ESTIMATION OF PARAMETERS IN THE GENERALIZED GRADED UNFOLDING MODEL USING A GENETIC ALGORITHM

A Dissertation
Presented to
The Academic Faculty

by

Elizabeth J. Williams

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

Georgia Institute of Technology
December, 2017

ESTIMATION OF PARAMETERS IN THE GENERALIZED GRADED UNFOLDING MODEL USING A GENETIC ALGORITHM

Approved by:

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

Dr. Daniel Spieler
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. Brian Habing
Department of Statistics
University of South Carolina

Date Approved: August 23, 2017

TABLE OF CONTENTS

LIST OF TABLES **v**

LIST OF FIGURES **vi**

SUMMARY **vii**

Chapter 1. Introduction **1**
Background **1**
Unfolding Item Response Theory (IRT) Models **1**
Generalized Graded Unfolding Model (GGUM) **2**
GGUM Parameter Estimation 4
Genetic Algorithms (GA) **7**
Details of GA 9
IRT Parameter Estimation Using GA 15
Performance of GA 17
Objectives of the Current Study **17**

Chapter 2. Method **21**
Parameter Recovery **21**
Experimental Design 21
Data Generation 21
Parameter Estimation **23**
Starting Values 23
Prior Distribution 23
GA-EM Parameter Item Estimation Procedure 23
Person Parameter Estimation 29
Preliminary GA Testing 32
Software 32
Analysis 33
Empirical Data Analysis **35**

CHAPTER 3. Results **36**
Parameter Recovery **36**
Item Parameters 36
Person Parameters 44
Grid Search **50**
Empirical Data Application **52**

Chapter 5. Discussion **60**

APPENDIX A. Abortion Attitude Statements **67**

Appendix B. Real Data Estimation Standard Errors **72**

REFERENCES **76**

LIST OF TABLES

Table 1	Mean RMSD of parameters estimates for GA preliminary tests	37
Table 2	Mean RMSD of parameters estimates by condition	38
Table 3	η_w^2 values for ANOVA effects	39
Table 4	Mean RMSD of person parameters estimates by condition	46
Table 5	Mean η_w^2 values for ANOVA effect	47
Table 6	Mean count of local maxima by simulation condition	51
Table 7	GGUM item parameter estimates of abortion attitude statements	56
Table A.1	Abortion attitude statements	67
Table A.2	Ordered abortion attitude statements	70
Table B.1	GGUM item parameter standard errors from abortion attitude data	73

LIST OF FIGURES

Figure 1	Genetic Algorithm Iterative Process	9
Figure 2	GA-EM Iterative Process	29
Figure 3	Mean RMSD of $\hat{\alpha}$ across response categories	40
Figure 4	Mean RMSD of $\hat{\delta}$ across response categories	41
Figure 5	Mean RMSD of $\hat{\tau}$ across response categories	42
Figure 6	Mean RMSD of $\hat{\alpha}$ across test length	44
Figure 7	Mean RMSD of $\hat{\theta}$ across estimation method	48
Figure 8	Mean RMSD of $\hat{\theta}$ across test length	49
Figure 9	Mean RMSD of $\hat{\theta}$ across response categories	49
Figure 10	Mean Count for $\hat{\theta}$ at each level of response categories across test length	52
Figure 11	Histogram of $\hat{\theta}$ for the abortion attitude data	58
Figure 12	Average expected response and average observed response across values of (θ-δ).	59
Figure B.1	$\widehat{\delta_i}$ and SEs from abortion attitude data	74
Figure B.2	$\widehat{\theta_j}$ and SEs from abortion attitude data	75

SUMMARY

In the current study, a genetic algorithm was used in conjunction with the expectation-maximization algorithm to estimate parameters in a polytomous unfolding IRT model known as the generalized graded unfolding model (GGUM). One advantage of using a genetic algorithm for IRT parameter estimation is that this global optimization procedure is not easily affected by local maxima in the likelihood function – a condition that is often encountered in unfolding IRT models including the GGUM. Additionally, because genetic algorithms do not use derivatives to maximize the likelihood function, it is computationally simple and could be deployed efficiently with higher dimensional data. The focus of this study was to implement the genetic algorithm in the context of the GGUM, and then evaluate the speed and accuracy of the resulting parameter estimates Program development was done with the R computer language, and the efficacy of estimates was examined with simulation methods, which systematically vary sample size, test length and number of response categories. The resulting estimation strategy was also illustrated with real data from an abortion attitude questionnaire.

CHAPTER 1. INTRODUCTION

Background

Unfolding Item Response Theory (IRT) Models

Traditional IRT models work under the assumption that the latent trait level of the respondent is monotonically related to the probability of endorsing an item. These models can be described intuitively as "more is better models," and, in psychometrics, they are often referred to as cumulative models. These types of models most appropriately describe item response data that follow the assumption mentioned above (i.e., result from a dominance-based response process). These data can be routinely found in the measurement contexts involving academic proficiency, personality traits, and clinical diagnoses. Such models would yield monotonically increasing item characteristic curves (ICCs) and test characteristic curves (TCCs).

However, not all item response data conform to a cumulative model. There are other areas in psychology in which item responses generally follow from a proximity-based process (an ideal point response process). These areas include measurement of attitudes, preferences, and certain developmental changes that occur in distinct stages (Noel, 1999; Roberts, Donoghue, & Laughlin, 2000; Roberts & Laughlin, 1996; Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009). Thurstone's (1928) work is a classical illustration that implicitly presumes that responses to attitude questionnaires specifically follow from an ideal point process. Following Thurstone's seminal work, there have been various confirmations throughout the years that responses to Thurstone and Likert style attitude questionnaires do indeed follow this

process (Andrich, 1996; Roberts, 1996; Roberts, Laughlin, & Wedell, 1999; Van Schurr & Kiers, 1994). The measurement of the aforementioned psychological constructs is a frequently researched area within psychology. As a result, there have been different models proposed for item response data that follow from an ideal point process (Andrich, 1988; Andrich & Luo, 1993; Roberts et al., 2000). Coombs (1964) referred to models for ideal point responses as "unfolding models" to describe the geometric analogy for resolving the different preference orders given by different respondents to a common set of stimuli.

The notion behind ideal point processes is that a person will endorse an item to the degree that the person and the item are located near each other on the underlying latent trait continuum or latent space. In other words, the endorsement probability increases as the distance between an item location and a person's ideal point approaches zero, and the probability decreases as this distance increases in any direction. The ICC of a unidimensional unfolding item would have a peak (fold) at the point on the latent trait continuum where the person and item locations are identical. It is at this point that an ICC reaches it maximum value.

Generalized Graded Unfolding Model (GGUM)

The GGUM is a unidimensional unfolding IRT model for polytomous item responses (Roberts et al., 2000; Roberts & Laughlin, 1996). The underlying premise of the GGUM is the assumption that the data follow the proximity-based response process described above. That is, as the location of an item on the latent continuum approaches that of the individual examinee, then greater agreement (i.e., higher item scores) will be

exhibited. The models in the GGUM family have been successfully applied to the measurement of attitudes, emotion faces, and physical attraction (Roberts, Barrett, & King, 2016; Roberts et al., 2000; Roberts & Sparks, 2015). Additionally, the GGUM has been used in the industrial and organizational psychology domain to explore the measurement of personality traits such as conscientiousness and neuroticism (Carter et al., 2014; Drasgow, Chernyshenko, & Stark, 2010).

Within the GGUM framework, each observed response category (ORC) available to the examinee is considered a combination of exactly two subjective response categories (SRCs). To illustrate this idea, consider an observed response of "strongly disagree" on a Likert scale to an item located at the center of the latent trait continuum. Assuming a proximity-based response process, the examinee could respond "strongly disagree" to this item because they are located above, and far away from the item on the latent trait continuum, or because they are located below, and also far away from the item on the latent trait continuum. Therefore, as seen in Equation 1, the numerator of the GGUM is made up of exactly two terms (i.e., the two SRCs that correspond to a particular ORC). Additionally, the GGUM is a divide by total model, which means the denominator of the equation is simply the sum of all SRC numerator terms.

The GGUM is explicitly defined as follows:

$$P[Z_i = \mathrm{z}|\theta_j] =$$

$$\frac{exp(\alpha_i[z(\theta_j-\delta_i)-\sum_{k=0}^{z}\tau_{ik}])+exp(\alpha_i[(M-z)(\theta_j-\delta_i)-\sum_{k=0}^{z}\tau_{ik}])}{\sum_{w=0}^{C}[exp(\alpha_i[w(\theta_j-\delta_i)-\sum_{k=0}^{w}\tau_{ik}])+exp(\alpha_i[(M-w)(\theta_j-\delta_i)-\sum_{k=0}^{w}\tau_{ik}])]} \quad (1)$$

where

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

$z = 0, 1, 2, \ldots, C$; $z = 0$ 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 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} subjective response category threshold for the i^{th} item.

The value of τ_{i0} is defined as zero, and the remaining SRC thresholds are constrained to be symmetric about the item location. However, these thresholds are not constrained to be constant across items, and are not forced to be ordered.

GGUM Parameter Estimation

In practice, the first and most important step in applying IRT models to response data is that of estimating the model parameters. Therefore, parameter estimation is paramount in the application of the GGUM, and a variety of estimation techniques have been explored over the years. Early, simpler versions of the model (i.e., the graded unfolding model; GUM) were estimated using joint maximum likelihood (JML;Roberts & Laughlin, 1996). JML is a two-step maximum likelihood procedure that first uses starting values for the item parameters, and estimates person parameters. Second, the item parameter estimates are updated constraining the person parameters to the values from step one. Then, the process iterates back and forth between the two steps until a

convergence (stopping) criteria is met (Birnbaum, 1968). Like other applications of JML, the resulting GUM estimates were not consistent, meaning the estimates did not converge to the true values as the sample size increased. Additionally, the algorithm often became stuck at various local maxima in the likelihood function, and a grid search was required whenever there were indications that this may have occurred. Consequently, the method overall was computationally intensive.

The GGUM has also been estimated using the marginal maximum likelihood (MML) procedure (Roberts et al., 2000; Roberts, Donoghue, & Laughlin, 2002). To implement MML, a prior distribution was placed on the person parameters, allowing these parameters to be integrated out of the likelihood equation. The resulting item parameter estimates were very accurate when based on a moderately large sample size (N=750 to 1000). Additionally, it was found that at least 15-20 equally spaced items with six ORCs were needed to get reasonably accurate person parameter estimates.

Moving forward to a fully Bayesian estimation methods, the GGUM was estimated using marginal maximum a posteriori (MMAP) (Roberts & Thompson, 2011). In addition to MMAP, the GGUM has also been estimated using the Markov-chain Monte Carlo (MCMC) method (De La Torre, Stark, & Chernyshenko, 2006; Roberts & Thompson, 2011; Wang, de la Torre, & Drasgow, 2015). The MMAP procedure produced parameter estimates that were generally more accurate than the estimates produced by the MML or MCMC procedures. Specifically, these differences were greatest in experimental situations where the number of ORCs was small (i.e., 2 to 4). Because the MMAP procedure is more efficient than the MCMC procedure, it was recommended as the best estimation procedure for the GGUM at the time. However, it

should be noted that the MMAP procedure has been shown to have some shrinkage in parameter estimates, as does the MCMC estimation procedure, as is the nature of using prior distributions in any Bayesian estimation procedure. As for the MCMC procedure, the disadvantage beyond computational intensity is the convergence criteria for the joint distribution of parameters, in that there is no sharply defined rule for making this decision. A researcher must decide how many burn-in iterations are necessary to achieve a stationary joint posterior distribution from which parameter values can be sampled; there is no exact number of these iterations that is required. Additionally, in both procedures, knowledge of appropriate priors is necessary, and misspecification can result in inaccurate estimates should the data be uninformative.

While each of these methods suffice in most experimental estimation settings, there are points on which estimation of the GGUM can be improved to provide another alternative for practitioners to use when applying this model. In estimation conditions where the joint likelihood function of model parameters contains many local maxima, which is a common problem when modelling item response data using the GGUM, estimation procedures like MML, MMAP, and MCMC can converge quickly to one of the local maxima if informative start values are not used. This is particularly true with respect to person parameters as opposed to item parameters because there is substantially less information in the data about an individual than an item. Such local minima result in inaccurate parameter estimates. The only recourse for the researcher to fix this problem is to re-estimate parameters from another set of starting values that are closer to the global maxima or to implement a procedure that does not require maximization. Unfortunately, it is not always apparent that a local maximum has been achieved in the case of a

particular GGUM parameter when this critical point is near that for the global maximum. Additionally, simulation studies have shown that some of the current estimation methods for the GGUM need very good starting values for item parameters, even with the use of prior distributions in some cases, to yield accurate parameter estimates. Therefore, a new estimation method is needed for these specific GGUM estimation situations to give researchers greater confidence that the global maximum has been found. The strengths of a genetic algorithm described below may overcome these specific weaknesses of current estimation methods for the GGUM. The current study developed and evaluated a genetic algorithm (GA) estimation procedure for the GGUM.

Genetic Algorithms (GA)

GAs were first developed by John Holland as an optimization procedure (Holland, 1973). It was in this early work that the traditional theory of schema (similarity templates) was used as a piece of the explanation of GA performance. Simply put, schema theory states that all knowledge is organized into units. Within these units of knowledge, or schemata, is stored information. By considering individual solutions to an optimization problem as knowledge broken down into schema, and then applying the principles of genetic algorithms, Goldberg (1989) showed through schemata analysis that the best solutions receive at least exponentially greater opportunities in successive generations. However, in the last decade of GA research, evidence has accumulated that GAs don't necessarily work the way Holland first described. This has resulted in several different perspectives on GAs, none of which can claim to be the complete answer as to why GAs work so well in solving optimization problems (Reeves & Rowe, 2003). Yet, the fact still remains that GAs have been successful in solving optimization problems in a

wide range of applications; for example, automotive design, telecommunications networks, and traffic signal timing (Castellani & Franceschini, 2003; Ceylan & Bell, 2004; Dengiz, Altiparmak, & Smith, 1997).

Broadly, from its beginning, GAs combine survival of the fittest among solution structures with a structured, but randomized, information exchange to form a search algorithm (Goldberg, 1989). A successful GA efficiently exploits historical information to venture onto new search points with the expectation of improved performance (Reeves & Rowe, 2003). As will be shown in the detailed description below, these algorithms are computationally simple, yet powerful in their search for improvement. Additionally, they are not, at the core, limited by restrictive assumptions about the search space (e.g., assumptions concerning the existence of derivatives).

The differentiating qualities of a GA are what lead to its strengths, and thus the motivation for its application in the present study. There are three major qualities that set GAs apart from traditional optimization methods: (1) GAs search from a population of points, not a single point, (2) GAs use payoff knowledge in the form of a fitness function, not derivatives for example, and (3) GAs use probabilistic transition rules, not deterministic ones. Characteristic (1) above gives a GA the ability to perform a parallel search, which allows the algorithm to not be stuck in one location in the search space. Additionally, because of characteristic (2), GAs are computationally simple to understand and implement due to the use of a payoff function to direct the search. However, the GA is not without its faults. The time taken for a GA to converge is generally longer than other optimization methods. Moreover, while the computation of GA mathematics are relatively simple, there are many GA parameters that must be specified and fine-tuned

with each new optimization application. This leads to more of a trial and error process in applying the GA. To better describe each of these emergent aspects of a GA, the following sections present the details of a traditional GA

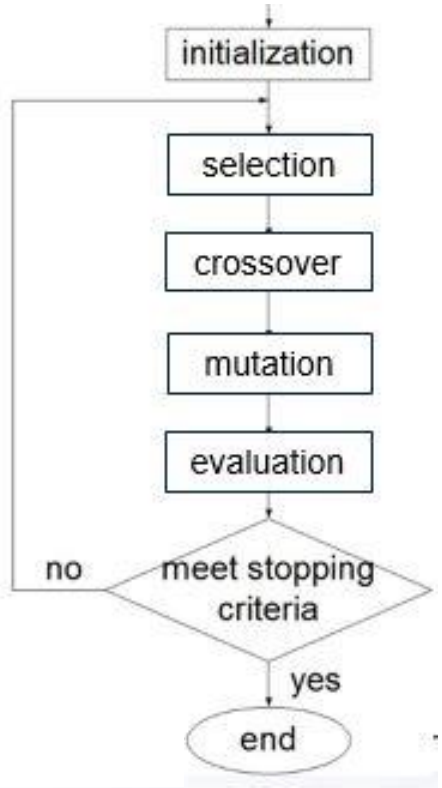

Figure 1. Genetic Algorithm Iterative Process

Details of GA

In general, GAs are iterative search algorithms based on the mechanics of natural selection and natural genetics. As an overview, a GA has four steps (1) initialization, (2) selection, (3) breeding, and (4) evaluation (Reeves & Rowe, 2003). Until the stopping criterion is met, steps 2-4 will be repeated iteratively as seen in Figure 1. When designing a GA, there are four main components of these steps that a researcher must consider: (a) a representation of potential solutions, (b) a method to create an initial population of

potential solutions, (c) an evaluation (fitness) function that plays the role of the environment, rating solutions in terms of their fitness, and (d) a set of well-defined genetic operators. These various components and how they relate to each step of a GA will now be described in more detail. To aid in understanding, GA estimation of the 3-PLM will be used as a very general running example, while specific empirical implementation of a GA in IRT is reviewed in the next section.

