the Time–Distance Relation of Ordinal Personality Items. *Applied Psychological Measurement, 45*(3), 178-194. doi:10.1177/0146621621990756
- Noel, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: Application to smoking cessation. *Psychological Methods, 4*, 173-191.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*(2), 146-178.
- Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.
- Post, W. J. (1992). *Nonparametric Unfolding Models: A Latent Structure Approach*: DSWO Press.

- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org>
- Revelle, W. (2019). *psych: Procedures for Personality and Psychological Research*. Northwestern University, Evanston. R package version 1.9.12
- Roberts, J. S. (1995). *Item response theory approaches to attitude measurement*. (Doctoral Dissertation). University of South Carolina, Columbia, SC.
- Roberts, J. S., Barrett, M. E., & King, D. R. (2016). *Measuring physical attraction with the multidimensional generalized graded unfolding model*. Paper presented at the International Meeting of the Psychometric Society, Asheville, N.C.
- 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*(1), 3-32.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimations in the generalized graded unfolding model. *Applied Psychological Measurement, 26*(2), 192-207.
- Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree scale. *Applied Psychological Measurement, 20*, 231-255.
- Roberts, J. S., Rost, J., & Macready, G. B. (2010). MIXUM: An Unfolding Mixture Model to Explore the Latitude of Acceptance Concept in Attitude Measurement. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 175-197). Washington, DC: American Psychological Association.
- Roberts, J. S., & Shim, H. (2010). *Multidimensional Unfolding with Item Response Theory: The Multidimensional Generalized Graded Unfolding Model*. Paper presented at the Psychometric Society, Athens, GA.
- Roberts, J. S., & Sparks, J. L. (2015). *Mapping the emotion space with the MGGUM*. Paper presented at the National Council on Measurement in Education, Chicago, IL.
- Roberts, J. S., & Thompson, V. M. (2011). Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model. *Applied Psychological Measurement, 35*(4), 259-279.
- Roskam, E. (1997). Models for Speed and Time-Limit Tests. In W. van der Linden & R. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 187-208). New York: Springer.

- Rouder, J., Province, J., Morey, R., Gomez, P., & Heathkote, A. (2015). The lognormal race: a cognitive- process model of choice and latency with desirable psychometric properties. *Psychometrika*, *80*, 491-513.
- Scheiblechner, H. (1979). Specifically objective stochastic latency mechanisms. *Journal of Mathematical Psychology*, *19*(1), 18-38.
- Smit, A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research Online*, *4*(3), 19-32.
- Storms, G., & Delbeke, L. (1992). The irrelevance of distributional assumptions on reaction times in multidimensional scaling of some same/different judgment tasks. *Psychometrika*, *57*, 599-614.
- Swaminathan, H., & Gifford, J. A. (1983). Estimation of Parameters in the Three-Parameter Latent Trait Model. In D. J. Weiss (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 13-30). New York: Academic Press.
- Thompson, V. M. (2014). *Marginal Bayesian parameter estimation in the multidimensional generalized graded unfolding model*. (Doctoral dissertation). Georgia Institute of Technology, Atlanta, GA.
- Thurstone, L. L. (1928). Attitudes can be measured. *The American Journal of Sociology*, *33*(4), 529-553.
- van der Linden, W. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287-308.
- van der Linden, W., Entink, R. K., & Fox, J. (2010). IRT Parameter Estimation with Response Times as Collateral Information. *Psychometrika*, *34*(5), 327-347.
- van der Linden, W., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195-210.
- van der Maas, H. L. J., Molenaar, D., Maris, G., Kievit, R., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: on the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*(2), 339-356.
- Van Schuur, W. H. (1984). *Structure in Political Beliefs: A New Model for Stochastic Unfolding with Application to European Party Activists*: CT Press.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., . . . Yutani, H. (2019). Welcome to the {tidyverse}. *Journal of Open Source Software*, *4*(43), 1686.

- Williams, E. (2017). *Estimation of parameters in the generalized graded unfolding model using a genetic algorithm*. (Doctoral dissertation). Georgia Institute of Technology,
- Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*, 281-294.