Initialization

A GA begins by specifying an initial population of potential solutions for the optimization problem at hand (Step (1) above). The initial population defines the areas of the solution space in which the GA will start its search. The GA maintains a population of solutions throughout the entire search process, which gives rise to the GA's strength in avoiding local maxima by conducting a parallel search. Therefore, the function of the initial population is to provide adequate coverage of the search space for a successful search to be achieved (Reeves & Rowe, 2003). To that end, there are user specified characteristics of initialization that the must be decided before beginning this process. These are: (1) population size, and (2) percentage of mutation (or, alternatively, cloning). Population size is the number of potential solutions (e.g., N=100) that the GA maintains and works with throughout the optimization process. The percentage of mutation (e.g., 20%) in an initial population helps to generate potential solutions to fill the aforementioned population of size N. The operational specifics of these two characteristics in achieving the goal of an initial population will be discussed next. But first, it should be noted that the values of these characteristics are largely specific to and dependent upon each individual optimization problem. Therefore, any explicit

recommendations with regard to the initialization of a GA should be taken as a starting point. For example, some empirical results indicate a population of size 30 are acceptable in some cases. However, later analyses led researchers to believe that there is a linear dependence between population size and the solution length (i.e., number of parameters to be estimated) (Goldberg, 1989). To more clearly illustrate this process, the initialization of a GA for IRT parameter optimization is described generally below.

First, the generation of an initial population begins with one candidate solution (CS). As mentioned above, if a GA is to be applied, then it must be possible to explicitly and simply represent this CS (i.e., a potential solution for the optimization problem) so that the GA can function optimally. In the case of IRT, this beginning CS would be the values of each item parameter for a given item *i* (e.g., CS 1 for item *i* would be $\beta_{i1} = (\alpha_{i1}, b_{i1}, c_{i1})$ for the 3-PLM). As you can see, IRT model parameter estimation contains this first GA component. Specifically, the IRT model parameters provide a clear representation of any solution.

Next in initialization, the CS is mutated according to the percentage of mutation specified at the start. That is, for example, 20% of the initial population will contain mutated solutions of β_{1i}. More specifically, in a population of size 100, the initial population will contain 20 mutated CSs, while the remaining 80 CSs in the population will be copies of the starting CS. The precise details of the mutation operator in a GA are described more fully later. For the purpose of initialization, it is most important to recognize that the function of this operator is to add variation to the search space. Consequently, the population size and the mutation percentage are specified to achieve

the appropriate level of differentiation the researcher needs in the areas of the search space. Once the algorithm has an initial population, it proceeds to the selection step.

Selection

In the selection step, CSs from the population are selected to survive to the next generation. The basic idea of selection is that it should be related to fitness (Reeves & Rowe, 2003). More specifically, the probability of being selected should be directly proportional to a CS's fit. In a GA, the fitness function should be defined to place value on the potential solutions in the population according to the GA's optimization goal. For example, when using a GA to solve a manufacturing optimization problem, the fitness function might be cost. Therefore, the candidate solution that minimized cost would be considered best "fitting". Conversely, when using a GA to solve a farming optimization problem, the fitness function might be crop yield. In this case, the CS that maximized crop yield (higher values of the fitness function) would be considered best "fitting".

After initialization of the population, the fitness function is calculated for each CS in the population. Then, the CSs are ranked based on their fitness function. In some GAs, the best (or top five best) candidate solution based on their fitness function value will automatically survive to the next generation; this is termed elitism. However, this is not a requirement of a GA. The GA description from this point forward will assume no elitism has been implemented.

Next, N CSs are selected with replacement to survive based on some user created probability based on these ranks, such that higher ranking equals a higher probability of selection, where N is the population size (Reeves & Rowe, 2003). While there is no

required definition of selection probability, it must achieve the goal of a GA's selection step: to reward better fitting CSs by having a higher chance of survival. The selection procedure and fitness function links the GA to the principle of survival of the fittest and the desired problem-specific optimization. As a final note on the selection procedure, it should be based on probability and not some deterministic rule (e.g., taking the top 50% best fitting solutions, and their copies). Deterministic rules like this undermine the randomness of the GA, and thus weaken the ability of a GA to continue an independent parallel search of the solution space.

Breeding

Following the selection step, there is a new population of size N that consists of a higher number of better fitting CSs than ill-fitting ones. At this point in the GA, new CSs (children) are created from these (parent) CSs using two genetic operators: (1) mutation, and (2) crossover (Reeves & Rowe, 2003). With either of these operators, the first decision point is whether mutation or crossover will occur at all. The researcher specifies two probabilities to select CSs for breeding; one for mutation (p_m), and another for crossover (p_c).

For the crossover operator, parent CSs are randomly selected from the population, without replacement, with probability p_c. These solutions are randomly paired together for crossover. Each pair of parent CSs generates two children, as described next. Consider this pair of 3-PLM item parameter solutions: $\beta_{i1} = (\alpha_{i1}, b_{i1}, c_{i1})$ and $\beta_{i2} = (\alpha_{i2}, b_{i2}, c_{i2})$. A crossover point is then randomly chosen. For example, if the crossover point was

after α_n, the new pair would be: $\beta_{iC1} = (\alpha_{i1}, b_{i2}, c_{i2})$, and $\beta_{iC2} = (\alpha_{i2}, b_{i1}, c_{i1})$. Finally, the pair of children then replaces the parents as candidate solutions.

Following the crossover procedure, the remaining CSs are randomly selected without replacement to undergo mutation. Selection for mutation is performed with a probability p_m, and is limited to those solutions that were not selected for crossover. The mutation operator is simpler than the crossover operator. Each selected parent CS is randomly altered according to some error probability distribution (Reeves & Rowe, 2003). More specifically, new CSs (children) are created by adding random noise (ε), where ε is randomly sampled from a standard normal distribution, for example. To illustrate, for the 3-PLM, if candidate solution 1 ($\beta_{i1} = (\alpha_{i1}, b_{i1}, c_{i1})$) is selected for mutation, then each parameter in this parent solution will have an independently and randomly sampled number added to it such that the child candidate solution is $\beta_{iC1} = (\alpha_{i1}+ \varepsilon_1, b_{i1}+\varepsilon_2, c_{i1+} \varepsilon_3)$. These mutated children survive to the next generation, while the parent CSs are discarded as before. Please note that in the initialization step, where 20% of the population were mutations of the first candidate solution, each mutated solution in the population was created using the random noise sampling described here.

At the end of the breeding step of a GA, the mutated children, crossover children, the candidate solutions that were selected to survive because of elitism, and any remaining CSs (that were selected, but did not breed) constitute the next generation population. At this point, the new generation is ready for evaluation. Through this step, the GA aims to have a new population of CSs that are approaching the global maximum by first selecting better fitting CSs at a higher rate, and then exchanging (changing)

characteristics of some of these CSs to produce offspring containing the best qualities of the parents.

Evaluation

In the evaluation step, the fitness function is calculated for each CS in the new population. The maximum value of the fitness function is recorded, and, more importantly, the CS associated with it. It is here that the stopping criterion for the GA is assessed. If the best fitting parameter values obtained at the end of the preceding iteration change by less than a pre-specified amount in the current iteration, then the GA stops. However, if the stopping criteria isn't met, then the algorithm continues (i.e., the steps above are repeated, beginning with selection from the current population) as shown in Figure 1 above.

As seen in the description above, GAs possess a number of strengths over traditional maximization methods such as parallel search, randomization of the search, and computational simplicity. Because of this, there are already a small number of researchers who have applied a GA to estimate IRT model parameters. In the next section, IRT model parameter estimation using a GA is discussed, as well as its performance and application thus far throughout the field.

IRT Parameter Estimation Using GA

IRT model parameter estimation is inherently an optimization problem. The nature of IRT model parameter estimation is to find the solution of parameter estimates that maximizes the likelihood of observing a particular set of response data. Two critical

components necessary for a GA to be applied are implicitly inherent in a typical IRT parameter estimation process. These two components are (1) an explicit representation of potential solutions (i.e., $\beta_{i1} = (\alpha_{i1}, b_{i1}, c_{i1})$), and (2) the likelihood function which can be used as a straightforward evaluation (fitness) function that plays the role of the environment, and rates solutions in terms of their fitness. While it seems that GA estimation of IRT model parameters is a natural extension of evolutionary optimization, GAs have only been used quite sparsely throughout the field of IRT. For instance, GAs have not been applied to polytomous IRT models, and have not been applied to unfolding models in the literature.

Despite the limited application of GA in IRT, there is empirical evidence that this method can be successfully applied in the above contexts. Specifically, GAs have been effectively applied to the 2-PLM and 3-PLM parameter estimation, albeit with a slight deviation from traditional GAs (Du & Chu, 2013; Jiang & Tang, 1998). In both cases, the GA was used at the maximization step of the E-M algorithm to significantly reduce the number of parameters estimated. By first integrating the person parameters out of the likelihood function (the E-step), the GA was then used to search for solutions for the item parameters, using the marginal likelihood function as the fitness function. Additionally, the marginal likelihood was maximized separately for each item parameter set due to the local independence assumption of IRT. The representation of candidate solutions, and the steps inherent in the GA are similar to the previous step-by-step example in that the steps are directly applied to the individual item parameter solutions. The next section describes the performance of GAs in IRT model estimation, as well as the benefits of using GAs above and beyond current IRT model estimation techniques.

Performance of GA

The extent of the use of GA in the IRT literature has been as calibration for generating start values for other estimation programs. It has been shown that the GA provides satisfactory results as a calibration method (Du & Chu, 2013; Jiang & Tang, 1998; Li, 1997). The results when using GA to get the starting points for use in BILOG are almost identical to the calibration results when using BILOG alone (Jiang & Tang, 1998). However, using a GA significantly reduces the number of EM cycles when input into that IRT estimation software. Additionally, when only the item characteristic curves or test characteristic curves are of primary concern, the GA alone provided acceptable results for most purposes. As noted in these studies cited above, the main disadvantage of the GA is computational speed. However, the benefit of using a GA in these instances, rather than rely on traditional estimation methods alone, lies squarely on the evolutionary programming strengths discussed previously. That is, because GAs are global optimization procedures, they won't be as easily susceptible to saddle points, or local maxima, in the likelihood function. Specifically, unlike a calculus hill-climbing method like Newton-Raphson, GAs allow the search of different areas of the solution space in parallel. Therefore, in situations where local maxima are an issue, the use of a GA could achieve a more successful search for the global maximum by being in two places at once.

Objectives of the Current Study

As an IRT model, when the GGUM fits the item responses, it offers researchers the advantages of person invariant interpretation of item parameters, item invariant interpretation of person parameters, and estimates of the standard errors of measurement

at the individual level. However, none of these benefits can be achieved if the model parameter estimation is not statistically sound or cannot be conducted in a reasonable time frame for most practical applications. Therefore, GGUM parameter estimation is an ever-evolving quest to realize statistically justifiable, accurate and computationally efficient results.

To this end, the goal of the current study was to build upon existing GGUM parameter estimation methods by exploring a potential solution to some of the issues plaguing all estimation methods currently available. First, the current study aimed to apply the GA-EM to GGUM parameter estimation and establish it as a procedure that results in estimates that are comparable to those explored in previous research (Roberts et al., 2000; Roberts & Thompson, 2011). Second, the current study sought to improve upon existing methods by developing a GA-EM estimation procedure of GGUM parameters that is not as easily fooled by local maxima, does not require a fully Bayesian solution for adequate parameter recovery (i.e., prior distribution for item parameters), and is computationally simple, without the need for calculation of complex derivatives. Third, it was expected that the computational speed would need to be sacrificed to achieve these goals, but that the potential benefits would outweigh this cost. Fourth, and finally, the current study sought to investigate what experimental conditions affect the performance of the GA, and therefore, determine the conditions in which it could be successfully applied.

The second point above deserves a bit more explanation in that previous marginal or fully Bayesian estimation methods applied for the GGUM obscured the fact that the GGUM likelihood is not generally single peaked. This was quite apparent in the work of

Roberts (1995) and Roberts and Laughlin (1996), which utilized joint maximum likelihood estimation of model parameters. Those researchers had to implement lengthy grid searches at various points within their algorithm to avoid local maxima. They reported that local maxima for both item locations and person locations were often encountered, but that the latter were more difficult to deal with using the Newton-Raphson maximization procedure. Specifically, local maxima for item locations tended to occur further away from the true item location and the log-likelihood at those local maxima was usually much smaller than that for the absolute maximum. Consequently, the Newton-Raphson procedure generally performed well as long as a judicious choice of start values was made. On the other hand, the local maxima for person locations typically occurred much closer to the corresponding global maxima, and the log-likelihood in these cases was often quite similar to the maximum log-likelihood. The Newton-Raphson procedure did not perform well in those circumstances, so they implemented a slow and tedious grid search for optimal person locations within the JML procedure.

More recent GGUM estimation algorithms that rely on marginal or fully Bayesian methods tend to skirt the issue of local maxima (Roberts et al., 2000; Roberts & Thompson, 2011; Thompson, 2014). That is somewhat justifiable for two reasons. First, anecdotal evidence suggests that the addition of single-peaked prior distributions mitigates the occurrence of local maxima to at least some extent (J.S. Roberts, personal communication, Spring, 2017). Second, all of the marginal or fully Bayesian estimation procedures used with the GGUM thus far have implemented expected a prior (EAP) estimates of person locations in which only the mean of the posterior likelihood for a

given person parameter, rather than the mode, is calculated. Thus, with respect to person parameters, the posterior likelihood has not been maximized and the issue of local maxima is moot. Whether this state of affairs is tolerable depends on the value of estimates based on the mode of the posterior likelihood of θ_j (i.e., the MAP estimate). For example, the MAP estimate of θ_j, like the EAP estimate, will always exist regardless of the person response vector in question (Baker, 1987). Moreover, the mode, rather than the mean, may be a more meaningful estimate when the posterior distribution of θ_j is skewed, and this will generally occur to some extent until the number of questionnaire items grows large. Thus, researchers may prefer the MAP over the EAP estimate in situations with smaller numbers of items. In such cases, local maxima in the log-likelihood for θ_j may lead to inaccurate results, and the benefits of the GA algorithm could be substantial.

Finally, it is worthwhile to note that some researchers may prefer maximum likelihood estimates (MLEs) of θ_j conditioned on item parameters obtained from a marginal Bayesian procedure. Those estimates may also be adversely affected by any local maxima that exist in the log-likelihood for θ_j, and thus, the GA would be advantageous in those contexts as well.

CHAPTER 2. METHOD

Parameter Recovery

Experimental Design

A simulation study was performed to assess the accuracy of parameter recovery when using the GA-EM to estimate the GGUM. The simulation study investigated the effects of three factors on parameter recovery. The three factors were (a) sample size (N=750, 1000, 1250), (b) test length (I=10, 20, 30), and (c) observed response categories (ORC=2, 4, 6). These three factors were fully crossed, resulting in a 3x3x3 experimental design. The simulation study had 10 replications in each of the resulting 27 cells.

Data Generation

The data was generated using the Generalized Graded Unfolding Model (GGUM) (Roberts et al., 2000). The GGUM is defined as follows:

$$P[Z_i = \mathrm{z}|\theta_j] =$$

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

where

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

$z = 0, 1, 2, \ldots, C$; $z = 0$ 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,

$M = 2C + 1$,

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} subjective response category threshold for the i^{th} item. Note that τ_{i0} is defined as zero in this model.

True Parameter Values

The true item parameter values were randomly sampled with replacement from a list of unidimensional GGUM parameter estimates for abortion attitude items described by Thompson (2014). The item parameter estimates from Thompson (2014) were separated into five intervals that span the latent trait continuum, and true item parameter values for this simulation were equally sampled, with replacement, from each interval. This strategy promotes realism with respect to the item population and the correlations among item parameters. The true person locations were sampled from a normal distribution with zero means and unit (1) variances, as the usual assumption is that examinee latent trait values are normally distributed.

Response Generation

Using the GGUM defined in Equation 2, an I x (1+C) matrix of item response probabilities was obtained for each individual in a given replication within the experimental design. The obtained category response probabilities for a given item

formed a multinomial distribution from which observed item responses were randomly sampled.

Parameter Estimation

Starting Values

The starting value for the item discriminations (α_i) were set to one for all items. The starting values for item location parameters (δ_i) were computed from a detrended correspondence analysis (DCA) of the simulated item responses in which only the first dimension was retained (Hill & H. G. Gauch, 1980). The starting values for thresholds (τ_{ik}) were obtained using a regression equation that was developed from previous MMAP estimation of the GGUM (King, 2017; Roberts & Thompson, 2011).

Prior Distribution

In this technique, person locations (θ_j) were integrated out of the likelihood function by specifying a prior distribution, $g(\theta)$, and then integrating over this distribution using numerical quadrature approximation. A standard normal distribution, $N(\mu=0, \sigma^2=1)$, was used as the prior distribution for θ_j. This prior distribution is traditionally used in EM approaches to parameter estimation (Baker, 1987; Bock & Aitkin, 1981).

GA-EM Parameter Item Estimation Procedure

Expectation

The GA was applied at the maximization step of the E-M algorithm, following the previous applications of GA to IRT model parameter estimation. To begin, at the

expectation step of the GA-EM, the expected frequency of response *z* for item *i* at quadrature point V_f was calculated using the following equation:

$$\bar{r}_{izf} = \sum_{s=1}^{S} \frac{H_{siz} r_s L_s(V_f) A(V_f)}{\widetilde{P_s}} \quad (3)$$

where

$$L_s(V_f) = \prod_{i=1}^{I} P(Z_i = x_{si} | V_f) \quad (4)$$

and

$$\widetilde{P_s} = \sum_{f=1}^{F} L_s(V_f) A(V_f) \quad (5)$$

where

s=1,…,S; S=total number of examinees,

f=1,…,F; F=total number of quadrature points,

$L_s(V_f)$= the conditional probability of response vector X_s at V_f,

$\tilde{P}_s$=the marginal probability of response vector X_s,

$A(V_f)$=the height of quadrature points V_f,

H_{siz}=dummy variable that is equal to 1 only when $z=x_{si}$.

$\bar{r}_{izf}$=expected frequency of response *z* for item *i* at V_f

The counts in matrix $\bar{r}_{izf}$ give the number of persons who are expected to be located at quadrature point X_f , and use category z in response to item *i*. These expected counts were calculated using the observed responses and the current item parameter estimates. In the subsequent maximization step, they were treated as known constants when calculating the log-likelihood equations, and then item parameters were solved for one item at a time. This reduces the computational burden of the log-likelihood calculation within the maximization step.

Step1: Initialization

For ease of understanding, the details of the GA are described for one item, although the procedure is identical for every item. To begin the maximization step and implementation of the GA, an initial population of size 100 was created from one starting solution. A candidate solution was represented as follows: β_{is}= (α_{is}, δ_{is}, and τ_{is1}, τ_{is2}, …,τ_{isC}). Of these 100 CSs, 20% were mutated solutions, while the other 80% were clones of the starting solution. The mutated solutions were created according to the mutation operator described below.

Step 2: Selection

The log-likelihood equation was evaluated for each CS according to the following equation, similar to Jiang and Tang (1998) and (Zhang, 2005):

$$Q_{is}(\beta_{is}) = \sum_{f=1}^{F} \sum_{z=0}^{C} \log(P_{isz}(V_f; \beta_{is})) * \bar{r}_{izf} \quad (6)$$

Equation 6 illustrates how the expected number of persons at a given quadrature point who use a particular response category for the *i*th item contributes to the marginal log-

likelihood of responses to that item. CSs were then ranked in ascending order according to their values of Q_{is}. To begin selection, the best fitting CS (rank 100) survived to the next generation. The remaining 99 CSs of the new population were selected, with replacement, from the current population according to roulette wheel selection, where the probability of selection is directly proportional to the rank. These 99 CS solutions constitute the selected population to begin this iteration of the GA.

Step 3: Breeding

The probability of mutation and crossover were 0.7 and 0.3, respectively, which are in line with common values from the literature. The crossover operator was conducted first. A random number from a uniform distribution (0, 1) was drawn for each of the remaining 99 CSs in the population to decide which CSs were selected for crossover. Once selected, the CSs were paired up, a random crossover point was selected, and two children were created using the scheme presented in Chapter 1. The parent CSs were then discarded.

Next, the mutation operator was applied to the remaining CSs in the population that were not selected for crossover. Once selected for mutation, each solution was altered by adding random noise to the parameter values. To demonstrate, consider the CS solution β_{i1}= (α_{i1}, δ_{i1}, and τ_{i11}-τ_{i13}) for a 4 ORC GGUM item. Independent random numbers were drawn from uniform distributions and added to each of the 5 parameters. The uniform distributions for δ_i and τ_1-τ_3 all ranged from -0.25 to +0.25. However, the mutation process for the α_i parameter was altered by sampling from a censored uniform distribution so that it was above the acceptable threshold (i.e., greater than zero). Once all

selected CSs were mutated, the children were saved to the new generation, and the parents were discarded. As described in the general GA steps above, the children of crossover, the mutated solutions, the elite solution that was saved from the selection step, and any remaining solutions that were not selected to breed made up the next generation of candidate solutions (totalling 100).

Step 4: Evaluation

Finally, the log-likelihood equation was evaluated for each CS in the next generation using Equation (6). The solution that yielded the maximum marginal log-likelihood function value in this population was retained and compared to the best fitting solution from the beginning of this iteration. If any of the parameter values for this item changed more than 0.0005, then steps 2 through 4 were repeated until this stopping criteria is met.

General Notes

It is conceptually simple to think of the GA-EM algorithm as two nested loops – the outer and inner loops. At the beginning of the outer loop, the values of $\bar{r}_{izf}$ were calculated, and then the inner loop was executed. Within the inner loop, the four GA steps described above were performed for each parameter type separately, while holding the other parameter types constant. The order of parameter estimation for a given item was item thresholds, followed by item locations, and then item discriminations. Each of these separate GAs, of which there are three, were performed in sequence within the inner loop. Because the maximum marginal log-likelihood associated with a given parameter depends on the values of other parameters for that item, the inner loop iterated

until the stopping criteria was met. Specifically, once each inner loop was finished, if any of the parameter estimates changed substantially from their respective values at the beginning of the loop, then then the inner loop iterated and the sequence of three GAs began again. The inner loop iterations continued until the change in any item parameter estimate was less than .0005. Once this criterion was met, then the parameters for the next item were calculated in an analogous fashion within the inner loop. At the conclusion of the inner loop for all items, control was passed back to the outer loop where the values of $\bar{r}_{izf}$ were recalculated with the most recent item parameters. The cycle of outer loop and inner loop updated continues until no item parameter change more than .0005 across successive iterations of the outer loop. A visual representation of this process can be seen in Figure 2.

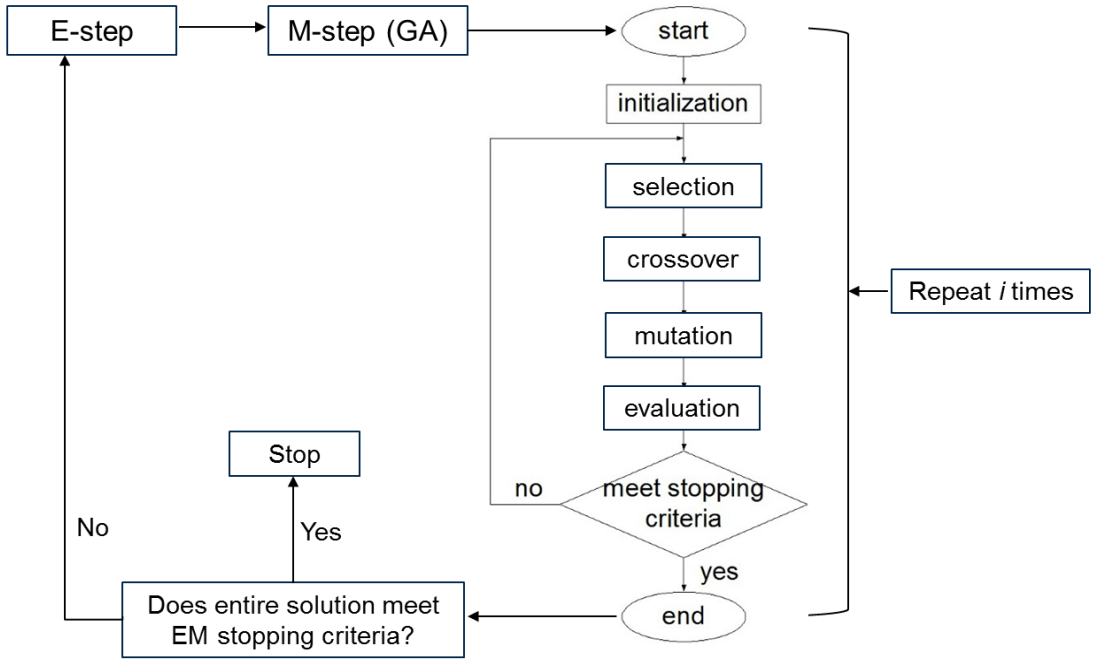

Figure 2. GA-EM Iterative Process

Person Parameter Estimation

The GA-EM estimates of the item parameters as well as the observed responses were used to calculate θ parameter estimates. These estimates were obtained using each of five procedures which included an expected a posteriori (EAP) procedure whereby the conditional mean of the individual's posterior distribution of θ_j was found, a maximum a posteriori (MAP) procedure in which the mode of the posterior likelihood was found with GA, a MAP procedure in which the mode of the posterior likelihood was found with a common Fisher scoring technique, a maximum likelihood (MLE) procedure in which the maximum was found with the GA, and an MLE procedure in which the mode was found

with the Fisher scoring method. Therefore, there were five estimates of θ_j produced for each individual examinee on a given replication.

Preliminary GA Testing

As mentioned above, there are specific parameters that define any particular GA. These parameters are optimization problem specific, and thus, have no predefined values. To test the appropriateness of the parameter values used in the GA-EM in this study (described above), five additional preliminary simulation tests were conducted.

In the simulation conditions described below, only the changes to the GA-EM detailed within each condition were implemented. Therefore, unless otherwise noted, all other GA-EM specifics that were presented above were not changed; that is, they were held constant across levels of the GA-EM parameter being investigated. Additionally, for each of the five tests, 10 replications of response data with six response categories, 30 items, and 1250 examinees were generated as described above. (i.e., a total of 50 independently simulated datasets). This constitutes a within-replications design for each test. Therefore, there will be five independent one-way within-replications ANOVAs performed on the five different sets of 10 replications of data. Please note that the factors manipulated in each of the five tests were not fully crossed. Finally, within each of the first four tests, one level of the manipulated factor constitutes the GA-EM described above as a control.

In the first test, the population size for the GA-EM item parameter estimation was varied in an attempt to investigate any improvement of parameter recovery performance by having a larger population of CSs or improvement of computational speed by having a

smaller population of CSs. The three levels of CS population size were (1) 50, (2)100, and (3) 200.

For the second test, the crossover probability was varied for item parameter estimation. There were three levels of crossover probability: (1) 0.1, (2) 0.2, and (3) 0.3. One concern with the current form of the GA-EM used in this study was that there might be too much variance in the CSs within each iteration, resulting in too random of a search. Therefore, only smaller crossover probabilities than the probability used in this study were tested.

In the third test, the factor that was manipulated was mutation probability. The four levels of this factor were (1) 0.05, (2) 0.1, (3) 0.25, and (4) 0.7. The choice to test smaller mutation probability was made to address any issue there might be with too much variance in the CS population during one iteration of the GA-EM.

The fourth test investigated the influence of the width of the uniform distribution from which random noise is selected when the mutation operator is applied. Specifically, the upper and lower bounds of the uniform distribution for mutating δ_i were manipulated. The motivation for only examining the effects of this factor on δ_i was because initial tests of the GA-EM yielded comparable parameter recovery with respect to δ_i, but not at the level that GGUM estimation simulation studies suggest. Therefore, by reducing the amount of random noise added to the mutated δ_i parameters, parameter recovery might increase as a result from smaller jumps in estimated δ_i parameter values within an iteration of the GA-EM. The three levels of this factor were (1) $\pm$0.05, (2) $\pm$0.1, and (3) $\pm$0.25.

For the final test of the GA-EM, all parameter types were estimated within the same GA. Additionally, the mutation operator was applied at the parameter level, instead of the level of the CS. These changes were made to mimic other common GA forms found outside of IRT parameter estimation. To better illustrate the change in mutation, consider the example CS from Chapter 2: $\beta_{i1}= (\alpha_{i1}, \delta_{i1}$, and $\tau_{i11}, \tau_{i12}, \ldots,\tau_{i1C})$ for the GGUM. During an iteration of the GA described above, each CS is chosen to mutate independently based on some probability (see mutation operator section above). If CS $\beta_{i1}= (\alpha_{i1}, \delta_{i1}$, and $\tau_{i11}, \tau_{i12}, \ldots,\tau_{i1C})$ is chosen to mutate in the τ estimation inner subloop, then every τ parameter is altered with random noise, which results in : $\beta_{i1}= (\alpha_{i1}, \delta_{i1}$, and $\tau_{i11}+ \varepsilon_1, \tau_{i12}+ \varepsilon_2, \ldots,\tau_{i1C}+ \varepsilon_k)$. However, when the mutation operator is applied at the parameter level, each parameter in each CS is evaluated separately according to the mutation probability, and thus, not all parameters within a CS are required to mutate together. In this method, it could be the case, for example, that only τ_{i11} is mutated in $\beta_{i1}= (\alpha_{i1}, \delta_{i1}$, and $\tau_{i11}, \tau_{i12}, \ldots,\tau_{i1C})$, resulting in a child CS $\beta_{i1}= (\alpha_{i1}, \delta_{i1}$, and $\tau_{i11}+\varepsilon_1, \tau_{i12}, \ldots,\tau_{i1C})$. In addition to estimating the parameter types at the same time, and mutating at the parameter level, the mutation probability was varied. The four levels of mutation probability were (1) 0.03, (2) 0.05, (3) 0.07, and (4) 0.1. To reiterate, these probability values were chosen to investigate whether reduction in variance in the CS population would increase parameter recovery.

Software

The GA-EM estimation and analysis will be implemented using the R statistical software package. Other than the EAP, MAP, and MLE estimation of person parameters, all other estimation algorithms were programmed specifically for this study by the author.

Analysis

In this study, parameter recovery was evaluated using the familiar root mean square deviation (RMSD) statistic, which is an index of the average discrepancy between estimated and true parameter values. RMSD for a given parameter type in the GGUM is defined as:

$$RMSD = \left(\frac{\sum_{t=1}^{T}[\widehat{\gamma_t}-\gamma_t]^2}{T}\right)^{1/2} \qquad (7)$$

where

$\widehat{\gamma_t}$= 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 (e.g., T=I for α or δ parameters; T=I*C for τ parameters; and T=N for θ parameters)

Within a given replication, RMSD was computed for each item parameter type, as well as each type of θ estimate produced.

For the GA preliminary tests, a one-way within subjects ANOVA was conducted for each test to examine the effects of the manipulated factor on the on the RMSD of item parameters. Similar to the primary study analysis, the RMSD was examined for each item parameter type (i.e., α_i, δ_i, and τ_{i1} to τ_{iC}) using three identical ANOVA models. The Type I error rate for each of these three ANOVA models was set to .05/3 = .0167.

In addition to a descriptive interpretation of RMSD for the larger simulation study, a three-way between replications ANOVA was computed to examine the effects of the three manipulated factors (i.e., sample size, test length and number of response categories) on the RMSD of item parameters. Similar to the above analysis, the RMSD was examined for each item parameter type (i.e., α_i, δ_i, and τ_{i0}-τ_{iC}) using three identical ANOVA models. In contrast, a split-plot ANOVA was conducted for the RMSD of alternative estimates of θ_j. The same three between replication factors were included in this split-plot ANOVA, and the type of θ_j estimation method used constituted the sole within replications factor. All main effects and interactions were entered into every ANOVA model. Moreover, the Type I error rate for each of the four ANOVA models was set to .0125 = .05/4. The power to detect even small effects in simulations such as the one proposed here is generally quite high. Therefore, an effect size estimate denoted as η_w^2 (Roberts & Thompson, 2011) was used to determine the largest effects in each ANOVA model. The η_w^2 index indicates the proportion of sums of squares within a family of effects tested by the same error term which can be attributed to a given effect within that family. As such, η_w^2 is like a traditional η^2 except that it decomposes within-family, rather than total, sums of squares. Across all ANOVA models, a given effect warranted interpretation only if it was both statistically significant and had a η_w^2 value greater than or equal to 0.10.

Although RMSD was the primary dependent variable in this study, a second set of ANOVA models analogous to those described above was conducted using a count of the number of local maxima seen in the corresponding likelihood for a given parameter. In the case of item parameters, this count was formed by examining the marginal likelihood

for each item parameter. A grid search was conducted to identify the number of times that 1) a critical point in the marginal likelihood occurs and 2) the pattern of the marginal likelihood to the left and right of that point indicates it is either a local or global maximum. This count minus 1 enumerated the number of local maxima in the likelihood. (The global maximum constitutes the largest of the local maxima.) Similar counts were constructed for θ_j. These counts were calculated based on the likelihood of θ_j.

Empirical Data Analysis

In addition to the simulation study described above, the GA-EM was applied in an empirical data set in an effort to assess the applicability and interpretability of parameter estimates derived from this procedure. The GA-EM was used to estimate GGUM parameter estimates from responses to attitude statements about abortion. These data are composed of responses to 40 statements obtained from approximately 1,500 college students, and all statements used a six point graded response scale. The GA-EM was implemented using the algorithm specifics that are described above.

CHAPTER 3. RESULTS

Parameter Recovery

Item Parameters

Following the GA-EM estimation of item parameters, interpretable effects were identified after analysing the mean RMSD using the ANOVA models described in Chapter 3 along with and the corresponding η_w^2. Before calculating RMSD values, it was necessary to match the proper signs corresponding to a particular end of the latent trait continuum. This is because the DCA process of assigning start values for δ_i occasionally reversed the poles of the dimension such that, for example, positive estimates corresponded with negative true generated values. It should be noted that this does not affect the GGUM likelihood other than reversing the sign of the domain for the corresponding parameters (i.e., location parameters δ_i and θ_j)

GA Preliminary Tests

Table 1 presents the mean RMSD for each parameter type within each test described above. There were no interpretable effects within any of the tests using the corrected Type I error rate, as evidenced by the similar mean RMSD seen in Table 1. It can be seen, descriptively, that the mean RMSD for each parameter type was higher in the fifth test than any of the other four tests. In addition to the combination of all parameter estimation into one GA-EM, test five had significantly lower mutation probabilities. Based on the results of this fifth test, decreasing the variance in the CS population to this degree does not result in acceptable parameter recovery. The

statistically insignificant results of the first four tests show that none of these factors, varied as they were here, improve parameter recovery beyond the GA-EM used in the larger simulation study. Therefore, the parameter values for the GA-EM described in Chapter 2 were sufficiently appropriate for use.

Table 1. Mean RMSD of parameters estimates for GA preliminary tests

Condition	$\hat{\delta}_i$	$\hat{\alpha}_i$	$\hat{\tau}_i$
Test 1: Population Size			
50	0.3822	0.15872	0.40672
100	0.37651	0.15878	0.3996
200	0.36768	0.15942	0.38991
Test 2: Crossover Probability			
0.1	0.37763	0.15628	0.40135
0.2	0.37907	0.15862	0.40983
0.3	0.37651	0.15878	0.3996
Test 3: Delta Distribution Bounds			
2	0.37701	0.15999	0.40007
4	0.37618	0.15999	0.39952
6	0.37651	0.15878	0.3996
Test 4: Mutation Probability			
0.05	0.37518	0.16114	0.4005
0.1	0.36249	0.15975	0.38794
0.25	0.37518	0.16114	0.4005
0.7	0.39051	0.15878	0.3996
Test 5: Mutation Probability (Combined GA)			
0.03	2.40075	0.52577	0.73549
0.05	2.42126	0.52452	0.74947
0.07	2.40437	0.52809	0.73652
0.1	2.42982	0.52862	0.7444

GA-EM Simulation

The mean RMSD of all parameter estimates across all conditions were equal to $\hat{\alpha}$=0.292, $\hat{\delta}$=0.3719, and $\hat{\tau}$=0.366. Table 2 displays the mean RMSDs in the factorial design conditions, while Table 3 portrays the results from the ANOVAs in the form of statistically significant effects and effect sizes (i.e., η_w^2).

Table 2. Mean RMSD of parameters estimates by condition

Factorial Condition	$\hat{\delta}_i$	$\hat{\alpha}_i$	$\hat{\tau}_i$
Sample Size			
750	0.36464	0.28346	0.36032
1000	0.38812	0.28763	0.38142
1250	0.36314	0.30661	0.35884
Test Length			
10	0.38816	0.37818	0.38376
20	0.37202	0.25154	0.36384
30	0.35536	0.24753	0.35267
Response Category			
2	0.43003	0.43114	0.38283
4	0.317	0.26303	0.32592
6	0.36809	0.18326	0.39122

Table 3. η_w^2 values for ANOVA effects

Effect	$\hat{\delta}_i$	$\hat{\alpha}_i$	$\hat{\tau}_i$
Sample Size	1.43%	0.39%	1.70%
Test Length	2.13%	**13.40%**	2.84%
Response Category	**24.76%**	**38.68%**	**14.19%**
Sample Size x Test Length	2.56%	1.00%	3.56%
Sample Size x Respsonse Category	1.49%	0.51%	1.61%
Test Length x Response Category	2.07%	**6.06%**	1.93%
Sample Size x Test Length x Response Category	**7.57%**	2.43%	**7.57%**

Note: Bolded values were statistically significant (p<0.0125). Values bordered in red were interpretable (η_w^2>0.10).

The main effect of response category was considered to be interpretable for all estimated item parameter types. This is similar to previous GGUM simulation study results (Roberts & Thompson, 2011). There was a decrease in mean RMSD for $\hat{\alpha}$ as response categories increased from two to six (F(2,243)=124.114, p<0.001) as seen in Figure 3. However, for $\hat{\delta}$, the mean RMSD decreased from 0.4300 to 0.3169 when the number of response categories increased from two to four, but then rose to 0.3680 when there were six response categories (F(2,243)=51.7280, p<0.001). For $\hat{\tau}$, the mean RMSD mimicked the pattern seen with $\hat{\delta}$. Specifically, it decreased from 0.3828 for two response categories to 0.3259 for four response categories, and then mean RMSD increased to 0.3912 when the number response categories reached six (F(2,243)=25.816, p<0.001). These results can be seen in Figures 4 and 5, respectively. Across these

parameters, there were statistically significant mean differences between all pairwise comparisons of the number of response categories, based on Tukey's HSD test ($p<0.0.125$), except for the pairwise comparison between two and six response categories with respect to $\hat{\tau}$. Therefore, the optimal accuracy was obtained with either 4 or 6 response categories depending on the type of item parameter in question. Two response categories consistently produced suboptimal accuracy across all item parameters.

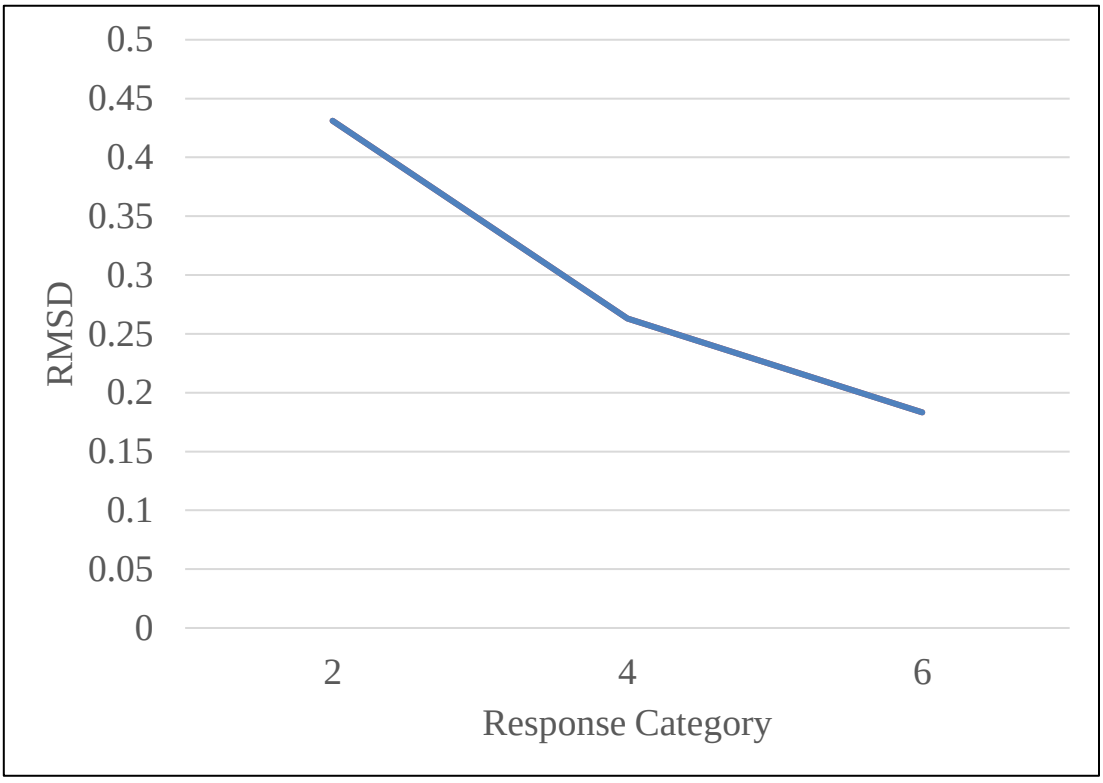

Figure 3. Mean RMSD of $\hat{\alpha}$ across response categories

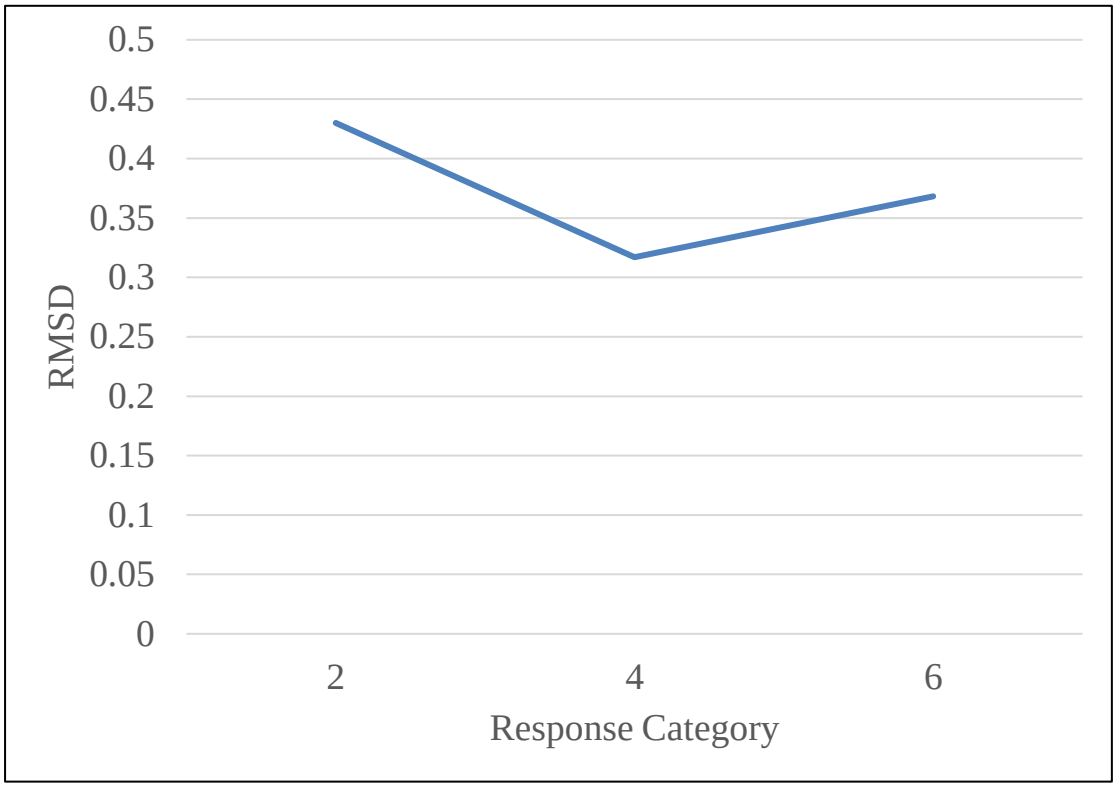

Figure 4. Mean RMSD of $\hat{\delta}$ across response categories

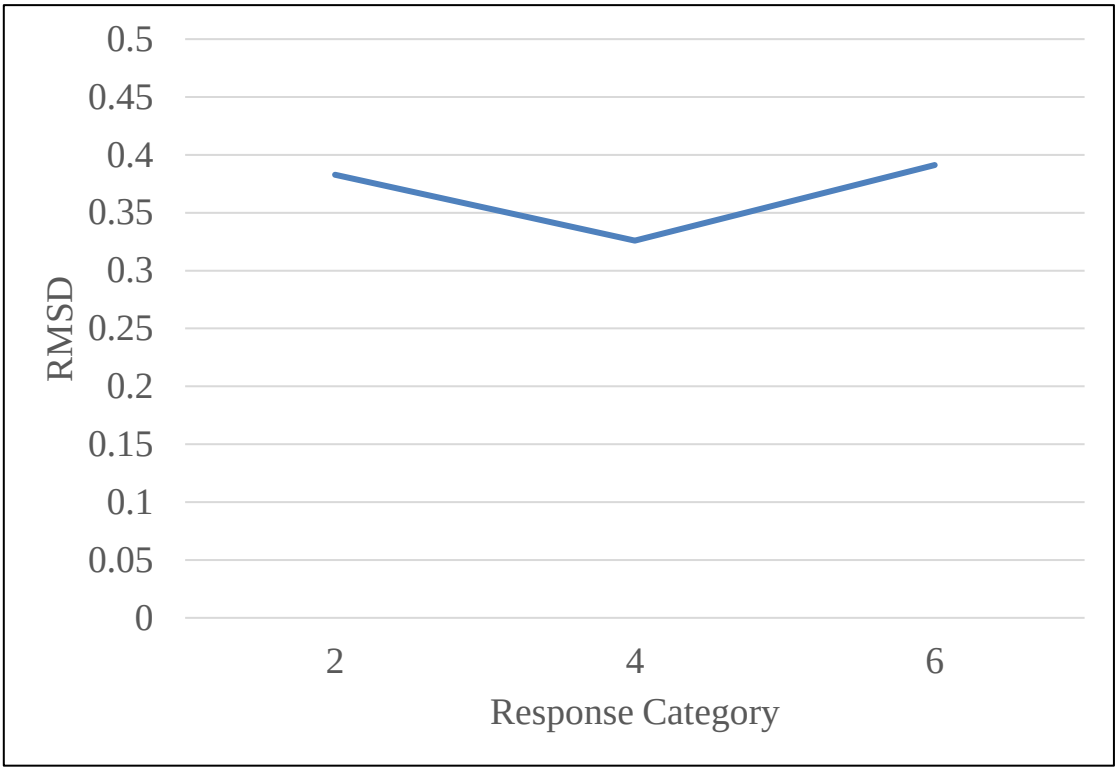

Figure 5. Mean RMSD of $\hat{\tau}$ across response categories

Beyond the effects of the number of response categories on each parameter type, the only other interpretable effect was the main effect of test length, but only when estimating $\hat{\alpha}_i$ ($F(2,243)=42.994$, $p<0.001$). As seen in Figure 6, mean RMSD decreased as the number of items increased, with the largest decrease occurring between 10 and 20 items and a very slight decrease thereafter. An increase in test length yields better approximations of $\bar{r}_{izf}$ counts in the expectation step of the algorithm, which can make item parameter estimation more precise. Moreover, because $\hat{\alpha}_i$ is a slope, and not a specific location, it benefits more from this increased precision more than the other item parameters; hence, the presence of an interpretable effect here. Based on the lack of interpretable effects beyond the ones mentioned, and based on percent of the total

variance in RMSD accounted for by the number of response categories effect, item parameter estimation accuracy benefits the most from an increase in the number of response categories. Additionally, when considering the item parameter estimation accuracy of $\hat{\alpha}_i$, the number of items should be greater than 10. It is interesting that, unlike other GGUM simulation study results, the GA-EM did not exhibit a benefit of increasing the sample size within these simulated conditions (Roberts & Thompson, 2011; Thompson, 2014). This is probably due to the fact that the minimum sample size in this study was 750, whereas the studies just mentioned also tested parameter recovery at a sample size of 500. The results found in Roberts et al. (2002) support this lack of sample size effect; that is, parameter recovery does not meaningfully diminish until sample size decreases below 750 examinees.

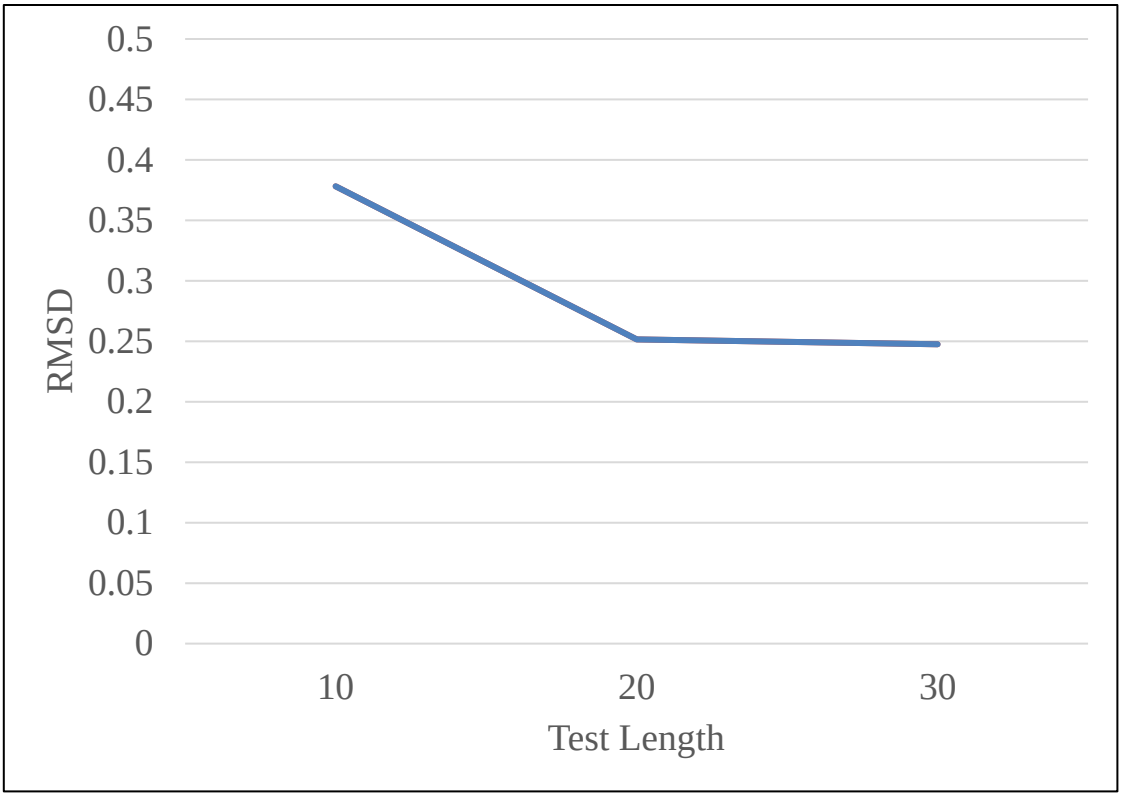

Figure 6. Mean RMSD of $\hat{\alpha}$ across test length

Person Parameters

As described in Chapter 2, person parameter estimates for each replication were obtained using five different methods: (1) an expected a posteriori (EAP) procedure whereby the conditional mean of the individual's posterior distribution of θ_j was found, (2) a maximum a posteriori (MAP) procedure in which the mode of the posterior likelihood of θ_j was found with GA, (3) a MAP procedure in which the mode of the posterior likelihood was found with a common Fisher scoring technique, (4) a maximum likelihood (MLE) procedure in which the maximum was found with the GA, and (5) an MLE procedure in which the mode was found with the Fisher scoring method. The mean

RMSD for each type of person parameter estimate is given in Table 4 for each factorial condition. The GA for θ_j estimation was implemented similarly as the GA for item parameters, albeit treating the estimated item parameters as known. Additionally, because each θ_j could be estimated separately (due to examinee independence), and the simulated data were unidimensional, the crossover operator was moot. The starting values for θ in each GA estimation procedure were set to the DCA starting values.

All between-subjects effects were examined for interpretability according to the same criteria as the ANOVAs for item parameters described previously. All within-subjects were evaluated for statistical significance using Huynh-Fedlt degree of freedom correction for any violation of the spherecity assumption. Based on these criteria, there were several statistically significant effects, and the main effect of number of response categories ($F(2, 243)=134.58$, $p<0.001$), and test length ($F(2, 243)=294.634$, $p<0.001$) accounted for the most variance in RMSD; by far more than the other effects in the model. The effect size calculations for each effect in the model can be found in Table 5. While the within subjects factor of estimation method was statistically significant ($F(4,240)=19.1203$, $p<0.001$), it did not account for as much variance in RMSD as the two aforementioned effects.The within-subjects main effect can be seen in Figure 7. Paired t-tests were conducted between each pairwise comparison of estimation method of θ, and were evaluated for statistical significance using a Bonferroni's corrected Type I error rate of 0.005=0.05/10. Based on this criteria, the only statistically significant comparisons were between MLE estimation method and all other methods. That is, all other methods reduced mean RMSD to a statistically significant degree.

With regard to the between-subjects main effects, mean RMSD decreased as response category increases from two to four ($p<0.0125$), and from two to six ($p<0.0125$). There was no statistically significant difference between mean RMSD of four and six response categories. Post-hoc analyses of the main effect of test length on mean RMSD shows that increasing test length will reduce the mean RMSD of person parameter estimation ($p<0.0125$ for each pairwise comparison). Figures 8 and 9 display these effects.

Table 4. Mean RMSD of person parameters estimates by condition

	Method of Estimation				
Factorial Condition	EAP	MAP	MLE	GA MAP	GA MLE
Sample Size					
750	0.33028	0.33215	0.37013	0.27864	0.28894
1000	0.27019	0.26964	0.28156	0.28523	0.29373
1250	0.27296	0.27166	0.26651	0.28901	0.28772
Test Length					
10	0.34792	0.35155	0.38483	0.3376	0.34849
20	0.27416	0.27353	0.28143	0.26968	0.27297
30	0.25113	0.24812	0.25161	0.24366	0.24700
Response Category					
2	0.33028	0.33215	0.37013	0.31447	0.32905
4	0.27019	0.26964	0.28156	0.26963	0.27946
6	0.27296	0.27166	0.26651	0.26712	0.26031

Table 5. Mean η_w^2 of person parameters estimates by condition

Effect	$\hat{\theta}_j$
Response Category	**1.03%**
Test Length	**2.23%**
Sample Size	0.00%
Test Length x Response Category	**0.13%**
Sample Size x Respsonse Category	0.03%
Sample Size x Test Length	0.01%
Sample Size x Test Length x Response Category	0.04%
Estimation Method	**0.06%**
Response Category x Estimation Method	**0.09%**
Test Length x Estimation Method	**0.05%**
Sample Size x Estimation Method	0.00%
Test Length x Response Category x Estimation Method	**0.05%**
Sample Size x Respsonse Category x Estimation Method	**0.04%**
Test Length x Response Category x Estimation Method	0.01%
Sample Size x Test Length x Response Category x Estimation Method	**0.06%**

Note: Bolded values were statistically significant ($p<0.0125$). Values bordered in red were interpretable based on relative η_w^2 values.

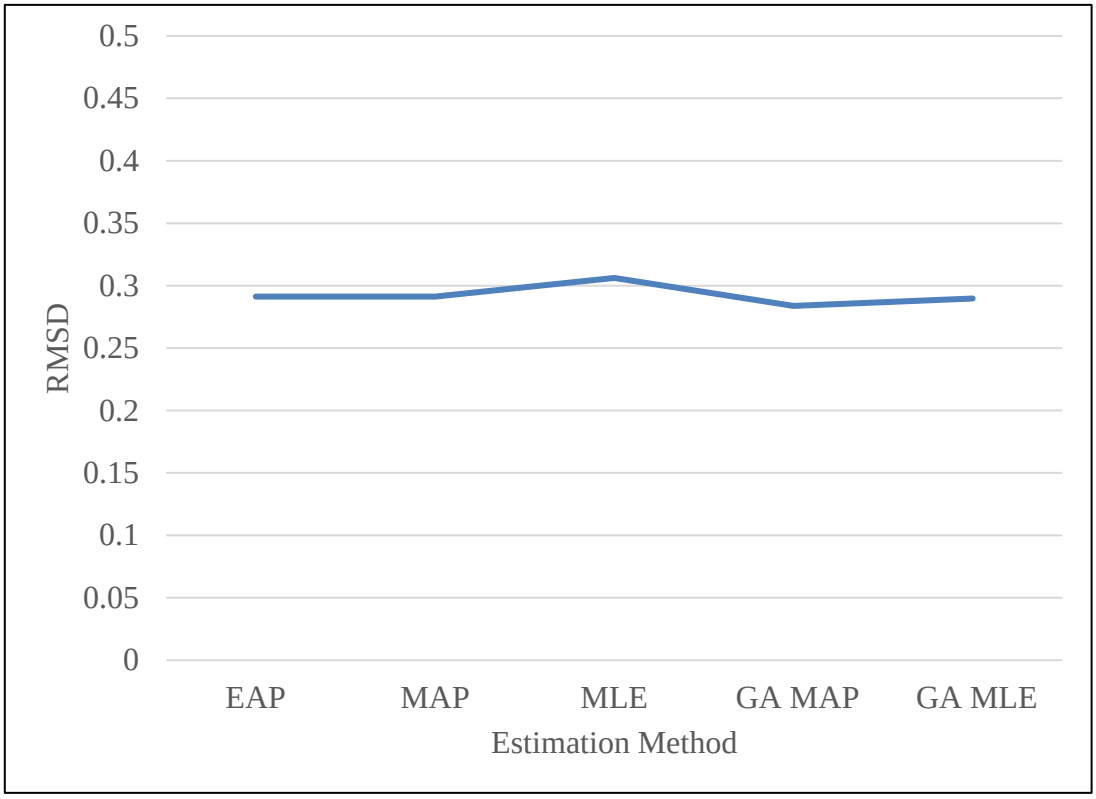

Figure 7. Mean RMSD of $\hat{\theta}$ across estimation method

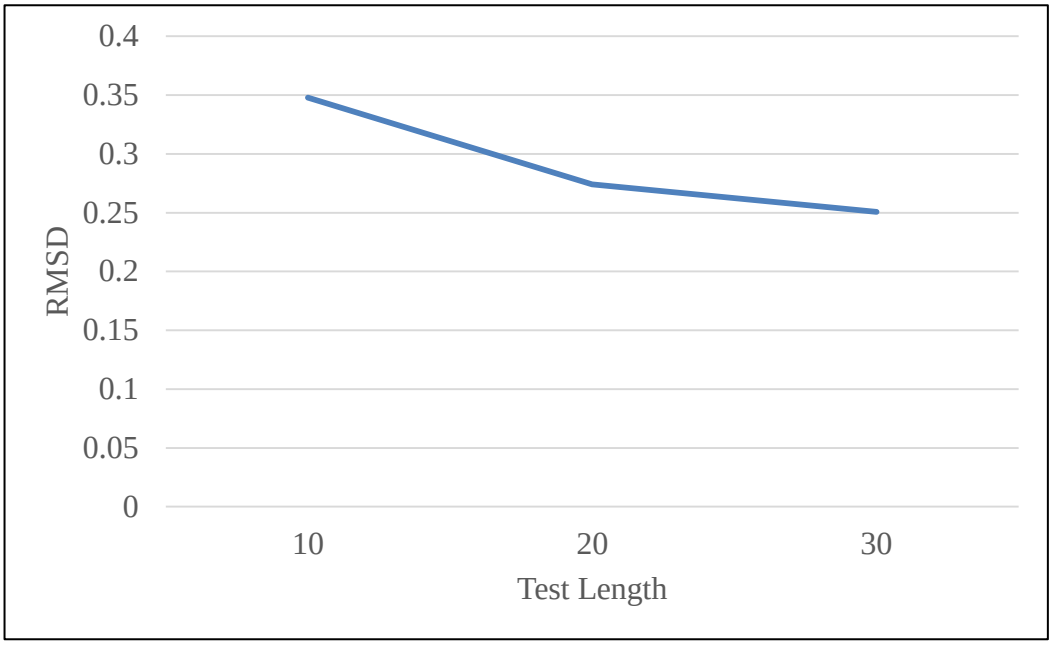

Figure 8. Mean RMSD of $\hat{\theta}$ across test length

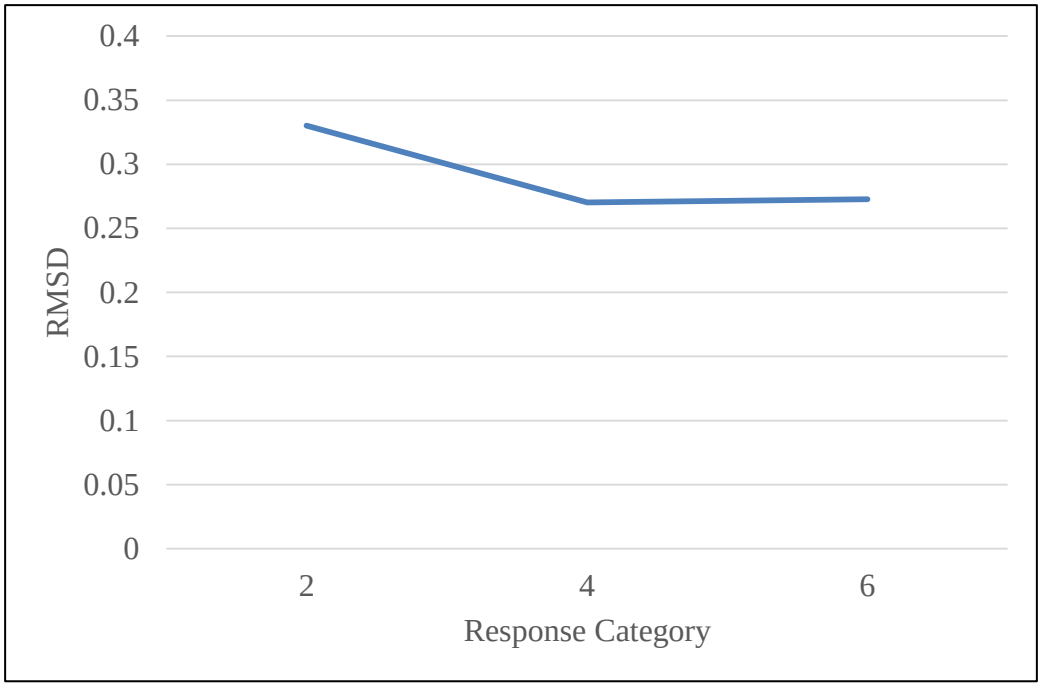

Figure 9. Mean RMSD of $\hat{\theta}$ across response categories

Grid Search

As described in Chapter 1, GA algorithms have been shown to avoid the pitfalls of local maxima in the likelihood function (or whatever the objective function might be). To investigate the existence, or lack thereof, of local maxima in the likelihood function and compare this to the simulation study results, a grid search was conducted. The grid search was conducted on the marginal likelihood for each item parameter type, and once on the likelihood with respect to $\hat{\theta}$. As mentioned previously, the count of maxima (minus 1) constitute the dependent measure for ANOVAs.

To illustrate the process of the grid search used here, consider conducting this search with respect to δ_i of a particular item within a given replication. To search the marginal likelihood function for maxima, the partial derivative of the marginal likelihood function, with respect to δ_i, is calculated for varying values of δ_i along the latent trait continuum of plausible values. Then, these values are investigated for any sign changes from positive to negative, which will constitute a count of 1. In the case of a maximum (local or global), the value of this partial derivative is equal to 0. Therefore, the value to the left and to the right of the maximum will have different signs. A count of these sign changes, from positive to negative is obtained across the successively larger values of δ_i for item *i*. Further, this count is averaged across items in a given replication. An analogous procedure was implemented for each parameter type in a given replication. All partial derivatives were calculated according the formulae given in Roberts et al. (2000).

All effects in the ANOVAs for maxima counts were evaluated for interpretability according to the effect size cutoff and a correct Type I error rate that have been discussed throughout this study. The mean count was highest for θ, and then τ, although none of the manipulated factors had an interpretable effect on the mean count for τ. The only interpretable effects found in any of the ANOVAs appeared within the analysis of the mean count of likelihood maxima with respect to θ. The main effect of number of response categories was statistically significant and met the effect size cutoff ($F(2,243)=19.798$, $p<0.001$). Additionally, the interaction between test length and sample size also had an interpretable effect on the mean count of maxima ($F(4,243)=15.4912$, $p<0.001$), which means that the effect of sample size on the mean count for person parameters depends upon the level of test length. However, these mean differences are too small to be meaningful to practitioners.

Table 6. Mean count of local maxima by simulation condition

	Parameter Type			
Factorial Condition	$\hat{\delta}_i$	$\hat{\alpha}_i$	$\hat{\tau}_i$	$\hat{\theta}_i$
Sample Size				
750	0.46646	0.00425	1.30874	1.516296
1000	0.46333	0.00423	1.34980	1.520111
1250	0.46858	0.00037	1.32038	1.518044
Test Length				
10	0.47611	0.00555	1.27563	1.522741
20	0.46419	0.0025	1.35625	1.515444
30	0.45777	0.001	1.35346	1.516367
Response Category				
2	0.44938	0.00888	1.28812	1.514844
4	0.47462	0	1.37108	1.527259
6	0.47259	0	1.31649	1.512348

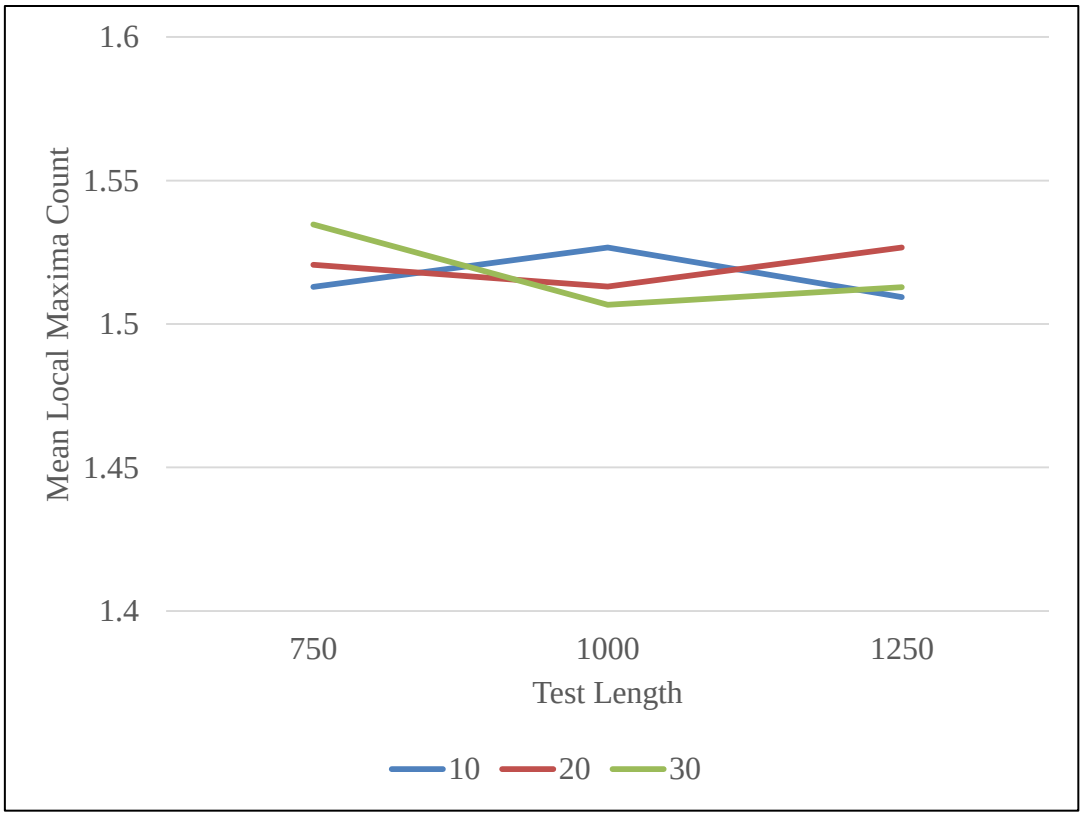

Figure 10. Mean Count for $\hat{\theta}$ at each level of response categories across test length

Empirical Data Application

The results of the simulation study have shown that the GA-EM is able to estimate GGUM parameters successfully in a variety of situations. To further investigate the applicability of the GA-EM, a set of real data from an abortion attitude questionnaire were analyzed. The questionnaire consisted of 40 statements ranging from pro-life to pro-

choice orientations. Each statement was associated with six response categories where 0=strongly disagree, 1=disagree, 2=slightly disagree, 3=slightly agree, 4=agree and 5=strongly agree. The data set contained responses from a random sample of 1,500 examinees, all of which were undergraduate college students from the Georgia Institute of Technology.

The estimation of GGUM parameters from this data set used the same procedure for start value calculation and implementation of the GA-EM as was described in Chapter 2. The item parameter estimates are shown in Table 7. The item parameters in Table 7 have been ordered by the value of $\hat{\delta}_i$ from smallest to largest. The contents of each statement can be found in Table A.1 in Appendix A. However, Table A.2 contains the same statements ordered according to the order in Table 7. When examining these statements and comparing them to the corresponding $\hat{\delta}_i$ values, it is clear that the statements range from pro-choice to pro-life in a logical fashion. Additionally, for all of the items in this dataset, the $\hat{\tau}_{ik}$ were ordinal. This is unlike previous applications of the GGUM to real data, where the $\hat{\tau}_{ik}$ were not generally ordered along the latent trait continuum (Roberts et al., 2000). The standard errors for item parameter estimates were calculated based on Roberts et al. (2000). Table B.1 displays the standard error estimates for each item and each parameter type. Additionally, a plot of the estimated standard errors against their corresponding $\hat{\delta}_i$ can be seen in Figure B.1. It is evident from this plot that the most extreme items (items 40 and 30) on either side of the latent trait continuum had the highest standard error estimates. This is due to the lack of examinees located around these items, thus making it harder to estimate these item locations. Figure B.2 displays a similar plot for $\hat{\theta}_i$. The bowl-shape nature of the plot is typical for a large

questionnaire with graded responses. Specifically, the standard error of an estimate generally increased as theta estimates become more distant from the mean. The magnitude of these changes is slight due to the large amount of information provided by the substantial number of polytomous questionnaire items.

Based on the estimation, Figure 12 displays the mean expected and observed responses as a function of (θ-δ). This plot illustrates the concordance between the average model predicted response and what was observed in the data. Therefore, when the estimated model parameters fit the data well, there will be a large overlap of the two lines. Additionally, the (largest) mode of both lines should be around (θ-δ)=0, because the closer an examinee is to an item, the higher mean expected and mean observed agreement should be. This figure shows that the estimated model fit the data relatively well based on these characteristics.

The person parameters for this dataset were estimated using the GA-MAP because this method led to the smallest observed mean RMSD across all five tested methods in the larger simulation study. A histogram of these estimates can be found in Figure 11. The distribution of $\hat{\theta}_j$ differed significantly from a normal distribution according to the Shapiro & Wilk criterion (W=0.96614, p<0.001). This distribution has a slight positive skew of 0.552, and a high degree of peakedness (kurtosis=2.692). The mean person parameter estimate was 0.006215, while the median was -0.150724. Based on this mean, the statement located around the average person in this sample was "7. My feelings about abortion are very mixed". Only 13 of the statements are located outside of the middle 50% of the person parameter estimates (25^{th} percentile of $\hat{\theta}_j$=-0.9899, 75^{th}

percentile of $\hat{\theta}_j$=0.912281). The statement located at the 25th percentile was statement "36. Although abortion on demand seems quite extreme, I generally favor a women's right to abortion", and the statement located at the 75th percentile was statement "33. Abortion should be illegal except in extreme cases involving incest or rape". Furthermore, the statement surrounding the most extreme $\hat{\theta}_j$ were "10. Society has no right to limit a woman's access to abortion" (at the 0.025 percentile of $\hat{\theta}_j$), and "6. Abortion is the destruction of one life for the convenience of another" (at the 0.975 percentile of $\hat{\theta}_j$).

Table 7. GGUM item parameter estimates of abortion attitude statements

Item	$\hat{\alpha}_i$	$\hat{\delta}_j$	$\hat{\tau}_{i1}$	$\hat{\tau}_{i2}$	$\hat{\tau}_{i3}$	$\hat{\tau}_{i4}$	$\hat{\tau}_{i5}$
40	0.8508	-2.875	-2.688	-2.667	-2.342	-2.3	-1.1016
10	1.1398	-1.563	-2.814	-2.308	-1.785	-1.488	-0.7407
9	0.9451	-1.499	-2.627	-2.324	-2.179	-1.231	-0.9742
39	1.371	-1.495	-2.852	-2.446	-2.44	-1.309	-0.9684
35	0.7858	-1.437	-2.742	-2.709	-2.686	-1.909	-1.5815
36	1.187	-1.173	-2.536	-2.097	-2.054	-1.268	-0.4318
38	1.1076	-0.949	-2.22	-1.765	-1.743	-1.23	-0.941
27	0.866	-0.94	-1.039	-1.037	-1.029	-0.891	-0.0784
19	1.1523	-0.838	-2.147	-1.277	-1.223	-0.013	-0.0037
20	0.7896	-0.79	-1.534	-1.303	-1.277	-0.166	-0.0046
4	1.5247	-0.762	-1.998	-1.388	-1.236	-0.407	-0.0003
37	0.8865	-0.76	-2.148	-1.997	-1.982	-1.318	-0.7675
22	1.4558	-0.714	-2.118	-1.543	-1.523	-0.784	-0.007
26	0.9475	-0.612	-1.109	-1.079	-0.768	-0.038	-0.0011
23	1.009	-0.542	-1.117	-0.653	-0.013	-0.005	-0.0041
25	1.1304	-0.531	-1.612	-0.718	-0.608	-0.031	-0.0018
3	0.9739	-0.512	-1.114	-0.846	-0.122	-0.008	-0.0012
24	0.8498	-0.503	-1.202	-1.054	-1.051	-0.027	-0.0089
21	0.9722	-0.363	-1.525	-1.384	-0.524	-0.012	-0.0022

Table 7. Continued

Item	$\hat{\alpha}_i$	$\hat{\delta}_i$	$\hat{\tau}_{i1}$	$\hat{\tau}_{i2}$	$\hat{\tau}_{i3}$	$\hat{\tau}_{i4}$	$\hat{\tau}_{i5}$
13	0.8742	-0.306	-1.551	-1.186	-0.232	-0.006	-0.0049
7	0.9386	0.0249	-1.188	-1.141	-1.13	-0.328	-0.3188
8	0.9509	0.0356	-1.315	-0.933	-0.931	-0.292	-0.2567
12	1.0285	0.3026	-1.609	-0.925	-0.146	-0.015	-0.0015
15	0.8944	0.4038	-2.14	-1.289	-1.217	-0.018	-0.004
18	1.3365	0.5531	-1.854	-1.083	-0.718	-0.054	-0.0172
17	1.2183	0.6512	-1.711	-1.202	-0.656	-0.031	-0.0095
1	1.2363	0.6696	-1.783	-1.047	-0.771	-0.034	-0.0185
2	1.1985	0.6756	-1.825	-1.322	-0.879	-0.202	-0.0149
14	1.3059	0.6809	-1.919	-1.129	-0.823	-0.032	-0.0003
16	1.4834	0.7269	-1.934	-1.449	-0.796	-0.268	-0.0004
11	1.0637	0.8011	-1.423	-1.212	-0.142	-0.009	-0.0042
33	1.3432	0.9593	-2.151	-1.541	-1.18	-0.876	-0.1844
32	1.1414	0.9988	-2.302	-1.492	-1.451	-0.894	-0.0027
31	0.9631	1.4367	-2.725	-1.96	-1.917	-1.278	-0.9912
34	0.4961	1.5362	-2.32	-2.3	-2.3	-1.691	-1.3189
5	0.7024	1.6222	-2.061	-2.058	-2.049	-1.089	-0.5254
28	0.9761	2.0083	-2.931	-2.605	-2.598	-1.633	-1.2358
29	1.1297	2.1892	-2.985	-2.764	-2.732	-1.785	-1.3306
6	0.7918	2.2278	-2.913	-2.882	-2.863	-2.079	-1.3049
30	0.8662	3.6955	-3.434	-3.427	-3.066	-2.958	-2.024

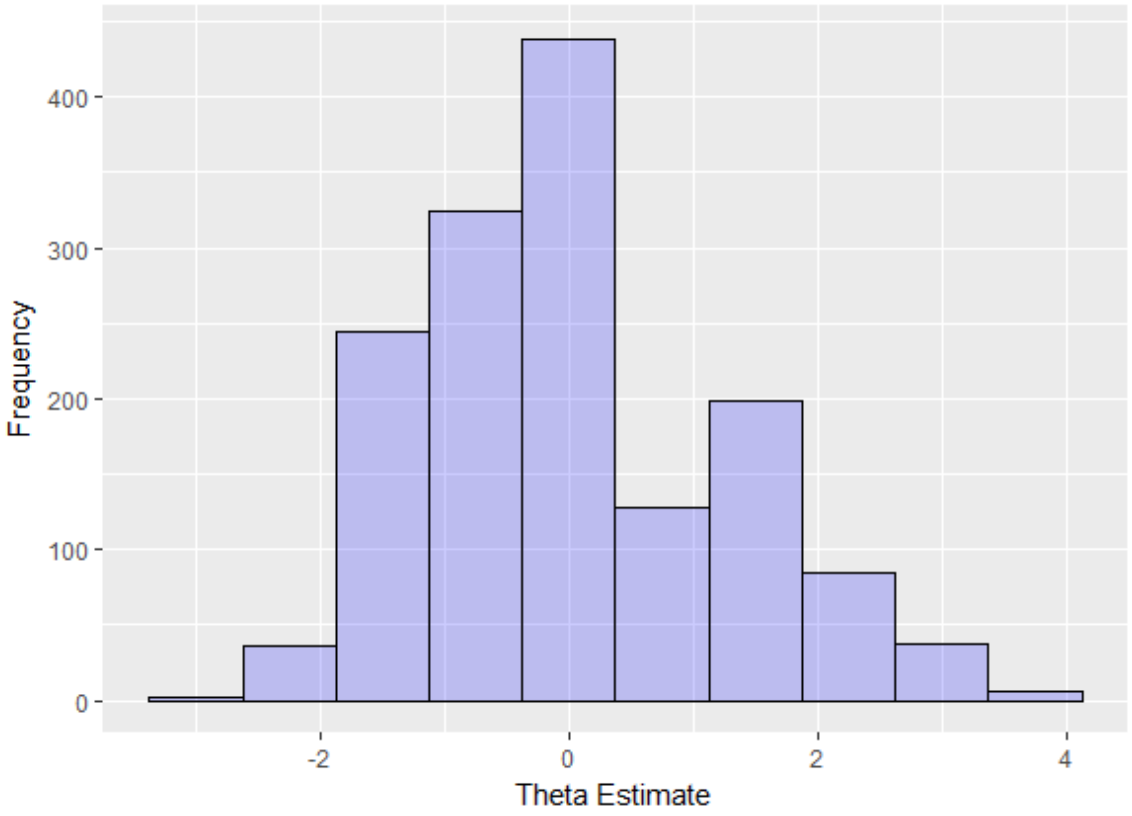

Figure 11. Histogram of $\hat{\theta}$ for the abortion attitude data

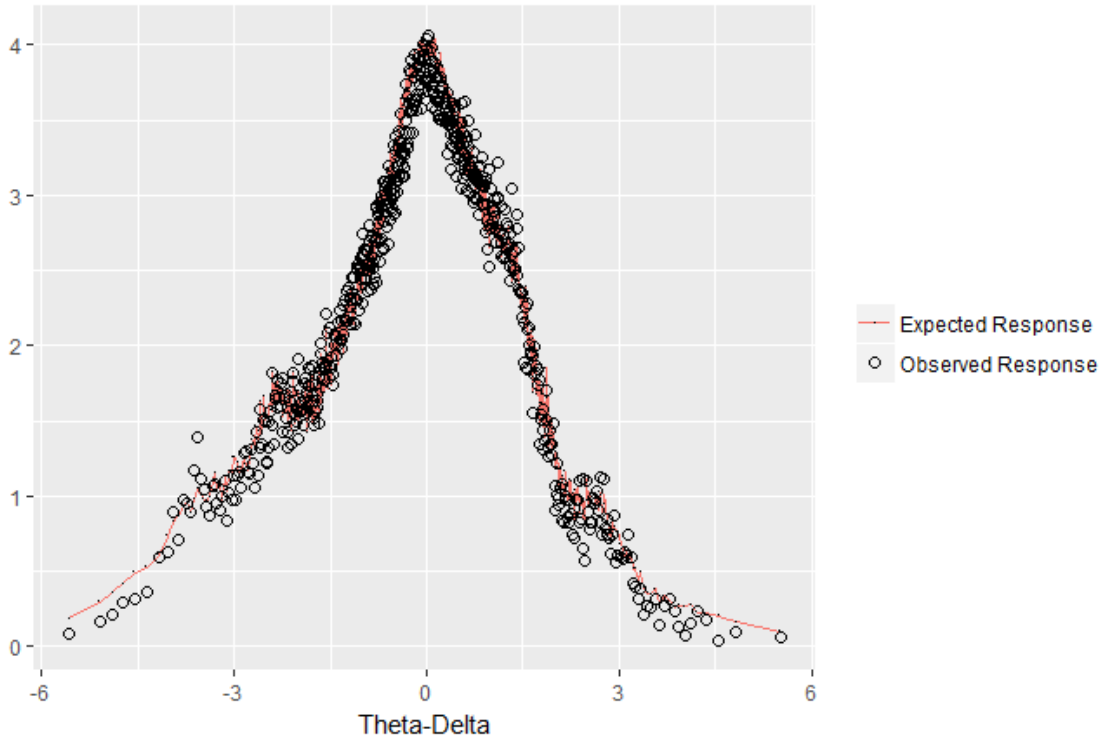

Figure 12. Average expected response and average observed response across aggregated values of (θ-δ).

CHAPTER 5. DISCUSSION

The primary goal of this study was to lay the groundwork for the GA application within a more complex IRT model generally, as well as specifically build a foundation for successful GA application within the GGUM. To achieve this goal, the current study developed a GA-EM for GGUM parameter estimation, and examined its performance using a simulation study and an empirical data application.

Based on the preliminary tests and the associated item parameter recovery results, the current GA-EM for GGUM parameter estimation seems relatively robust to manipulations of data and GA specific characteristics. Specifically, four different parameters of the GA-EM were varied independently, and did not result statistically different item parameter recovery. One contributing factor to the robustness of the GA-EM to changes in the algorithm could be the informed starting values used (i.e., for δ_i and τ_{ik}). Because these starting values place the algorithm close to the global solution, it is not necessary for the GA-EM to have a larger population size for example. Additionally, the GA-EM is robust to changes in mutation probability above a minimum value (i.e., 0.1 tested here), which shows that the GA-EM is more sensitive to not enough variance in the population than too much variance among the candidate solutions. With respect to improvements to the GA-EM, computational speed is an area where there could be significant strides made. The GA-EM was implemented in this study without a maximum number of iterations, only a cutoff value based on estimated parameter differences. By capping the number of iterations for the innermost cycle of the algorithm, the GA-EM

could be executed much quicker by eliminating iterations of the algorithm that occur after the estimations are relatively stable.

Roberts and Laughlin (1996) briefly addressed the fact that the likelihood function for the GGUM generally has local maxima, especially with respect to person parameters. However, this aspect of the GGUM has never been systematically studied until now. These results suggest that the GGUM marginal likelihood function may have local maxima with respect to all item parameters, but it is most problematic for τ_{ik} parameters and least noticeable for α parameters. On average there were 1.33 local maxima in the marginal likelihood function for a given τ_{ik} parameter. It is possible that the GA-EM has helped find the absolute maximum in these cases, but this is only conjecture because the standard EM algorithm utilizing Fisher scoring was not included as a comparison condition.

The mean number of local maxima in the likelihood function with respect to person parameters was equal to 1.52, and was the highest seen for any GGUM parameter. Unlike the count of local maxima for item parameters, the number of local maxima for θ parameters was related to the design factors implemented in the simulation study. However, the magnitude of mean differences lacked practical significance.

With respect to GGUM item parameter estimation in the main simulation, the current version of the GA-EM was found to efficiently recover reasonably accurate item parameter estimates from simulated GGUM response data, without the use of item prior distributions or complex start values derivations beyond DCA calculations. Moreover, the only experimental factor that resulted in interpretable parameter recovery differences

for all item parameter types was the number of response categories. These results suggest that the GA-EM can recover item parameters relatively well when there are between 4 to 6 response categories. This finding is consistent with those of Roberts and Thompson (2011) who suggested that little gain in precision can be obtained once the number of response categories exceeds 5. The current simulation also showed that the precision of item parameter estimates obtained with the GA-EM does not substantially improve with sample sizes of more than 750 simulees. This is also consistent with traditional EM results reported by Roberts et al. (2002) who showed that sample sizes greater than N=750 led to noticeably smaller corresponding increases in item parameter estimation accuracy. Obviously, if substantially smaller samples had been implemented in the current simulation, then one would expect degraded parameter recovery with decreasing sample size. Finally, the current results indicated that once the number of items reached 20, there was little improvement in accuracy for any item parameters. This is consistent with past studies of EM parameter estimation using Fisher scoring where 15 to 20 items were recommended (Roberts et al., 2002). Presumably, the test length effect emerges in the discrimination parameter estimates due to increased accuracy of the estimated theta distribution in the EM algorithm with larger test lengths. The marginal likelihood with respect to all item parameters is developed using this estimated distribution, and thus, it too is better estimated with longer tests.

With respect to person parameter estimation, the GA-MAP and GA-MLE produced a smaller observed mean RMSD than their non-GA counterparts (MAP and MLE, respectively). Indeed, the GA-MAP was the most precise theta estimate of those investigated here. As mentioned above, this research showed that the person parameter

likelihoods contained the highest mean count of local maxima relative to other GGUM parameters. So, although the mean RMSD differences between GA-MAP and GA-MLE and their non-GA counterparts did not lead to an interpretable effect, it does make sense that GA estimation of theta parameters would have better observed parameter recovery given the relatively higher incidence of local maxima with these parameters. In short, the differences observed in mean RMSD across estimation methods, while not deemed interpretable, made logical sense and were in the expected direction. Furthermore, it could be the case that in measurement situations where the frequency of local maxima is potentially higher (e.g., multidimensional GGUM applications), this improvement in parameter recovery could be magnified. The same would be true if random starting values for GGUM parameters were used in the algorithm rather than the informed starting values that are traditionally used.

Another feature of the GA that is pertinent to IRT models in general and the GGUM in particular concerns it mathematical simplicity. As new models are developed at a higher rate, a means to quickly test the performance of these models becomes more valuable. Approaches like MCMC and GA are particularly attractive in that they avoid the need to calculate and subsequently program partial derivatives for model parameters, and therefore, the speed with which new models can be investigated is potentially enhanced.

The GA procedure implemented in this research was also used to estimate GGUM parameters using real responses to an abortion attitude questionnaire. The resulting item locations were strongly related to the pro-choice or pro-life orientation of item content. Moreover, the estimates of subjective response category thresholds were generally

ordered in a logical fashion. This appealing result is rarely obtained when six response categories are used in the traditional EM estimation procedure (J.S. Roberts, personal communication, Summer, 2017). Indeed, when a standard EM procedure was used to obtain item parameters with the same data (not reported), the thresholds were noticeable disordinal for many of the items.

A note about the computational speed of this GA-EM application to item parameter estimation in the GGUM is in order. The average time for item parameter estimation across simulation replications using the GA-EM was approximately 40 minutes, and thus, the algorithm is not as fast as traditional EM estimation in the GGUM which generally requires only a few minutes if that long. It does, however, improve upon the computational speed of MCMC item parameter estimates, with the benefit of a much simpler mathematical form (Roberts et al., 2002). Moreover, with faster compiled languages like FORTRAN or C++ along with more streamlined code, the computational efficiency of the GA-EM algorithm implemented in this study could certainly be improved.

Obviously, the GA-EM is no panacea. In addition to its relative computational inefficiency compared to traditional EM, it is generally not portable across problems. It typically requires some tweaking depending on the optimization problem under study. When the data are believed to follow a GGUM model, then this GA approach can be used to estimate accurate item and person parameters. However, there are many other psychometric measurement models for which a GA-EM procedure may be effective in theory, but cannot be applied without first tailoring the algorithm to the model and the typical data that are encountered. This makes the application of the GA technique more

problematic for mathematical programmers and measurement practitioners than other estimation techniques that can easily be applied to a variety of measurement models in a more or less "canned" fashion.

The primary goal of this work was to develop and implement a GA approach to estimate parameters in the GGUM. Another relevant goal was to broaden the GA literature, and in doing so, merge it with the field of IRT parameter estimation. Based on a literature search, the current study is the first successful application of a GA-EM to polytomous IRT data, with respect to item parameter estimation. Furthermore, it is the first instance of a GA-EM being successfully applied to IRT data that do not follow a cumulative model. Therefore, the benefits of the GA can be realized beyond the binary 3-PLM case that has formed the basis of all published GA applications to date. Furthermore, by taking the first step of extending the application of GA across more varied IRT model estimation problems, psychometricians can then investigate and take advantage of other evolutionary processing optimization advancements beyond the traditional GA. For example, the ability of the GA to hybridize with other optimization methods (Gehlhaar & L.J., 1995; Wieland, 1990).

Overall, the current study has helped further the development of the GGUM and increase the practicality of the model for use in common psychological research. By increasing the accuracy of estimation and ease of implementation of the GGUM, practitioners and researchers beyond psychometrics will more readily apply the model in research situations where the measurement of a proximity-based response process is necessary for correct interpretation of the psychological constructs being investigated. It is only when psychological constructs are validly and reliably measured that they are

truly useful for investigating important research questions involving individual differences.

The current study laid the groundwork for eventual GA application to estimate the more complicated versions of GGUM. These include GGUMs for multidimensional latent constructs or those in which responses are from a mixture of several latent populations or both. In these situations, the number of local maxima associated with model parameters is expected to increase and sequentially oriented search algorithms will become less efficient and perhaps less accurate. This research has shed light on a potential solution to avoid local maxima in GGUM parameter estimation by capitalizing on a parallel search and randomization strategy. It is hoped that future research with more elaborate GGUMs will use this work as a roadmap for successful parameter estimation.

APPENDIX A. ABORTION ATTITUDE STATEMENTS

Table A.1. Abortion attitude statements

01. Abortion should not usually be allowed unless the child will be extremely mentally retarded.
02. Abortions should not normally be performed unless there is medical evidence that the baby will die before one year of age.
03. Abortions should typically be allowed, but only when both biological parents are legal adults.
04. Abortions should generally be allowed, but only when the woman obtains counseling beforehand.
05. Abortion could destroy the sanctity of motherhood.
06. Abortion is the destruction of one life for the convenience of another.
07. My feelings about abortion are very mixed.
08. I cannot whole-heartedly support either side of the abortion debate.
09. Outlawing abortion violates a woman's civil rights.
10. Society has no right to limit a woman's access to abortion.
11. Abortions should generally be illegal except in cases involving women in prison.
12. Abortion should not usually be allowed except when the woman is financially unable to support the child.
13. Abortion should not typically be illegal except in cases where the woman is not emotionally capable of rearing the child.
14. As a general rule abortion should be illegal unless the woman is mentally incapable of caring for a child.
15. Abortion should be avoided unless the woman is not physically able to raise the child.
16. Abortion is generally unacceptable except when the child will never be able to live outside a medical institution.

Table A.1. Abortion attitude statements continued

17. Abortions should generally be prohibited except when there is medical evidence that the child will be unable to hear, speak, and see.
18. Abortions should not usually be permitted unless the child will never be able to care for itself.
19. Abortions should generally be legal unless the woman is mentally incapable of making a decision to undergo the procedure.
20. Abortion is acceptable in most cases, but it should not be supported with tax dollars.
21. Abortions should generally be legal unless a sonogram has detected a heartbeat.
22. Abortions should usually be permissible, but other alternatives must be explored first.
23. Abortions should be acceptable most of the time unless the pregnant person is a minor.
24. Abortions should be allowed only if both biological parents agree to it.
25. Abortions should typically be permitted unless the parents are fully capable of providing a good home life for the child.
26. Abortions should be permitted unless the woman has had multiple abortions in the past.
27. Abortion should usually be legal except when it is performed simply to control the gender balance in a family.
28. Abortion can be described as taking a life unjustly.
29. Abortion is inhumane.
30. Abortion is unacceptable under any circumstances.
31. Even if one believes that there may be some exceptions, abortion is still generally wrong.
32. Abortion is basically immoral except when the woman's physical health is in danger.
33. Abortion should be illegal except in extreme cases involving incest or rape.
34. Abortion should not be made readily available to everyone.

Table A.1. Abortion attitude statements continued.

35. Regardless of my personal views about abortion, I do believe others should have the legal right to choose for themselves.
36. Although abortion on demand seems quite extreme, I generally favor a woman's right to choose.
37. Abortion should be a woman's choice, but should never be used simply due to its convenience.
38. Abortion should generally be legal, but should never be used as a conventional method of birth control.
39. A woman should retain the right to choose an abortion based on her own life circumstances.
40. Abortion should be legal under any circumstances.

Table A.2. Ordered abortion attitude statements

40.	Abortion should be legal under any circumstances.
10.	Society has no right to limit a woman's access to abortion.
9.	Outlawing abortion violates a woman's civil rights.
39.	A woman should retain the right to choose an abortion based on her own life circumstances.
35.	Regardless of my personal views about abortion, I do believe others should have the legal right to choose for themselves.
36.	Although abortion on demand seems quite extreme, I generally favor a woman's right to choose.
38.	Abortion should generally be legal, but should never be used as a conventional method of birth control.
27.	Abortion should usually be legal except when it is performed simply to control the gender balance in a family.
19.	Abortions should generally be legal unless the woman is mentally incapable of making a decision to undergo the procedure.
20.	Abortion is acceptable in most cases, but it should not be supported with tax dollars.
4.	Abortions should generally be allowed, but only when the woman obtains counseling beforehand.
37.	Abortion should be a woman's choice, but should never be used simply due to its convenience.
22.	Abortions should usually be permissible, but other alternatives must be explored first.
26.	Abortions should be permitted unless the woman has had multiple abortions in the past.
23.	Abortions should be acceptable most of the time unless the pregnant person is a minor.
25.	Abortions should typically be permitted unless the parents are fully capable of providing a good home life for the child.
3.	Abortions should typically be allowed, but only when both biological parents are legal adults.
24.	Abortions should be allowed only if both biological parents agree to it.
21.	Abortions should generally be legal unless a sonogram has detected a heartbeat.
13.	Abortion should not typically be illegal except in cases where the woman is not emotionally capable of rearing the child.
7.	My feelings about abortion are very mixed.

8.	I cannot whole-heartedly support either side of the abortion debate.
12.	Abortion should not usually be allowed except when the woman is financially unable to support the child.
15.	Abortion should be avoided unless the woman is not physically able to raise the child.
18.	Abortions should not usually be permitted unless the child will never be able to care for itself.
17.	Abortions should generally be prohibited except when there is medical evidence that the child will be unable to hear, speak, and see.
1.	Abortion should not usually be allowed unless the child will be extremely mentally retarded.
2.	Abortions should not normally be performed unless there is medical evidence that the baby will die before one year of age.
14.	As a general rule abortion should be illegal unless the woman is mentally incapable of caring for a child.
16.	Abortion is generally unacceptable except when the child will never be able to live outside a medical institution.
11.	Abortions should generally be illegal except in cases involving women in prison.
33.	Abortion should be illegal except in extreme cases involving incest or rape.
32.	Abortion is basically immoral except when the woman's physical health is in danger.
31.	Even if one believes that there may be some exceptions, abortion is still generally wrong.
34.	Abortion should not be made readily available to everyone.
5.	Abortion could destroy the sanctity of motherhood.
28.	Abortion can be described as taking a life unjustly.
29.	Abortion is inhumane.
6.	Abortion is the destruction of one life for the convenience of another.
30.	Abortion is unacceptable under any circumstances.

APPENDIX B. REAL DATA ESTIMATION STANDARD ERRORS

Table B.1. GGUM item parameter standard errors from abortion attitude data

Item	$\hat{\alpha}_i$	$\hat{\delta}_i$	$\hat{\tau}_{i1}$	$\hat{\tau}_{i2}$	$\hat{\tau}_{i3}$	$\hat{\tau}_{i4}$	$\hat{\tau}_{i5}$
40	0.0486	1.0833	1.1124	1.1615	1.4641	3.8238	2.6639
10	0.0100	0.0015	0.0076	0.0865	0.3853	0.8786	0.7313
9	0.0066	0.0003	0.0117	0.0889	0.3632	0.6698	0.5180
39	0.0421	0.0007	0.0552	0.0353	0.4286	1.1289	0.9680
35	0.0031	0.0003	0.0068	0.0748	0.2560	0.5839	0.5342
36	0.0165	0.0041	0.0219	0.1034	0.3204	0.6697	0.5392
38	0.0115	0.0019	0.0158	0.0806	0.1854	0.3902	0.3577
27	0.0372	0.0118	0.0647	0.1209	0.1943	0.3004	0.2748
19	0.0349	0.0033	0.0484	0.1453	0.3231	0.4386	0.3138
20	0.0506	0.0168	0.0913	0.1270	0.2379	0.3330	0.2703
4	0.0320	0.0041	0.0240	0.0974	0.3074	0.5874	0.4324
37	0.0325	0.0077	0.0528	0.0653	0.1898	0.3798	0.3475
22	0.0309	0.0009	0.0272	0.0875	0.2730	0.6007	0.4720
26	0.0566	0.0144	0.0706	0.1166	0.2019	0.2665	0.2381
23	0.0512	0.0171	0.0772	0.1288	0.1783	0.2441	0.2553
25	0.0511	0.0097	0.0715	0.1252	0.2173	0.3055	0.2455
3	0.0550	0.0177	0.0780	0.1348	0.1869	0.2313	0.2199
24	0.0477	0.0144	0.0770	0.1267	0.2161	0.2820	0.2325
21	0.0488	0.0138	0.0659	0.1262	0.2102	0.3051	0.2826

Table B.1. Continued

Item	$\hat{\alpha}_i$	$\hat{\delta}_i$	$\hat{\tau}_{i1}$	$\hat{\tau}_{i2}$	$\hat{\tau}_{i3}$	$\hat{\tau}_{i4}$	$\hat{\tau}_{i5}$
13	0.0470	0.0193	0.0822	0.1500	0.2109	0.2822	0.2914
7	0.0425	0.0142	0.0518	0.0992	0.1898	0.2978	0.2564
8	0.0469	0.0182	0.0650	0.1063	0.1942	0.3135	0.2621
12	0.0466	0.0151	0.0794	0.1580	0.2295	0.3259	0.3087
15	0.0294	0.0102	0.0661	0.1374	0.2964	0.4495	0.3612
18	0.0582	0.0061	0.0603	0.1490	0.3230	0.5295	0.4469
17	0.0505	0.0063	0.0632	0.1572	0.3214	0.4578	0.3621
1	0.0487	0.0076	0.0663	0.1557	0.3258	0.4815	0.3888
2	0.0380	0.0041	0.0564	0.1556	0.3322	0.4904	0.3883
14	0.0478	0.0047	0.0549	0.1592	0.3687	0.5562	0.4357
16	0.0623	0.0021	0.0505	0.1623	0.3934	0.6605	0.5417
11	0.0560	0.0127	0.0809	0.1765	0.2808	0.3528	0.3171
33	0.0282	0.0022	0.0305	0.1423	0.3196	0.7055	0.6204
32	0.0264	0.0059	0.0428	0.1324	0.3476	0.7002	0.5733
31	0.0070	0.0004	0.0209	0.1731	0.4622	0.7265	0.6046
34	0.0119	0.0072	0.0680	0.2383	0.4457	0.6397	0.5386
5	0.0158	0.0044	0.0448	0.1717	0.3325	0.5640	0.5027
28	0.0045	0.0002	0.0110	0.1094	0.4444	0.9897	0.8411
29	0.0118	0.0002	0.0074	0.0576	0.4692	1.2538	1.0514
6	0.0091	0.0003	0.0107	0.1653	0.6023	1.0793	0.8606
30	0.0137	1.1549	1.1519	1.1913	1.0306	5.4300	3.4651

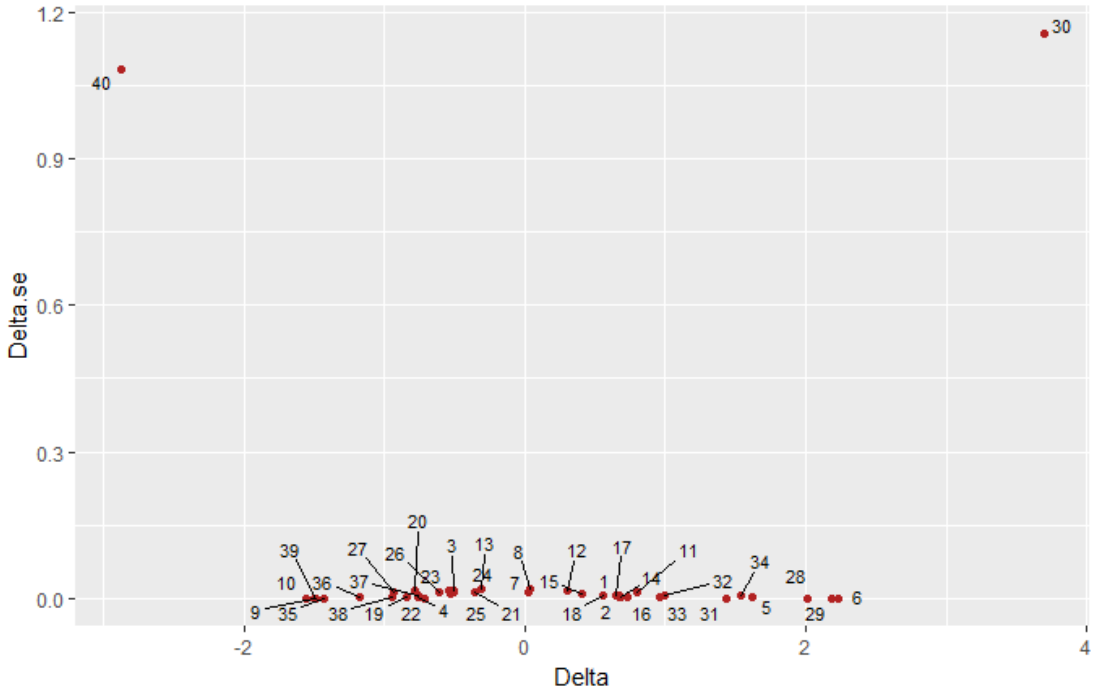

Figure B.1. $\widehat{\delta_\iota}$ and SEs from abortion attitude data

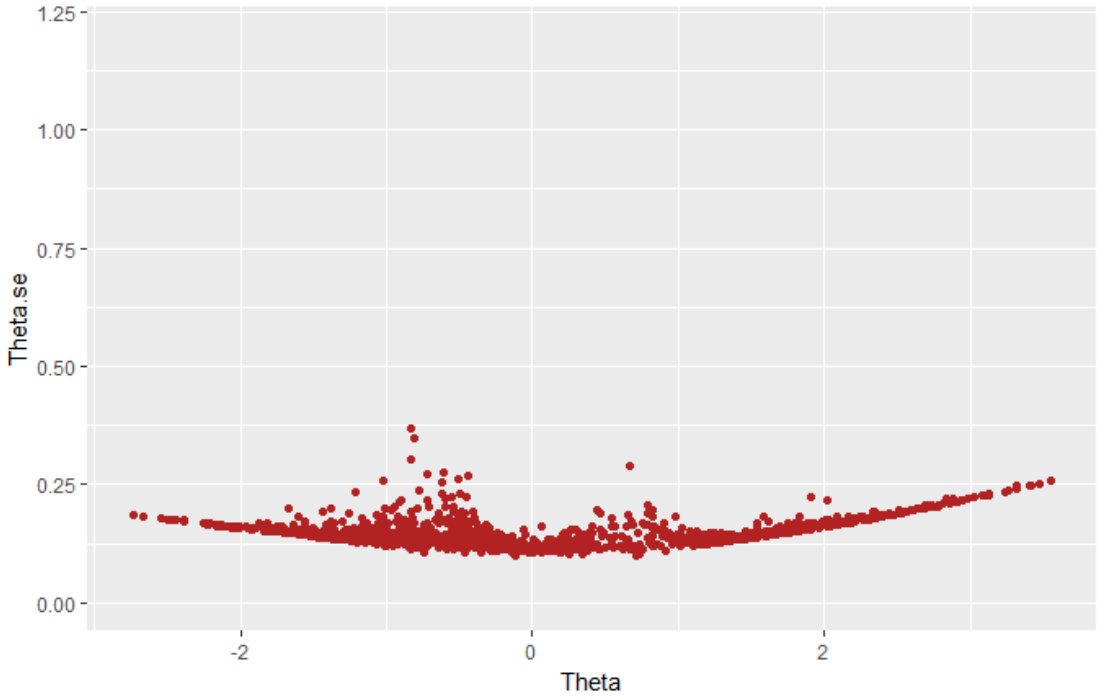

Figure B.2. $\widehat{\theta_j}$ and SEs from abortion attitude data

REFERENCES

Andrich, D. (1988). The application of an unfolding model of the PIRT type for the measurement of attititude. *Applied Psychological Measurement, 12*, 33-51.

Andrich, D. (1996). A hyperbolic 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*(3), 253-276.

Baker, F. B. (1987). Methodology Review: Item Parameter Estimation Under the One-, Two-, and Three-Parameter Logistic Models. *Applied Psychological Measurement, 11*(2), 111-141.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-472). Reading, MA: Addison Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application of an Em Algorithm. *Psychometrika, 46*(4), 443-459.

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.

Castellani, F., & Franceschini, G. (2003). *Use of Genetic Algorithms as an Innovative Tool for Race Car Design*. http://dx.doi.org/10.4271/2003-01-1327

Ceylan, H., & Bell, M. G. (2004). Traffic signal timing optimisation based on genetic algorithm approach, including drivers' routing. *Transportation Research Part B: Methodological, 38*(4), 329-342.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

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.

Dengiz, B., Altiparmak, F., & Smith, A. E. (1997). Local search genetic algorithm for optimal design of reliable networks. *IEEE Transactions on Evolutionary Computation, 1*(3), 179-188.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology, 3*(4), 465-476.

Du, P., & Chu, Y. (2013). The improved genetic algorithm applied on paramter estimation of two parameter logistic model on item repsonse theory. *Advanced Materials Research*, 756-759.

Gehlhaar, D. K., Verkhivker, G.M., Rejto, P.A., Sherman, C.J., Fogel, D.B., Fogel,, & L.J., F., S.T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease:

conformationally flexible docking by evolutionary programming. *Chem. & Biol., 2*, 317-324.

Goldberg, D. E. (1989). *Genetics algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley Pub. Co.

Hill, M. O., & H. G. Gauch, J. (1980). Detrended Correspondence Analysis: An Improved Ordination Technique. *Vegatio, 42*, 47-58.

Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trials. *SLAM Journal on Computing, 2*(2), 88-105.

Jiang, H., & Tang, L. (1998). *New method of calibrating IRT models*. Paper presented at the National Conference on Measurement in Education, San Diego, CA.

King, D. R. (2017). *Stocahstic approximation of the multidimensional generalized graded unfolding model with the Metropolis-Hastings Robbins-Monro algorithm.* (Doctoral Dissertation), Georgia Institute of Technology, Atlanta, GA.

Li, H. H. (1997). *Using genetic algorithms to estimate IRT item parameters*. Paper presented at the Annual Meeting of Psychometric Society, Gatlinburg, TN.

Noel, Y. (1999). Recovering unimodal latent patterns of change by unfolding analysis: Application to smoking cessation. *Psychological Methods, 4*(2), 173.

Reeves, C. R., & Rowe, J. E. (2003). *Genetic algorithms: Principles and perspectives*. Boston, MA: Kluwer Academic Publishers.

Roberts, J. S. (1996). *Item response theory approaches to attitude measurement.* (Doctoral Dissertation), University of South Carolina, Columbia, South Carolina.

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 IRT 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 Estimates 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 Response Scale. *Applied Psychological Measurement, 20*(3), 231-255.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity Issues in the Likert and Thurstone Approaches to Attitude Measurement. *Educational and Psychological Measurement, 59*(2), 211-233.

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.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item repsonding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(5).

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology, 94*(5).

Thompson, V. M. (2014). *Marginal Bayesian estimation in the multidimensional generalized graded unfolding model*. Georgia Institute of Technology. Atlanta, GA.

Thurstone, L. L. (1928). Attitudes Can Be Measured. *The American Journal of Sociology, 33*(4), 529-554.

Van Schurr, W. H., & Kiers, H. A. (1994). Why factor analysis often is the incocrect model for analyzing bipolar concepts, and what model to use instead. *Applied Psychological Measurement*, 97-110.

Wang, W., de la Torre, J., & Drasgow, F. (2015). MCMC GGUM: A new computer program for estimating unfolding IRT models. *Applied Psychological Measurement, 39*(2), 160-161.

Wieland, A. P. (1990). *Evolving controls for unstable systems.* Paper presented at the Connectionist Models: Proceedings of the 1990 Summer School, Morgan Kaufmann, San Mateo, CA.

Zhang, J. (2005). *Estimating multidimensional item response models with mixed structure (ETS Research Report 05-04).* Retrieved from Princeton, NJ: Educational Testing Service: