A BAYESIAN MARKOV CHAIN MONTE CARLO APPROACH TO THE GENERALIZED GRADED UNFOLDING MODEL ESTIMATION:

THE FUTURE OF NON-COGNITIVE MEASUREMENT

BY

WEI WANG

B.S., Beijing Normal University, 2006

M.A., University of Illinois at Urbana-Champaign, 2010

M.S., University of Illinois at Urbana-Champaign, 2010

Dissertation

Submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy in Psychology

in the Graduate College of the

University of Illinois at Urbana-Champaign, 2013

Urbana, Illinois

Doctoral committee

Professor Fritz Drasgow, Chair

Associate Professor Daniel A. Newman

Professor Hua-Hua Chang

Professor Jeffrey A. Douglas

Professor Jim Rounds

# ABSTRACT

Accurately measuring individual differences underpins psychological research, educational and clinical decision-making, personnel selection and managerial practices. Previous research has concluded that ideal point models are more appropriate for measuring *non-cognitive* variables such as personality, vocational interests, attitudes, person-environment fit (e.g., person-job fit), etc. Although a couple of ideal point IRT models have been proposed in the literature, the only one model with executable estimation software available to the public is the generalized graded unfolding item response model (GGUM) and its corresponding software GGUM2004. However, this software unfortunately possesses serious estimation deficiencies due to the marginal maximum likelihood (MML) method that it utilizes. Therefore, this dissertation research is aimed at developing a new computer program estimating the GGUM model by using a state-of-the-art estimation method––Bayesian Markov Chain Monte Carlo estimation. A series of studies will be conducted to test the estimation accuracy of the new software and compare its properties to GGUM2004. It is expected that the MCMC GGUM will outperform GGUM2004 in many ways, especially for detecting differential functioning items. Implications and future research directions are discussed.

# TABLE OF CONTENTS

[ABSTRACT ii](#_Toc348301023)

[TABLE OF CONTENTS iii](#_Toc348301024)

[LIST OF TABLES vi](#_Toc348301025)

[LIST OF FIGURES vii](#_Toc348301026)

[INTRODUCTION 1](#_Toc348301027)

[A Historical Overview of Modern Psychological Measurement Advances 3](#_Toc348301028)

[Ideal Point Response Process IRT Models 6](#_Toc348301029)

[The Rationale and Importance of Using Ideal Point Models for Non-Cognitive Measurement 7](#_Toc348301030)

[The Currently Available Ideal Point IRT Model and Its Estimation Deficiency 9](#_Toc348301031)

[Summary 13](#_Toc348301032)

[BAYESIAN MARKOV CHAIN MONTE CARLO AS A NEW ESTIMATION APPROACH 14](#_Toc348301033)

[A Brief Introduction to Markov Chain Monte Carlo Estimation 14](#_Toc348301034)

[The Bayesian Approach to Markov Chain Monte Carlo Estimation 17](#_Toc348301035)

[The Advantages of Using the Bayesian MCMC Approach to Estimating the GGUM Model 18](#_Toc348301036)

[Summary 20](#_Toc348301037)

[THE DEVELOPMENT OF THE MCMC GGUM SOFTWARE 21](#_Toc348301038)

[Building the Likelihood Function 21](#_Toc348301039)

[Identifying the Prior, Posterior, and Conditional Distribution Functions 22](#_Toc348301040)

[Constructing the MCMC Iterations 24](#_Toc348301041)

[Dealing with Missing Values 27](#_Toc348301042)

[Analyzing Output to Generate Estimates 28](#_Toc348301043)

[Programming environment 28](#_Toc348301044)

[Summary 29](#_Toc348301045)

[STUDY I: ITEM PARAMETER ESTIMATION ACCURACY 30](#_Toc348301046)

[Introduction 30](#_Toc348301047)

[Methods 31](#_Toc348301048)

[Results 34](#_Toc348301049)

[Summary 34](#_Toc348301050)

[STUDY II: PERSON PARAMETER ESTIMATION ACCURACY 35](#_Toc348301051)

[Introduction: 35](#_Toc348301052)

[Methods 35](#_Toc348301053)

[Results 35](#_Toc348301054)

[Summary 35](#_Toc348301055)

[STUDY III: DETECTING DIFFERENTIAL FUNCTIONING ITEMS 36](#_Toc348301056)

[Introduction: 36](#_Toc348301057)

[Method 37](#_Toc348301058)

[Results 40](#_Toc348301059)

[Summary 40](#_Toc348301060)

[STUDY IV: APPLICATION OF THE NEW SOFTWARE TO REAL DATASETS 41](#_Toc348301061)

[Introduction: Testing Goals 41](#_Toc348301062)

[Methods 42](#_Toc348301063)

[Analysis and Results 43](#_Toc348301064)

[Summary 43](#_Toc348301065)

[GENERAL DISCUSSION AND CONCLUSIONS 44](#_Toc348301066)

[Future Research 44](#_Toc348301067)

[REFERENCES 45](#_Toc348301068)

[TABLES 56](#_Toc348301069)

[FIGURES 59](#_Toc348301073)

# LIST OF TABLES

[Table 1 Estimates by GGUM2004 for a Personality Scale of Industriousness 56](#_Toc348301070)

[Table 2 Type I Error Rates and Power for DIF Detection by GGUM2004 When Impact = .50 *SD* 57](#_Toc348301071)

[Table 3 Sample characteristics of the two subject groups 58](#_Toc348301072)

# LIST OF FIGURES

[Figure 1. A Typical Item Response Function (IRF) for a Dominance Model 59](#_Toc348301074)

[Figure 2. The Pacifism-Militarism Continuum and Six Item Locations on the Continuum 60](#_Toc348301075)

[Figure 3. Item Endorsement Frequency in Attitude Measurement 61](#_Toc348301076)

[Figure 4. Response Probability from Thurstone (1929) 62](#_Toc348301077)

[Figure 5. A Typical Item Response Function (IRF) for an Ideal Point Model with Neutral Location 63](#_Toc348301078)

[Figure 6. The Item Characteristic Curve for Item #8 in Table 1 64](#_Toc348301079)

[Figure 7. The Item Characteristic Curve for Item #13 in Table 1 65](#_Toc348301080)

[Figure 8. The Item Characteristic Curve for Item #17 in Table 1 66](#_Toc348301081)

[Figure 9. A Comparison of the Normal Distribution and a Four-Parameter Beta Distribution *Beta*(2, 2, –5, 5) 67](#_Toc348301082)

**CHPATER 1**

# INTRODUCTION

Accurate measurement is pivotal for psychological research, educational and clinical decision making (Cohen & Swerdlik, 2002), and personnel selection (Hough & Oswald, 2000) and managerial practices. Indeed, accurately measuring and comparing individual differences underpins individual difference theory (Ackerman & Humphreys, 1990) and these processes are fundamentally important to the study of domains such as job performance (Motowidlo, Borman, & Schmit, 1997) and work motivation (Mitchell & Daniels, 2003). The most recent advances in measurement science have led to the conclusion that the ideal point process is the most appropriate for measuring non-cognitive variables, including typical behaviors and traits such as personality, vocational interests, and attitudes (Drasgow, Chernyshenko, & Stark, 2010a, 2010b; Roberts, Laughlin, & Wedell, 1999; Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009; Tay & Drasgow, 2012).

Although a couple of ideal point IRT models have been proposed in the literature, the only model with executable estimation software available to the public is the generalized graded unfolding item response model (GGUM) and its corresponding software GGUM2004. However, this software unfortunately possesses serious estimation deficiencies due to the marginal maximum likelihood (MML) method that it utilizes. For example, it generates unreasonably large standard errors which lead to fundamental problems in detecting differential functioning items, especially under conditions with mean differences across the groups being compared (i.e., impact; Wang, Tay, & Drasgow, in press). Therefore, this dissertation research aims to develop a new computer program for estimating the GGUM model by using a state-of-the-art estimation method––Bayesian Markov Chain Monte Carlo estimation, and will conduct a series of simulation studies as well as analyses of real datasets to test the estimation accuracy of the new software and compare it to GGUM2004.

The dissertation has eight chapters. Chapter 1 presents a brief overview of the historical development of applied psychological measurement and the most recent advances in this area. It also contrasts two different response processes––dominance and ideal point––and highlights the rationale for using ideal point response models for measuring non-cognitive individual differences. Practical estimation problems with the existing software are also described. Chapter 2 proposes the Bayesian Markov Chain Monte Carlo method as a new approach to estimating the most widely used ideal point model: the generalized graded unfolding model (GGUM). Chapter 3 elaborates the technical details underpinning the new executable software MCMC GGUM, including the details of item parameter estimation and latent trait estimation. The programming environment is also outlined.

Chapters 4 to 7 present a series of studies testing the new MCMC GGUM software. Specifically, Chapter 4 examines estimation accuracy by comparing the estimated item parameters with the generated/targeted values in a series of simulation studies by manipulating various MCMC-related and data characteristic factors including prior distributions, informativeness of starting values, number of MCMC iterations, sample size, scale length, and number of categories. Chapter 5 investigates the estimation accuracy of person parameters. Chapter 6 explores the capacity of the new estimation approach to detecting item differential functioning (DIF), especially under the conditions where impact (i.e., group difference) exists. Chapter 7 uses real data to further test the performance of the new estimation software and DIF analysis. Finally, Chapter 8 discusses the strengths and limitations of the dissertation study and provides future directions for the research in this area.

## A Historical Overview of Modern Psychological Measurement Advances

Measuring individual differences has a history of thousands of years. It began in China about 3,000 years ago when an emperor decided to assess the competency of his officials. This government-developed measurement method gradually evolved a sophisticated system with a multistage process to select for various government administrative positions, and this measurement covered a wide range of topics including music, horsemanship, civil law, writing, Confucian principles, and knowledge of public and private ceremonies (see Du Bois, 1970, for a detailed overview of this ancient measurement system). The system had been in use for 3,000 years until 1905, when Britain and the United State started to develop civil-service exams as a fair way of selecting applicants for government jobs.

However, the modern development of psychological measurement has a mere 100 year’s history. Rooted in psychophysics from the late 19th century, which experimentally studied the lawful relation between the measure of a physical stimulus and the measure of an observer’s sensation, the origin of psychometrics indeed focused on individual differences in physical features and physical abilities. For example, Frances Galton, the widely renowned pioneer of measurement, collected data of bodily dimensions on over 10,000 individuals in London for six years after the establishment of his Anthropometric Laboratory in 1884. Later Galton’s American supporter James McKeen Cattell studied the measurement of hand strength, rate of movement, reaction time, etc, which he referred as a “mental test” (Cattell, 1890).

The real mental testing, however, started from Alfred Binet and Théophile Simon in France (1905) who developed the first test that directly measured an individual’s cognitive ability as intelligence. This endeavor of measuring individual cognitive ability was advanced by many prominent psychometricians, such as Charles Spearman (1904), who analyzed the data from a battery of cognitive tests and proposed the concept of the general intelligence factor, *g*. Lewis M. Terman (1916) introduced the idea of the intelligence quotient (IQ) and proposed a new psychometric method for its calculation. Arthur S. Otis developed procedures to score multiple-choice items which enabled group (rather than individual) testing. Of all the early psychometricians, Louis Leon Thurstone perhaps made the most significant contributions to measuring individual differences. He founded the American Council on Education (ACE), developed methods for measuring attitudes, and studied the structure of cognitive ability (Thurstone, 1924), where he found several primary factors. Moreover, he devised methods for multiple factor analysis (Thurstone, 1931b, 1935).

Two important psychological measurement theories have been developed for assessing cognitive ability and scoring tests. The first one is classical testing theory, which assumes that the test score *X* is the sum of a true score *T* and a random error *E*: *X* = *T* + *E*. This theory, originated from Spearman (1904) and further developed in Thurstone’s (1931b) book *The Reliability and Validity of Tests*, underpins the estimation of test reliability and validity, and serves as the foundation for much of modern psychological measurement. The other theory is item response theory (IRT), which was first comprehensively introduced in Frederic Lord and Melvin Novick’s *Statistical Theories of Mental Test Scores* (Lord & Novick, 1968). The development of item response theory makes a quantum leap in psychological measurement because it facilitates many important psychometric practices such as equating, automated test assembly, computerized adaptive testing (CAT), differential item functioning (DIF) study, etc. Thus item response theory has drawn tremendous attention in psychometrics in the past few decades. Many key IRT models have been proposed with various functionalitites. For example, Birnbaum (1968) proposed the three-parameter logistic model (3PL) to account for guessing in the testing; Masters (1982) offered the partial-credit model (PCM) and Samejima (1969, 1997) introduced the graded response model (GRM) to analyze items with more than two response categories.

All these IRT models––designed primarily for measuring cognitive abilities––although differing in form, have the same theoretical assumption: the probability that an individual correctly responds to an item is *positively* related to his/her ability, denoted as . That is, the higher one’s ability is, the higher the probability that he/she responds to the item correctly or endorses the item positively. This relationship is depicted in Figure 1. Thus an individual with a high level of ability likely answers *all* the easy items correctly, answers *all* the moderately difficult items correctly, and *some* of the most difficult items correctly, which suggests that the individual *dominates* the easy and moderately difficult items. This is like a weight lifter in a clean and jerk competition (Drasgow, Chernyshenko, & Stark, 2010a): a stronger weight lifter would dominate the weights under his/her strength and have a high probability of lifting the weight. These response processes were called dominance response processes by Coombs (1964). Correspondingly, these IRT models are considered as dominance models and they are widely used for measuring cognitive abilities.

However, more and more recent research has demonstrated that cognitive ability dominance IRT models are ill-suited for measuring *non-cognitive* variables, which typically require introspection or capture self-reported typical behaviors and involve a response process that Coombs (1964) labeled as an ideal point response process (Drasgow, Chernyshenko, & Stark, 2010a, 2010b; Roberts, Laughlin, & Wedell, 1999; Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009; Tay & Drasgow, 2012).

## Ideal Point Response Process IRT Models

Perhaps because the history of psychometrics has been overwhelmingly dominated by *cognitive* measurement, research on *non-cognitive* measurement has been largely ignored. Nevertheless, Thurstone (1927, 1928, and 1929) pioneered research on non-cognitive measurement through his well-known work on the measurement of attitudes, which was later widely adopted as measurement tools for research in social psychology and applied psychology. His landmark publication, titled “*Attitudes Can Be Measured*,” articulated eight steps to measure attitudes. Taking the militarism-pacifism attitude as an example, Thurstone (1928) believed that there existed an attitudinal continuum from extreme pacifism to extreme militarism and each item (i.e., attitudinal statement) measured a specific attitude strength located on the continuum (see Figure 2). Importantly, on an attitudinal continuum (e.g., on a scale from value 0 to value 8, where 0 represented extreme pacifism and 8 represented extreme militarism), Thurstone noticed “those readers who indorse statements in the vicinity of 4.0 on the scale will not often indorse statements that are very far away from that point on the scale,” which resulted in the endorsement frequencies presented in Figure 3. Thurstone (1929) further argued that, mathematically, for *N*1 people with an attitude value of *S*1, the probability for these people to endorse another attitude statement with an attitude value *S*2 was inversely related to |*S*2–*S*1|, which is illustrated in Figure 4 and underpins the modern research on ideal point response process item response theory models.

Unlike dominance IRT models, ideal point models assume that the probability of endorsing an item is *negatively* related to , where denotes the standing of person *i* on the latent trait and  denotes the extremity of item *j*. The item endorsement probability is consistent with the pattern depicted in Figure 5. Hence ideal point models posit that individuals are most likely to endorse items that are closest to their latent trait standing, and that the probability of a positive response is non-monotonic and is the highest when the item location matches the latent trait. A typical item response function (IRF) for an ideal point item with a neutral location is presented in Figure 5. Intuitively, such relational model is more appropriate for measuring non-cognitive variables. For example, an item “I enjoy chatting quietly with a friend at a café” measuring extraversion personality is likely to be rejected by both groups of individuals who are too introverted, because they are uncomfortable in public places, and those who are too extraverted, because they find chatting quietly at a café is boring (Drasgow, Chernyshenko, & Stark, 2010a).

## The Rationale and Importance of Using Ideal Point Models for Non-Cognitive Measurement

As mentioned, the cognitive ability focused dominance IRT models are ill-suited for measuring non-cognitive variables, which usually require introspection or self-report of typical behaviors. Indeed, more and more research has concluded that an ideal point response process is appropriate for non-cognitive measurement (Drasgow, Chernyshenko, & Stark, 2010a, 2010b; Roberts, Laughlin, & Wedell, 1999; Stark, Chernyshenko, Drasgow, & Williams, 2006; Tay, Drasgow, Rounds, & Williams, 2009; Tay & Drasgow, 2012), and the rationale is two-fold, from both theoretical and empirical perspectives.

Theoretically, it is believed that individuals use introspective cognitive processes when rating items, asking themselves “Does this statement closely describe me?” In this introspection process, an individual considers his/her behaviors, attitudes, feelings, or whatever that is being assessed, and also consider what the item asks (Zinnes & Griggs, 1974). Therefore individuals actively compare their own standing with the item location, the closer an item’s location on the latent trait continuum to the individual’s standing, the greater the probability that the individual endorses the item, and the probability of endorsing an item decreases as items’ locations are further away from an individual’s ideal point (Drasgow, Chernyshenko, & Stark, 2010a). The example item “I enjoy chatting quietly with a friend in a café” given in the last section well demonstrates this rationale. Since this item can be rejected by both high introverted and high extraverted individuals, it violates the monotonicity assumption of dominance models (i.e., that the item response function is monotonically increasing). Moreover, when measuring non-cognitive variables, dominance models can only handle with items with extreme positions (i.e., with either very low or very high item locations); intermediate items are considered as bad items because they are equivocal and thus such items are not permitted. For example, an item “I don’t believe in capital punishment but I am not sure it isn’t necessary” measuring attitude towards capital punishment is not allowed for dominance models because of “equivocality” by the dominance models’ standards (Andrich, 1996). However, it is possible that an individual has such an equivocal attitude, which is well documented as attitudinal ambivalence in the attitude literature, as one typically holds both negative and positive evaluations towards an entity (Fabrigar, MacDonald, & Wegener, 2005; Kaplan, 1972; Scott, 1969; Thomposon, Zanna, & Grifin, 1995).

Empirically, Chernyshenko, Stark, Chan, Drasgow and Williams (2001) fitted several IRT models to data from the Sixteen Personality Factor (16PF) Questionnaire, and found the best fitted model was not monotonic, which is the hallmark of dominance models, but, instead, found non-monotonies in the probabilities of endorsement, suggesting an ideal point response process. In addition, a study by Carter and Dalal (2010) has demonstrated that, because of its unique flexibility, an ideal point model fits work satisfaction data better than dominance models. Tay, Ali, Drasgow and Williams (2011) further conducted a simulation study and found that ideal point models are not simply more flexible (e.g., fit many types of response), but their goodness of fit index had substantial power to detect model misspecification (fitting an ideal point model to dominance data or vice versa).

More importantly, using dominance models for measuring non-cognitive variables may inevitably lead to mistakes and inaccuracy. For example, the factor analysis techniques are based on the dominance response process assumption. However, Davison (1977) found that factor analysis of a set of unidimentional ideal point items produced two factors (see also Tay & Drasgow, 2012). Thus misspecified models lead to spurious results. Similar mistakes can result if we use dominance IRT models to conduct a differential item functioning (DIF) analysis. In addition, intermediate items in scale development process are likely have to be removed because of low dominance model statistics (e.g., low item-total correlations), which results in great waste and may even lead less accurate measurement results. Therefore, it is necessary and important to use ideal point models for non-cognitive measurement.

## The Currently Available Ideal Point IRT Model and Its Estimation Problems

Only a few ideal point IRT models have been proposed and developed in the psychometric literature. The most widely used ideal point model is the generalized graded unfolding model (GGUM; Roberts, Donoghue, Laughlin, 2000, 2002), which has the following form

|  |  |
| --- | --- |
| , | (1) |

where *j* denotes the *j*th item, *Zj* = a random variable denoting the response to the *i*th item, *z* = 0, 1, 2, …,*C*, where *z* is the observed response, 0 represents the strongest level of disagreement, *C* represents the strongest level of agreement and *C* = the number of observable response categories minus 1, and *M* = 2 × *C* + 1. *θi* = the location of *i*th individual on the latent continuum, *δj* = the location of the *j*th item on the latent continuum, *αj* = the discrimination of the *j*th item, *τjk* = the *k*th subjective category threshold parameter associated with the *j*th item, Based on the GGUM model, Roberts and colleagues developed a computer program to estimate item parameters and latent abilities. This software was first published in 2001 as GGUM2000 (Version 1.0; Roberts, 2001) and then in 2006 as GGUM2004 (Version 1.1; Roberts, Fang, Cui, & Wang, 2006). Both GGUM2000 and GGUM2004 use marginal maximum likelihood (MML) estimation and they both estimate polytomously scored unidimensional items. This is the only software publically available so far for estimating ideal point models and has led to important research advances.

Unfortunately, the only-available software for ideal point model estimation has shown some serious deficiencies. First, it sometimes yields unacceptably large estimates for *δ*’s and their associated standard errors, especially for the items with a low discrimination parameter *α*. For instance, when  < .5,  can be greater than 10 and *SE*() can be greater than 100 (see Table 1 for sample results from a real dataset of an industriousness scale, a subscale of the Big Five personality measure). These inflated estimates by GGUM2000 were also reported by de la Torre, Stark, & Chernyshenko (2006), where GGUM2000 generated *SE*() as large as 31.37 and *SE*() as large as 32.49 (see Table 2 in de la Torre, Stark, & Chernyshenko, 2006).

The reason for such large estimates by GGUM2000 and GGUM2004 likely results from the estimation method that is used: marginal maximum likelihood (MML). This estimation method requires taking second derivatives and matrix inversion, which can encounter difficulties when estimating complicated models and in estimating the standard error (*SE*), especially when the ICC is flat. To illustrate this problem with MML, I present the item curves of expected scores for Items #8, #13, #17 from Table 1 in Figures 6, 7, and 8, respectively. It is noted that when the discrimination parameter is small, the curves tends to be very flat, which leads to difficulties when taking a derivative to determine the delta estimate (). Not surprisingly, these items with small discrimination parameters yielded large estimates for location parameters () and their standard errors *SE*(), as well as the threshold parameter estimates . For example, for Item #8 with  = .0089, GGUM2004 generated infinite estimates for the location parameters () and its standard error *SE*(). For Item #13 with  = .0530, GGUM2004 generated –26.3010 and 729.5049 as the estimates for the location parameters () and its standard error *SE*() , respectively.

Because accurate standard errors and covariances are crucial for comparing the functioning of items across groups, not surprisingly, Wang, Tay, and Drasgow (in press) found DIF detection with GGUM2004 problematic when there was impact greater than .25. Impact occurs when two groups or populations differ in the measured mean of a construct of interest. Previous research (e.g., Mullis, Dossey, Owen & Phillips, 1993) shows that it is quite common for the reference and focal groups to have a mean discrepancy of one standard deviation (1 *SD*). The impact issue is especially prevalent for constructs that require ideal point models. These constructs include a variety of personality traits, attitudes and vocational interests; they often vary quite a lot among subpopulations (e.g., genders, cultures, SES classes, age subgroups, etc.). For example, data from over 200,000 participants from 53 nations have revealed that women score significantly higher than men in extraversion, agreeableness and neuroticism, and that women and men have different occupational preferences (Lippa, 2010). Recent data (McCrae et al., 2010) from 24 cultures have shown that Westerners are significantly more extraverted than non-Westerners, and Hong Kong Chinese and Japanese are significantly more neurotic than mainland Chinese and South Koreans, indicating that subcultures differ on some personality traits even in the same geographic area.

Besides the GGUM model, there are three other ideal point IRT models in the literature; nevertheless they all have clear weaknesses compared to the GGUM model and none of them has been popularly used or extensively studied. These three models include the squared simple logistic (SSLM) model proposed by Andrich (1988), the hyperbolic cosine (HCM) model that was independently developed by Andrich and Luo (1993) and Verhelst and Verstralen (1993), and the normal probability density function (Normal PDF) model introduced by Maydeu-Olivares, Hernandez and McDonald (2006). The SSLM model has the form

|  |  |  |
| --- | --- | --- |
|  | , | (2) |

and can only accommodate binary data and estimate one item parameter. The HCM model has the form

|  |  |  |
| --- | --- | --- |
|  | , | (3) |

where  is called the item-unit parameter (Andrich & Luo, 1993) and it measures the likelihood that a respondent located at *θ* endorses item *j*. Again, this model only accommodates binary data yet estimates two item parameters. Although the HCM was extended to the general hyperbolic cosine model (GHCM; Andrich, 1996) to model polytomous Likert scales, nevertheless, it is not widely used in psychometrics and very few papers in the literature have examined this model. Lastly, the Normal PDF model has the form

|  |  |  |
| --- | --- | --- |
|  | . | (4) |

Although it fits multidimensional ideal point models, it only accommodates binary data. In addition, estimation software for these three ideal models is not widely available for researchers and practitioners. Therefore, this dissertation focuses on the GGUM model and proposes a new estimation method. Specifically, the dissertation adopts a Bayesian Markov Chain Monte Carlo approach to estimate parameters for the GGUM model. It is hoped that this new estimation approach overcomes the problems encountered by GGUM2000 and GGUM2004.

## Summary

The historical legacy of psychological testing has overwhelmingly focused on the measurement of cognitive abilities, with little attention to non-cognitive measurement. Perhaps another reason for little study of non-cognitive variables is that researchers have taken for granted that dominance models would be well suited for them. However, as discussed in this chapter, more and more evidence, both theoretical and empirical, has converged on the conclusion that the dominance models are indeed ill-suited for the measurement of non-cognitive variables such as personality, attitudes, vocational interests, etc. The generalized graded unfolding model (GGUM) and its corresponding estimation computer program GGUM2000 and GGUM2004 constitute important advances for the measurement of non-cognitive traits. However, it appears that the estimation method implemented in GGUM2000 and GGUM2004 could benefit from Bayes priors. Therefore, this dissertation aims to explore a new estimation approach to the GGUM model that adopts a Bayesian Markov Chain Monte Carlo (MCMC) approach to estimate parameters for the GGUM model.

**CHAPTER 2**

# BAYESIAN MARKOV CHAIN MONTE CARLO AS A NEW ESTIMATION APPROACH

## A Brief Introduction to Markov Chain Monte Carlo Estimation

Markov chain Monte Carlo (MCMC) is a promising estimation method that has been extensively used in physics and has become increasingly popular in statistics (Chib & Greenburg, 1995; Gilks, Richardson & Spiegelhalter, 1996). Although its full capacity for estimation is yet to be explored, several pioneering psychometricians have been amazed by its effectiveness in estimating complicated models (e.g., Béguin, & Glas, 2001; Bolt & Lall, 2003; de la Torre & Douglas, 2004; de la Torre, Stark, Chernyshenko, 2006; Johnson & Junker, 2003; Kim, 2001; Patz & Junker, 1999a, 1999b; Shi & Lee, 1998). Edwards (2010), for example, argued that MCMC will be an important estimation method in the decades to come. In this section I will briefly introduce this method and its associated algorithms; more detailed introductions to MCMC estimation methods are available from many excellent sources, such as *Markov Chain Monte Carlo in Practice* by Gilks, Richardson and Spiegelhalter (1996) and *Monte Carlo Method in Bayesian Computation* by Chen, Shao, and Ibrahim (2000).

Originally, the Monte Carlo method was proposed for estimating the expectation of a function *f*(*X*). This method assumes that, given a function *f*(*X*), its expectation *E*[*f*(*X*)] can be approximated as

|  |  |  |
| --- | --- | --- |
|  | , | (5) |

where the series of {*Xt*}, *t* = 1, …, *n*, is a set of randomly and independently sampled values from , which is the support of the function *f*(*X*). When the sample size *n* is large enough, the estimation approximation can become as accurate as desired because of the law of large numbers.

The *Markov chain* type of Monte Carlo (MCMC) changes the manner of generating the series {*Xt*} from complete independence to mild dependence: the generation of the next state’s value *Xt+*1 slightly depends on the current state’s value *Xt* of the chain yet it is conditionally independent of all the other previous states’ values {*X*0, *X*1, *…*, *Xt*–1}. Mathematically,, where the conditional probability , or more generally,, is called the *transition kernel* of the chain. It is used to restrict the generation of the next value *Xt+*1. The transition kernel, sometimes also called the *moving probability*, or the *probability of move*, is considered as a general mechanism to describing the probability that the current chain status *X*t moves to (or is updated to) the next state’s value *Xt*+1. As *t* sufficiently increases, the conditionally sampled values {*Xt*} will increasingly look like dependent samples from *f*(*X*), especially after a sufficiently long *burn-in* (e.g., *m* iterations) is discarded, then the remaining samples {*Xt*}, *t* = *m*+1, …, *n*, will highly approximate samples from *f*(*X*), thus these sampled values {*Xt*}, *t* = *m*+1, …, *n*, will be appropriate to be used to appropriately describe the distribution of *f*(*X*). This is the rationale of the MCMC estimation method. As such, the expectation of *f*(*X*) can be better estimated by a Markov chain with a burn-in:

|  |  |  |
| --- | --- | --- |
|  | . | (6) |

Although the idea of MCMC estimation method is straightforward, it is crucially challenging to find such an algorithm to generate the sample values {*Xt*} that eventually approximate the distribution of *f*(*X*) as the chain is sufficiently long. The well-known algorithm for generating such a Markov chain is the Metropolis–Hastings (M–H) algorithm, which was originally proposed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and later generalized by Hastings (1970). This algorithm has been widely used in physics and statistics. According to the Metropolis–Hastings (M–H) algorithm, given the current state’s value *Xt* at iteration *t*, the sampling of the next state’s value *Xt*+1 is based on a probability  that is conditional on the current value *Xt*. Specifically, to choose a value for the next state *Xt*+1, a candidate value Y is first sampled from a proposal distribution, then the moving probability (i.e., transition kernel) that the current value *Xt* can be updated to the candidate value *Y* is  and calculated by

|  |  |  |
| --- | --- | --- |
|  | , | (7) |

where  is commonly considered as a proposed prior distribution (introduced in the following section) and  is probability conditional on the value of *X*. The details of ,  and  for the MCMC GGUM are introduced in Chapter 3. After the moving probability is calculated, then a random value *u* is generated from a uniform distribution *U*(0, 1). If *u* ≤ , then the value of the next state is set to *Xt*+1 = *Y*. Otherwise, the chain does not move and goes back the current state, that is, *Xt*+1 = *Xt*, and a new candidate from the proposal distribution is generated to calculate the moving probability. This process is repeated for a sufficient number of iterations to generate a long enough chain of {*Xt*}. And all the values of {*Xt*} are recorded for use in estimating the distribution of the posterior function.

Once the Monte Carlo output {*Xt*}, *t* = *m*+1, …, *n*, is generated, the means and variances (and even the correlations if multiple series of {*Xt h*}where *h* ≥ 2 are generated) can be estimated by using

|  |  |  |
| --- | --- | --- |
|  | , | (8) |
| and | . | (9) |

## The Bayesian Approach to Markov Chain Monte Carlo Estimation

Two approaches have been developed to estimate parameters (or a parameter vector ): the frequentist as the traditional approach and the Bayesian as the newer approach. Although both approaches build a model based on the observed data , the frequentist approach considers the parameter to be fixed yet unknown, whereas the Bayesian approach treats both the data  and parameters  as random and builds a joint distribution model for both the data  and parameters . Thus the Bayesian approach has advantage over the frequentist approach in that it takes account of the distribution information of the parameters . Before the observed data  is obtained, statisticians or psychometricians or subject matter experts usually have a good guess about the distribution of the parameters based on theories or previous research. This guess about the distribution of the parameters is called the prior distribution, denoted as . Once the data  is obtained, an updated distribution is then built based on the observed data . This updated distribution, denoted as *P*(*θ*|*X*=*x*), is called the posterior distribution, which is the probability of the parameters *θ* given the evidence of the data *X*. The Bayesian approach to estimation is based on the posterior distribution with the consideration of the prior distribution. This approach has tremendous advantages over the traditional frequentist approach. The prior distributions provided by subject matter experts are essential in building the models for estimation. In many cases, the guess about the prior distributions can be primarily theory based with little subjective significance attached, thus it avoids subjective influences and can provide important and useful results and insights (Bickel & Doksum, 2001). Because of its unique advantage, the Bayesian approach has been widely adopted in estimating complicated models and become increasingly popular in psychometrics.

In the Bayesian approach to MCMC estimation, the posterior distribution is determined by the prior distribution of *π*() together with the likelihood function :

|  |  |  |
| --- | --- | --- |
|  | . | (10) |

The likelihood function is the probability that the data *X*=*x* is observed given the parameter *θ*. Typically, the likelihood function is denoted as *L*(X), thus the posterior distribution can be written as

|  |  |  |
| --- | --- | --- |
|  | . | (11) |

## The Advantages of Using the Bayesian MCMC Approach to Estimating the GGUM Model

The Bayesian MCMC approach has been found to estimate various IRT models accurately (de la Torre & Douglas, 2004; de la Torre, Stark, & Chernyshenko, 2006; Johnson & Junker, 2003; Kim, 2001; Patz & Junker, 1999a, 1999b). Patz and Junker (1999a) first introduced this approach to estimate the two-parameter logistic (2PL) IRT model and they (Patz & Junker, 1999b) then extended this new estimation approach to more complicated IRT models including the 3PL model, the generalized partial credit model (GPCM), and the generalized linear logistic test model (GLLTM). They believed the application of MCMC to IRT estimation was promising. Kim (2001) used four datasets and the one-parameter logistic model (1PL) to compare various estimation methods including MCMC, conditional maximum likelihood, marginal maximum likelihood, and joint maximum likelihood. He found that item parameter estimates from the four methods were almost identical. de la Torre and Douglas (2004) utilized MCMC to estimate higher-order latent traits models for cognitive diagnosis.

Notably, Johnson and Junker (2003) explored the capacity of the Bayesian MCMC method for estimating the Hyperbolic Cosine (HCM) ideal point IRT model. They argued that any maximum likelihood method for estimating ideal point response models tends to have problems because the likelihood function may display bi-modality, thus maximum likelihood estimation may find the wrong mode or even anti-mode. They found that the Bayesian MCMC methods “have clear advantages over maximum likelihood estimation procedures for unfolding models” (p. 226). In addition, de la Torre, Stark and Chernyshenko (2006) have preliminarily tried using MCMC to estimate the GGUM model. This is the only publication that has estimated the GGUM model parameters by using an MCMC approach. In their simulation study, they found both the MCMC and MML methods produced equally accurate item parameter estimates, but only the MCMC method generated reasonable standard error estimates for all items, whereas the MML method resulted in very large standard errors.

Despite the groundbreaking contributions of de la Torre et al. (2006), their exploration of the MCMC method with the GGUM was preliminary with many important questions untouched. For instance, they did not explore how the estimation accuracy was influenced by various MCMC related factors such as the prior distribution, informativeness of starting values, and the number of iterations, etc. Another important question is whether this estimation method can be used to identify DIF, especially when impact greater is .50 *SD* or larger. Besides these unanswered questions, their software was limited in scope; many important functionalities that are necessary in psychometric applications were not included. For example, because the code was developed for a simulation study, their code only accommodates four-response category data (i.e., *K* = 4) and does not work with missing values. Furthermore, their code does not estimate of person parameters (*θ*). This research, building on the pioneering contributions of de la Torre et al. (2006), aims at creating software with many additional functionalities that are necessary for psychometric research and applications and examining its many properties such as how MCMC-related factors affect estimation accuracy. In addition, this research conducts a series of studies to test the software and examine its performance.

## Summary

This chapter briefly introduces the MCMC estimation method and the Metropolis–Hastings (M–H) algorithm that is commonly used for the MCMC estimation. It also discusses the Bayesian approach to MCMC estimation. This chapter further reviews the IRT literature that explores the MCMC estimation method, highlighting the strengths of this method in estimating ideal point IRT models and stressing problems not yet addressed in the literature.

**CHAPTER 3**

# THE DEVELOPMENT OF THE MCMC GGUM SOFTWARE

## Building the Likelihood Functions

As discussed in Chapter 2, the construction of a MCMC procedure necessitates the calculation of moving probability , where  denotes the current estimate of the parameter at Iteration *t* and  denotes a sampled candidate of the parameter estimate for the next Iteration *t*+1. The calculation of moving probability  further requires the determination of prior distributions, the likelihood functions, and the calculation of the posterior distributions for all the parameters.

Suppose *N* respondents answer *J* ideal point items, the response vector for the respondent *i* is  and the entire response matrix can be denoted as ***X***, and

|  |  |  |
| --- | --- | --- |
|  | . | (12) |

Thus, for the respondent *i*, the likelihood function is

|  |  |  |
| --- | --- | --- |
|  | . | (13) |

And the likelihood function for the entire response matrix is

|  |  |  |
| --- | --- | --- |
|  | , | (14) |

where *k* is the response category selected by the respondent *i* (with a latent value ) on item *j*, and  is probability of respondent *i* (with a latent value ) selecting category *k* on item *j* and is calculated through the GGUM model given in Equation 1.

## Identifying the Prior, Posterior, and Conditional Distribution Functions

**Prior distribution.** In dominance models, the item discrimination parameter, difficulty parameter, and guessing parameter are often assumed to respectively follow lognormal, normal, and beta prior distributions (Baker, 1992; Swaminathan & Gifford, 1986). However, the lognormal, normal, and beta prior distributions may be replaced by the four-parameter beta because it is more flexible and possesses many attractive features (de la Torre et al., 2006; Zeng, 1997). Denoted as Beta (*υ*, *ω*, *a*, *b*), the four-parameter beta distribution can alter the support area [*a*, *b*] by adjusting the parameters *a* and *b*; the variance of the four-parameter beta can also be modified based on the values of *υ* and *ω*, where larger values are associated with smaller variances. For example, Beta (12.66, 12.66, –5, 5) approximates the standard normal distribution *N*(0, 1) very well, with the maximum difference between the two distributions merely about .005. And when *υ* = *ω* = 1, the Beta (1, 1, *a*, *b*) is a uniform distribution *U*(*a*, *b*) with the boundaries of *a* and *b*. Thus the normal distribution and uniform distribution are just special cases of the four-parameter beta distribution. In addition, the symmetry of the four-parameter beta distribution is changeable based on the values of *υ* and *ω*: when *υ* = *ω*, the distribution is symmetric; when *υ* > *ω*, the distribution is left skewed, and when *υ* < *ω*, the distribution is right skewed. This flexibility is important for prior distributions for estimating ideal point model parameters. Therefore, this research adopts the four-parameter beta distribution for the parameters in the GGUM model.

Specifically, the parameters , , , , …,  in the GGUM model, as well as person parameter , are assumed to follow the four-parameter beta prior distributions as below:

|  |  |  |
| --- | --- | --- |
|  | , | (15) |
|  | , | (16) |
|  | , | (17) |
| and | . | (18) |

**Posterior distribution.** According to Equation 10, the posterior distribution of the parameters  can be written as

|  |  |
| --- | --- |
| , | (19) |

and further,

|  |  |  |
| --- | --- | --- |
|  | , | (20) |

where *N* represents the number of respondents, *J* represents the number of items, and *K* represents the number of response categories;, , , , , ; parameters *θ*, *α*, *δ*, *τ* are introduced in Equation 1. Equation 20 is the object of the Bayesian analyses.

**Conditional distribution.** Notably, the posterior function in Equation 20 is too complicated to be analytically evaluated and to calculate the moving probability in order to conduct MCMC iterations, and for mathematical convenience, the samples that are supposed to be drawn from the posterior distribution would be better drawn iteratively from the full conditional distributions of each parameter (de la Torre et al., 2006). Specifically, the full conditional distributions of the parameters *θ*, *α*, *δ*, *τ*1, …, and *τK*–1 are, respectively,

|  |  |  |
| --- | --- | --- |
|  | , | (21) |
|  | , | (22) |
|  | , | (23) |
|  | , | (24) |
| and | . | (25) |

These full conditional distributions are then used to build the Metropolis–Hastings algorithm and grow the MCMC iterations (see the details in the section below).

## Constructing the MCMC Iterations

In constructing the MCMC iterations, it is important to first select the prior distributions and determine the starting values for Iteration 0. Then Iteration *t*+1 is constructed based on Iteration *t* (t ≥ 0). As discussed in the preceding section, the four-parameter beta distribution is used as the prior functions for the item parameters , , , , …, . Following de la Torre et al. (2006), Beta(1.5, 1.5, .25, 4), Beta(2, 2, –5, 5), and Beta(2, 2, –6, 6) are used for the priors of , , and , respectively. These distributions are much flatter than the normal distribution (see the comparison of the standard normal distribution and Beta(2, 2, –5, 5) in Figure 9). However, this research will also vary the prior distributions to examine the priors’ impact on estimation accuracy. For example, if one item is rated by subject matter experts to have a high (or low) level location, then using Beta(2, 2, 0, 5) (or Beta(2, 2, –5, 0)) as the prior distribution for item  may yield a more precise estimate, because the prior function provides a more accurate support interval. Similarly, if an item is rated as an intermediate item, using Beta(2, 2, –3, 3) may help produce better results, because the support interval is adjusted based on subject matter experts’ judgment. Moreover, it is known that subject matter experts can accurately discover a personality statement’s extremity (Stark, Chernyshenko, & Guenole, 2011).

For Iteration 0, the starting values are set as follow: , , , , …, . Here *θ* is a vector with *N* elements, *N* is the number of respondents, and all the item parameters are contained in vectors with *J* elements, and *J* is the number of items. For the *δ* vector, all the elements are equally spaced in the interval [–2. 45, 2.45]. This procedure also follows de la Torre et al. (2006).

For Iteration *t*+1 (*t* ≥ 0), a candidate (e.g., ) is first randomly sampled from a normal distribution with mean equal to the value at Iteration *t* (i.e., ). Then Equation 7 is used to calculate the moving probability (i.e., ), the probability that that the current estimate value  can be updated to a new estimate value . Operationally, if the moving probability  ≥ *u*, where *u* is randomly generated from a uniform distribution *U*(0, 1), then the candidate  is set to the value at Iteration *t*+1, which means the chain moves from  to . Otherwise, the chain goes back to the state at Iteration *t*, and the value  is set back to equal to and the chain does not move for this step. Specifically, the calculation of the moving probability for parameters , , , , , …,  are as follow:

For parameter , draw a candidate , and calculate the moving probability

|  |  |
| --- | --- |
| . | (26) |

For parameter , draw a candidate , and calculate the moving probability

|  |  |
| --- | --- |
| . | (27) |

For parameter , draw a candidate , and calculate the moving probability

|  |  |
| --- | --- |
| . | (28) |

For parameter , draw a candidate , and calculate the moving probability

|  |  |
| --- | --- |
| . | (29) |

For parameter , draw a candidate , and calculate the moving probability

|  |  |
| --- | --- |
| . | (30) |

In practice, the standard deviation of the distribution for drawing candidates is set to 0.15 rather than 1, thus the drawn candidate is closer to the current value. In addition, the calculation of the moving probability is based on logs of the above equations, transforming the multiplicative forms to additive forms, which are easier for computer programming. For instance, to calculate  for person *i*’s parameter in Equation 26, it is easier if the values are calculated as

|  |  |  |
| --- | --- | --- |
|  | , | (31) |
| and | . | (32) |

Then the moving probability is calculated by

|  |  |  |
| --- | --- | --- |
|  | , | (33) |

where num denotes the numerator and den denotes the denominator of Equation 26. Calculating the moving probability for item parameter is similar to Equations 31 and 32 except for different subscripts, which deal with columns instead of rows. For example, to calculate  for item *j*’s location parameter in Equation 28, the numerator and denominator are calculated by

|  |  |  |
| --- | --- | --- |
|  | , | (34) |
| and | . | (35) |

And again, the moving probability is then calculated by

|  |  |  |
| --- | --- | --- |
|  | , | (36) |

This process repeats until it finishes all the designated the number of iterations (e.g., 60,000).

## Dealing with Missing Values

Dealing with missing values in a dataset is one of the important features of the new software MCMC GGUM. This feature is important as it is common to have unanswered questions in personality test, and more importantly, the data rearrangement for some DIF detection methods––for example, the free baseline approach and the constrained baseline approach (see Wang, Tay, & Drasgow, in press, for details)––necessitate the functionality of missing values. However, including missing values can interfere with the calculation of likelihoods and further affect the moving probabilities, which are crucial for the MCMC method, so it is crucial for the MCMC GGUM to deal with missing values properly. In programming MCMC GGUM, a strategy wass developed to make sure that missing values are appropriately excluded for the calculation of the likelihoods. Specifically, MCMC GGUM recodes missing values to –9 (or any other numerical values that are not equal to *k*, where *k* = 0, 1, 2, …, *K*–1) and does not include them when calculating the likelihoods (see Equations 31, 32, 34, and 35 for example). Because the missing values do not contribute to the likelihoods in either the numerator or the denominator, the calculation of the moving probability will not be affected.

## Analyzing Output to Generate Estimates

After completing all the iterations, the estimates and the standard errors can be calculated by using Equations 8 and 9, with a designated number of *burn-in* iterations. In this research, the effect of the number of iterations and corresponding burn-in iterations on estimation accuracy will be examined.

## Programming environment

All the programming work in this research is done with the Python programming language. Python is a powerful programming language and it has been more and more popular in computer science. Python was chosen in this research for three main reasons. First, Python code can be easily compiled to generate an executable applications which can be easily distributed and shared with other researchers; second, Python code can be easily called to run in R, which is another programming language that is widely used in statistics; lastly, although Python is a high-level language, when it is implemented in NumPy and PyPy, its speed can be as fast as C, which is a powerful low-level programming language. de la Torre et al.’s (2006) programming code was written in the Ox programming language, and consequently this research will compare the results from Python with those from the Ox code provided by Dr. de la Torre.

## Summary

This chapter elaborates on the technical details of the MCMC GGUM software. Starting with the likelihood functions, this chapter discusses the prior distributions of the parameters and the appropriateness of using the four-parameter Beta distribution. It also details the posterior distributions and full conditional distributions that are used to calculate the moving probability for growing the MCMC chain. Finally, it addresses the missing value issues and the output analysis; it also discusses the programming language issues and strategies. The chapters that follow will present a series of studies, both with simulations and real datasets that will test the performance of this new software.

**CHAPTER 4**

# STUDY I: ITEM PARAMETER ESTIMATION ACCURACY

## Introduction

The goal of this study is to test the new software by examining parameter estimation accuracy and how it might be impacted by various factors. Three factors that are uniquely associated the MCMC method may affect estimation accuracy: (1) item location prior distributions; (2) starting values; and (3) the number of iterations. The prior distributions can vary in both their variance and range (i.e, the four parameters *υ*, *ω*, *a*, and *b* in the Beta distribution); the range of the prior distribution determines the boundaries of the estimates. By appropriately setting the boundaries, the prior distributions can become more informative. For example, if an item is apparently positive in terms of the item location (e.g., the item “I am competitive and play to win” measuring industriousness/achievement can be easily judged as a positive item), then setting the prior boundaries to be greater than 0 is more informative and should facilitate estimation accuracy. In fact, previous research (Stark, Chernyshenko, & Guenole, 2011) has shown that subjective matter experts make relatively precise judgment of item locations. The judgment of item locations can also help determine a good starting point for MCMC, which may further facilitate estimation accuracy and shorten the chains required to obtain desired estimation accuracy. Thus all these factors are examined in this study.

In addition, data characteristics may also affect estimation accuracy. These characteristics include: (a) the sample size (i.e., *N*); (b) scale length (i.e., the number of items *J*), and (c) the number of response categories (i.e., *K*).

The MCMC output will first be examined based on the Gelman and Rubin (1992) criterion. Then the estimates and corresponding standard errors will be calculated, and these results will further compared to the results from GGUM2004 with the MML estimation method. The empirical standard deviations of MCMC and MML estimates based on 50 replications will also be studied and compared.

## Method

**Design.** The study examines seven factors and their manipulations as follow:

(1) Item location prior distribution (two levels: informative vs. noninformative);

(2) Starting values of item locations (two levels: informative vs. noninformative);

(3) Starting values of person parameter (two levels: informative vs. noninformative);

(4) The number of iterations (six levels: 10,000; 20,000; 40,000; 60,000; 80,000; and 100,000, all with the first 25% as burn-in iterations);

(5) Sample size (two levels: 500 vs. 1,000);

(6) Number of items (two levels: 10 vs. 20);

(7) Number of response categories: (two levels: 2 vs. 4).

**Item and person parameters.**In order to better examine estimation bias, this study uses fixed item parameters across all the 50 replications. For the conditions with 10 items, item parameters are obtained from the paper published by de la Torre et al. (2006; see their Tables 1, and 5). Using their published item parameters makes it possible to compare our results with their results. Note that in de la Torre et al.’s (2006) paper, only the three data characteristic factors were examined––sample size (400, 800, and 1,200), number of items (10 vs. 20), and number of response categories (2 vs. 4), without examining the MCMC related factors such as item location prior distribution, starting values of item location and person parameter, and the number of iterations. With the 12 conditions examined in their study, they only reported item parameter estimation accuracy for two conditions: one condition with 2-response categories, 10-items, and 400-simulees, and the other condition with 4-response categories, 10-items, and 400-simulees. Nevertheless, their reported estimation accuracy in the two conditions can be compared in this study. All person parameters are sampled from a *N* (0, 1) distribution. The person parameters are sampled separately for each of the 50 replications.

**Response data generation.**Response data are generated for each replication based on the obtained item parameters and the sampled person parameters. To generate response data, Equation 1 is used to compute the probability that each category of a given item is endorsed by a respondent with a simulated theta value (*θi*). Then, the cumulative probabilities from the first category to the fifth category of an item (i.e., a vector ) are calculated and a number is randomly generated from a uniform distribution *U*(0, 1). The simulated response to an item is determined by the location of the randomly generated in the cumulative probability vector.

**Starting values.** Approximate starting values are provided for item location parameters and person parameters for the conditions where informative starting values are required. In this study, the starting values for the item location parameters are determined by setting the first item’s location parameter to –2.45, and the last item’s location parameter to 2.45, and set all the other items’ location parameter to be about equally and symmetrically spanned in the interval of [–2.45, 2.45]. For example, for the 10-item conditions, the starting values of the item locations are set as [–2.45, –1.90, –1.35, –0.80, –0.25, 0.25, 0.80, 1.35, 1.90, 2.45].

The starting values of person parameters are determined by a rough nonparametric method. This method was first proposed by Post (1992) and further developed by Johnson (2006). This study uses its most simplistic version of this method because all the items are sorted by item location parameters from negative to positive; this data arrangement is possible because either the item location parameters are known in the simulation study, or they can be roughly judged by subject matter experts. The calculation of the starting values for person parameters involves two steps. The first step is to calculate each respondent’s raw score after recoding their responses based on the method proposed by Post (1992). Second, the raw scores of all the respondents are then standardized. And each respondent’ standardized score is used as their starting value.

**Procedures.** The response data generated in each replication is used by both MCMC GGUM and GGUM2004, and the output for each replication is also saved for further analyses and comparison. This study will focus on the estimates, estimation bias, and standard errors, and the empirical standard deviations across 50 replications.

**Convergence diagnostics.** All the separately generated chains will be examined for convergence before further conducting analyses to evaluate estimation accuracy. The convergence diagnostics will be conducted based on Brooks and Gelman’s (1998) method, which was extended from the univariate version by Gelman and Rubin’s (1992) method. This method assumes that the statuary posterior distribution is normal and is based on the comparison of within-chain and between chain variances, which is similar to the analysis of variance (ANOVA) testing method. According to this method, the within-chain covariance matrix *W* and between-chain covariance matrix *B/n* are first calculated by

|  |  |  |
| --- | --- | --- |
|  | , | (37) |
|  | , | (37) |

where  denotes the parameter vector in chain *h* at time *t*, and *n* represents the number of iterations and *m* represents the number of separate chains. The estimated posterior variance-covariance matrix is

|  |  |  |
| --- | --- | --- |
|  | . | (37) |

The convergence diagnostic index is

|  |  |  |
| --- | --- | --- |
|  | , | (37) |

where

|  |  |  |
| --- | --- | --- |
|  | , | (37) |

and index values *R* substantially above 1 indicate lack of convergence.

In this research, this diagnostic is conducted with CODA, an R statistical package that is designed to process MCMC output.

## Results

## Summary

**CHAPTER 5**

# STUDY II: PERSON PARAMETER ESTIMATION ACCURACY

## Introduction:

This study is designed to examine the estimation accuracy of the person parameters (*θ*) that are simultaneously estimated with item parameters. In the literature, there has been no published research trying to estimate and describe the accuracy of estimates of the person parameters for the GGUM model using the MCMC approach. However, Johnson and Junker (2003) found that argued that the MCMC approach was able to derive the posterior distributions of items and persons simultaneously for the hyperbolic cosine model (HCM). In this study, it is expected that MCMC GGUM will also perform well on estimating person parameters.

## Methods

Because person parameters are simultaneously estimated with item parameters, the method for this study is identical with Study I in Chapter 4.

## Results

## Summary

**CHAPTER 6**

# STUDY III: DETECTING DIFFERENTIAL FUNCTIONING ITEMS

## Introduction:

Study 3 was designed to examine the performance of MCMC GGUM estimation for detecting differential item functioning. This study focused conditions with impact greater than .25 *SD*, because serious problems for GGUM2004 DIF detection were found for those conditions (see Wang, Tay, & Drasgow, in press). Impact occurs when the mean of the latent trait in the focal group differs from the mean in the reference group; it is commonly observed in psychological research (Donoghue, Holland & Thayer, 1993; Mullis, Dossey, Owen, & Phillips, 1993). Although impact has little influence on DIF detection for dominance models (e.g., Chang, Mazzeo & Roussos, 1996; Finch, 2005; W-C Wang & Su, 2004a, 2004b), recent research has revealed that it poses a serious threat to DIF detection for ideal point models estimated with GGUM2004 (Wang, Tay, & Drasgow, in press). Therefore, this study aimed to examine if the MCMC GGUM performed well in conditions where DIF detection was problematic.

The DIF detection method used in this study was Lord’s (1977, 1980) chi-square, for it is a well-known and widely used method for detecting DIF (Lim & Drasgow, 1990; McLaughlin & Drasgow, 1987; Raju, Drasgow, & Slinde, 1993) and it has been considered as effective in detecting DIF items (Donoghue & Isham, 1998). However, the method was found to be deficient in detecting DIF item for ideal point models with GGUM2004. As presented in Table 2, Lord’s chi-square had both high Type I error rates and low power for the uniform DIF when impact was .50 *SD*. The main reason for the poor performance of Lord’s chi-square was apparently the large standard errors which are crucial for the method yet poorly estimated by GGUM2004. If the new software accurately estimates standard errors, Lord’s chi-square should work well. Therefore it is expected that with MCMC GGUM, the Lord’s chi-square may be able to effectively detect DIF items even under conditions where impact was greater than .25 *SD*.

## Method

This study examined two variables: (a) type of DIF––No DIF, uniform DIF, and nonuniform DIF; (b) level of impact––no impact (i.e., 0 *SD*), .25 *SD*, .50 *SD*, and 1.00 *SD*. Other settings for this study included the use of 10 items with 4 (40%) DIF items with 4 response categories and 1000 simulees for the both focal and reference groups.

**Item and person parameter generation.**The parameter generation process followed the procedures by Roberts et al. (2002) because they proposed the GGUM model and their simulation procedures has been followed in numerous other studies (e.g., Tay, Ali, Drasgow, & Williams, 2011; Wang, Tay, & Drasgow, in press). Specifically, *αi* was generated from a uniform distribution *U*(.5, 2). The location parameters *δi* were evenly distributed in the interval [-2.5, 2.5]. The responses to each item had *K* categories: 0 to *K*–1, where *K* is the number of categories, and 0 represented the strongest disagreement and *K*–1 represented the strongest agreement. The thresholds () were generated as follow: *τi,K–1* was generated from a uniform (-1.4, -.4) distribution and the other thresholds (i.e., ) were calculated using Equation 37,

|  |  |  |
| --- | --- | --- |
|  | , for *k* = 4, 3, 2, | (37) |

where  represents a random error term generated from a *N*(0, .04) distribution. All latent trait values were sampled from a *N* (0, 1) distribution.

**Response data generation.**After item parameters were generated, Equation 1 was used to compute the probability that each category of a given item was endorsed by a respondent with a simulated theta value (*θj*). Then, the cumulative probabilities from the first category to the fifth category of an item (i.e., a vector ) were calculated and a number was randomly generated from a uniform distribution *U*(0, 1). The simulated response to an item was determined by the location of the randomly generated in the cumulative probability vector.

**DIF and impact implementation.** The reference group item parameter generation exactly followed the procedures used by Roberts et al. (2002). The focal group item parameter generation was the same as that of the reference group except for the 4 DIF items, which were randomly selected for each replication. For uniform DIF conditions, *δi* values of the selected DIF items were incremented by adding by .50; for nonuniform DIF conditions, *αi* values of the selected DIF items were incremented .50.

When generating the latent trait values in this study, all latent trait values for the reference group were sampled from a *N* (0, 1) distribution. For the 0 impact conditions, the latent trait values for the focal group were also sampled from a *N* (0, 1) distribution; For the conditions of .25 *SD*, .50 *SD*, and 1.00 *SD* impact, the latent trait values for the focal group were sampled from *N* (.25, 1), *N* (.50, 1), and *N* (1.00, 1) distributions.

**Procedure.** The implementation of the Lord’s chi-square involved four steps. First, the response data for the reference group and the focal group were simulated by using the generated item parameters and person parameters. Second, the new software MCMC GGUM was used to separately estimate item parameters for the two response datasets. In this step, two sets of item parameters were estimated: , , , , , and , , , , . Third, a linking method was used to link focal group item parameters , , , ,  to reference group item parameters , , , ,  and obtained , , , , . The linking procedures follow the technical manual of the software program GGUMLINK, Version 1.0 (Roberts, 2002) and the technical details are described below. Finally, Equation 35 was used to calculate a chi-square statistic for each item with five degrees of freedom (because five item parameters were involved in the calculation).

|  |  |  |
| --- | --- | --- |
|  | , | (38) |

where

|  |  |  |
| --- | --- | --- |
|  | , | (39) |
| and | . | (40) |

Here which is the variance-covariance matrix of the difference in item parameter estimates, is the variance-covariance matrix of the item parameter estimates for the reference group, and is the variance-covariance matrix of the calibrated item parameter estimates for the focal group after linking.

The linking method used in this study was the mean-sigma method based on item locations. By this method, two constants  and  were first derived through Equations 38–39,

|  |  |  |
| --- | --- | --- |
|  | , | (41) |
| and | . | (42) |

where and  are the means of the item locations estimates for the common items from the focal and reference groups respectively, and and  are the corresponding standard deviations. Then the item parameters *δ*, *α*, and *τ* were transformed by

|  |  |  |
| --- | --- | --- |
|  | , | (43) |
|  | . | (44) |
| and | , | (45) |

where parameter , , and refer to the estimates after having been transformed from the metric of the focal group to the metric of the reference group, and the subscript *j* refers to the *j*th common item across the scales being calibrated.

## Results

## Summary

**CHAPTER 7**

# STUDY IV: APPLICATION OF THE NEW SOFTWARE TO REAL DATASETS

## Introduction: Testing Goals

The last study of this dissertation research applied the new software to real datasets that require ideal point IRT models and examine its performance. Specifically this study utilized two datasets for a comprehensive personality scale. This comprehensive personality scale consists of 440 items covering a full set of personality facets that were derived from the traditional Big-Five Personality Model. One group of respondents consisted of university undergraduates and the other group was online crowdsourcing participants––specifically from the Amazon Mechanical Turk (MTurk; [www.mturk.com](http://www.mturk.com)). University undergraduates have been the gold standard for psychology research participants for decades; however, with the development of information technology, online crowdsourcing has become popular in the past decade for survey and experiment studies in many social science fields, including I/O psychology and political science (Barger, Behrend, Sharek, & Sinar, 2011; Berinsky, Huber, & Lenz, 2012). Compared to collecting student sample data, collecting data from online crowdsourcing is more efficient and convenient and may cost much less. In addition, online crowdsourcing subjects have been found more representative (e.g., older, more diverse) and have more work experiences than university student subjects, which seems better suited to employee-focused research (Barger, Behrend, Sharek, & Sinar, 2011).

Despite the popularization of online outsourcing for I/O research, it has become increasingly important to understand the equivalence of the two sources of subjects. For example, will similar psychometric properties be observed when assessment tools are administered to the two types of samples? A very few pioneering studies have been conducted. For instance, Behrend, Sharek, Meade, and Wiebe (2011) examined the measurement equivalence of a personality measure and an attitude survey between a university student sample and MTurk sample and they found only a few DIF items. One limitation of previous research lies in its use of an apparently inappropriate item response theory (IRT) model for analyzing DIF. Specifically, they used Samejima's Graded Response Model (SGR, Samejima, 1969), which assumes a dominance response process and might be methodologically inappropriate for the equivalence study.

Thus, this study aimed to examine the measurement equivalence of personality measures across online crowdsourcing and university student samples. It was expected that, by using the appropriate model (i.e., the GGUM model) with an improved estimation method (i.e., the MCMC approach), more valid results would be obtained.

## Methods

**The comprehensive personality scale**. This personality scale was developed in Dr. Fritz Drasgow’s lab through years of work. It is based on the Big-Five model and includes 22 facets underlying the Big Five dimensions. For example, the traditional conscientiousness dimension was extended to six facets: industriousness, order, self-control, traditionalism, responsibility, and virtue; the extraversion dimension was extended to dominance, sociability, excitement, and energy facets; etc. 100 items were written to capture each facet and then 20 items were carefully selected for each facet (440 items in total) to comprise a complete measure of personality. The 20 items in each facet consisted of approximately equal numbers of statements reflecting high, medium and low trait levels.

**Participants**. The 440 items were administered via computers to undergraduates through a university subject pool and to internet users through MTurk. At the end of the personality survey, demographic data were also collected. To ensure response quality, four quality control items were included (e.g., “For quality control purposes, please select ‘Strongly disagree’ as the answer to this item”); they were located in the middle to the end of the survey. 839 undergraduate participants and 673 MTurk participants completed the survey. After deleting cases with at least one error on the quality control items, 733 undergraduate cases and 529 MTurk cases remained. The demographic characteristics of the two sample sources are summarized in Table 3.

## Analysis and Results

The data analyses for DIF detection followed the procedures in Study 3. While using the MCMC GGUM software for the DIF detection, the GGUM2004 was also used as a comparison.

## Summary

**CHAPTER 8**

# GENERAL DISCUSSION AND CONCLUSIONS

## Future Research

# REFERENCES

Ackerman, P. L., & Humphreys, L. G. (1990). Individual differences theory in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (pp. 223–282).

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium on Information Theory.*In B. N. Petrov and F. Csaki (Eds.), 267–281. Budapest:Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19 (6),* 716–723.

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

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: Reconciling Thurstone and Likert methodologies. *British Journal of Mathematical and Statistical Psychology, 49,* 347–365.

Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Meaurement, 17,* 253–276.

Barger, T., Behrend, T.S., Sharek, D. J., & Sinar, E. F. (2011). I-O and the Crowd: Frequently Asked Questions About Using Mechanical Turk for Research. *The Industrial-Organizational Psychologist, 49(2).* 11-18.

Behrend, T. S., Sharek, D. S., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43,* 800-813.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Political Analysis, 20 (3)*, 351-368. doi:10.1093/pan/mpr057.

Bickel, P. J., & Doksum, K. A. (2001). *Mathematical Statistics: Basic Ideas and Selected Topics.* Prentice Hall, NJ: Upper Saddle River.

Binet, A., Simon, T. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Annee Psychologique 11,* 191–244.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In: Lord, F.M., Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores.* Addison-Wesley, Reading, MA, pp. 395–479.

Brooks, S. P. & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7,* 434–455.

Candell, G., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,*253–260.

Carter, N. T., & Dalal, D. K. (2010). An ideal point account of the JDI work satisfaction scale. *Personality and Individual Differences, 49(7),* 743–748.

Carter, N. T. & Zickar, M. (2011). A comparison of the LR and DFIT frameworks of differential functioning applied to the generalized graded unfolding model*. Applied Psychological meaurement, 35,* 623–642.

Cattell, J.M. (1890). Mental tests and measurements. *Mind, 15,* 373–380.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedures.*Journal of Educational Measurement.33,*333–353.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36,* 523–562.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155−159.

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist*, *49*, 997−1003.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335–350.

Cohen, R. J., & Swerdlik, M. (2002). *Psychological testing and assessment: An introduction to tests and measurement*: McGraw-Hill.

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

Davison, M. L. (1977). On a metric, unidimensional unfolding model for attitudinal and developmental data. *Psychometrika, 42,* 523-548.

de la Torre, J. (2003). *Improving the accuracy of item response theory parameter estimates through simultaneous estimation and incorporation of ancillary variables. Unpublished manuscript,* University of Illinois at Urbana-Champaign.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika 69,* 333–393

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,* 216-232.

Donoghue, J., Holland, P., & Thayer, D. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* (pp. 137–166).Hillsdale, NJ: Erlbaum.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22,* 33–51.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010a). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3(4),* 465-476. doi:10.1111/j.1754-9434.2010.01273.x

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010b). Improving the measurement of psychological variables: Ideal point models rock! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 515-520.

Du Bois, P. H. (1970). *A Hisotry of Psychological Testing.* Boston: Allyn & Bacon.

Edelen, M.O., Thissen, D., Teresi, J.A., Kleinman, M., Ocepek-Welikson, K., 2006. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: application to the Mini-Mental State Examination. *Medical Care 44,* S134–S142.

Fabrigar, L. R., MacDonald, T. K., & Wegener, D. T. (2005). The structure of attitudes. In D. Albarracín, B. T. Johnson & M. P. Zanna (Eds.), *The Handbook of Attitudes.* Mahwah, NJ: Lawrence Erlbaum.

Finch, H. ( 2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 29,* 278–295.

Fonseca-Pedrero, E., Wells, C., Paino, M., Lemos-Giraldez, S., Villazon-Garcia, U., Sierra, S., et al. (2010). Meaurement invariance of the Reynolds Depression Adolessent Scale arocss gender and age. *International Journal of Testing, 10,* 133–148.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science, 7,* 457–511.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika, 57,* 97–109.

Hough, L. M., & Oswald, F. L. (2000). Personnel Selection: Looking toward the future—Remembering the past. *Annual Review of Psychology, 2000*, 631–664.

Johnson, M. S. (2006). Nonparametric estimation of item and respondent locations from unfolding-type items. *Psychometrika, 71,* 257–279.

Johnson, M. S., & Junker, B. W. (2003). Using data augmentation and Markov Chain Monte Carlo for the estimation of unfolding response models. *Journal of Educational and Behavioral Statistics, 28,* 195–230.

Kaplan, K. J. (1972). On the Ambivalence-Indifference Problem in Attitude Theory and Measurement: A Suggested Modification of the Semantic Differential Technique. *Psychological Bulletin, 77*, 361-372.

Kim, S-H. (2001). An evaluation of a Markov Chain Monte Carlo method for the Rasch Mode. *Applied Psychological Measurement, 25,* 163–176.

Kirk, R. E. (2006). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference*. DOI: 10.1016/j.jspi.2006.09.011.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 1-55.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parmaters when assesing differential item functioning. *Journal of Applied Psychology, 75,* 164–174.

Lippa, R. A. (2010). Sex differences in personality traits and gender-related occupational preferences across 53 nations: testing evolutionaryand social-environmental theories. *Archive of Sex Behaviors, 39,* 619–636.

Lord, F. M. (1977). A study of item bias using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic Problems in Cross-Cultural Psychology (pp.19–29).* Amsterdam, The Netherlands: Swets & Zeitlinger.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale NJ: Erlbaum.

Lord, F.M., Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, MA.

Masters, G. N. (1982). A Rasch model for partial-credit scoring. *Psychometrika, 47,* 149–174.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif)to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41,* 55–64.

Maydeu-Olivares, A., Hernandez, A. & McDonald, R. P. (2006). A multidimensional ideal point item response theory model for binary data.*Multivariate Behavioral Research,41,*445-471.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95,* 728-743.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087–1092.

McCrae, R. R., Terracciano, A., De Fruyt, F., De Bolle, M., Gelfand, M. J. Costa, P. T. (2010). The validity and structure of culture-level personality scores: data from ratings of young adolescents. *Journal of Personality, 78,* 815–838.

McLaughlin, M. E., & Drasgow, F. (1987). Lord’s chi-square test of item bias with estimated and with known person paraeters. *Applied Psychological Measurement, 11,* 161–173.

Mitchell, T. R., & Daniels, D. (2003). Motivation. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Eds.), *Handbook of industrial and organizational psychology* (pp. 225–254). Hoboken, NJ: John Wiley & Sons.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71–83.

Mullis, I., Dossey, J., Owen, E., & Phillips, G. (1993).*NAEP 1992 mathematics report card for the nation and the states.*Washington, DC: National Center for Education Statistics.

Nye, C. D. (2011). The development and validation of effect size measures for IRT and CFA studies of measurement equivalence (Doctoral dissertation). University of Illinois at Urbana-Champaign, Champaign, IL.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Maxkov chain Monte Carlo methods for item response theory. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R., & Junker, B. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Penfield, R. D. (2010). Distinguishing between net and global DIF in polytomous items*. Journal of Educational Measurement, 47,* 129–149.

Post,W.J. (1992). *Nonparametric unfolding models: a latent structure approach.* M&T Series, Leiden, The Netherlands: DSWO Press.

R Development Core Team (2011). *R: A language and environment for statistical computing. R Foundation for Statistical Computing,* Vienna, Austria. URL <http://www.R-project.org/>.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53,*495–502.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993).An empirical comparison of the area methods, Lords’ chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning.*Educational and Psychological Measurement, 53,* 301–314.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 353-368.

Reynolds, W. M. (2002). *Reynolds Adolescent Depression Scale—2nd Edition. Professional Manual.* Odessa: Psychological Assessment Resources, Inc.

Roberts, J. S. (2001). GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 25,* 38.

Roberts, J. S. (2002). *GGUMLINK version1.0 user’s manual.* Atlanta, GA: Georgia Polytech Uni.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211–233.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24,* 3–32.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002).Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model.*Applied Psychological Measurement, 26,* 192–207.

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2006). GGUM2004: A Windows-based program to estimate parameters of the generalized graded unfolding model. *Applied Psychological Measurement, 30,* 64–65.

Roberts, J. S., & Gordon, J. D. (March, 2008). *Identification of differentially functioning items in the generalized graded unfolding model*. Paper presented at the American Educational Research Association Meeting, New York.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the Likert and Thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211–233.

Roberts, J. S., & Shim, H. S. (2008). *GGUM2004 technical reference manual* (Version 1.1). Atlanta, GA: Georgia Polytechnic University.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement 17* .

Samejima, F. (1997). Graded response model. In: van der Linden, W., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory.* Springer, New York, pp. 85–100.

Scott, W. A. (1969). Structure of national cognitions. *Journal of Personality and Social Psychology, 12,* 261–278.

Spearman, C. (1904). “General intelligence”, objectively determined and measured. *American Journal of Psychology 15,* 201–293.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497–508.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292–1306.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point models be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25–39.

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can subject matter experts’ ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scale? *Organizational Research Methods, 14,* 256–278.

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods, 15,* 363-384.

Tay, L., Ali, U. S., Drasgow, F., & Williams, B. A. (2011). Fitting simulated dichotomous and polytomous data: Examining the difference between ideal point and dominance models. *Applied Psychological Measurement, 35*, 280–295.

Tay, L., & Drasgow, F. (2012). Theoretical, statistical, and substantive issues in the assessment of construct dimensionality: Accounting for the item response process. *Organizational Research Methods, 15*, 363-384.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal? *Journal of Applied Psychology, 94*, 1287–1304.

Terman, L.M. (1916). The Measurement of Intelligence. Houghton Mifflin, Boston.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, 67–113.

Thissen, D. & Wainer, H. (2001). *Test Scoring.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, M. M., Zanna, M. P., & Griffin, D. W. (1995). Let’s not be indifferent about (attitudinal) ambivalence. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 361–386). Hillsdale, NJ: Erlbaum

Thurstone, L.L. (1924). *Psychological Examination for High School Graduates and College Freshmen.* American Council on Education, Washington, DC.

Thurstone, L.L. (1927). *The law of comparative judgment.* Psychological Review 34, 278–286.

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

Thurstone, L. L. (1929). *Theory of attitude measurement*. *Psychological Review, 36*, 222-241.

Thurstone, L.L. (1931a). *Multiple factor analysis.* Psychological Review 38, 406–427.

Thurstone, L.L. (1931b). *The Reliability and Validity of Tests.* Edwards Brothers, Ann Arbor, MI.

Thurstone, L.L. (1935). *Vectors of the Mind.* University of Chicago Press, Chicago.

Wang, W., Tay, L., & Drasgow, F. (in press). Assessing differential item functioning in polytomous items with Generalized Graded Unfolding Model (GGUM). *Applied Psychological Measurement.*

Wang, W.-C., & Su, Y.-H. ( 2004a). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education, 17,* 113–144.

Wang, W.-C., & Su, Y.-H. ( 2004b). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28,* 450–480.

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic,multidimensional unfolding analysis. *Psychometrika, 39,* 327–350.

# TABLES

## Table 1 Estimates by GGUM2004 for a Personality Scale of Industriousness

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Items | Estimates by GGUM2004 | | | |
|  | *SE* |  | *SE* |
| 1. I am competitive and play to win | -6.5314 | 32.6895 | .2697 | .1148 |
| 2. I find it easy to stick to my plans | -1.8232 | .2364 | .7591 | .1079 |
| 3. I am average at the things I do | 8.9254 | 139.1369 | .5286 | .0789 |
| 4. I frequently make-up believable excuses for not finishing my work | -1.7675 | .1672 | .4851 | .1589 |
| 5. I finish my work on time but try not to work more than I have to | 5.6237 | 43.9729 | 1.1153 | .1105 |
| 6. I work hard, but I know when it's time to quit | -.1682 | .1385 | .6160 | .1139 |
| 7. I enjoy the process of doing things and don't care much about the results | 7.7249 | 79.0826 | .6667 | .0873 |
| 8. Being successful is more important than most other things in my life | –– † | –– † | .0089 | .0700 |
| 9. I don't care very much about the quality of my work | -1.7711 | .1564 | 1.8203 | .1912 |
| 10. I hardly ever finish the tasks I start | -1.9164 | .1806 | 1.7881 | .1821 |
| 11. I tend to do just what is expected of me when doing a job | -3.8410 | 3.9383 | .6464 | .0773 |
| 12. I always want to be better than others in the things I do | -6.6873 | 36.5166 | .2713 | .1379 |
| 13. There is too much to be done to waste time relaxing | -26.3010 | 729.5049 | .0530 | .1387 |
| 14. When I set my mind on achieving a goal, I can always reach it | -3.0543 | .8457 | .8181 | .1006 |
| 15. I always try to do my best work even when no one will know | -3.3816 | 6.2144 | 1.3149 | .1747 |
| 16. If I am interested in something I don't mind working hard | -3.1412 | 4.3088 | .8750 | .1474 |
| 17. To me, being moderately successful is enough | 10.4909 | 74.1443 | .3878 | .0746 |
| 18. I don't really care about being successful | -3.9027 | 1.5415 | .4071 | .0668 |
| 19. People should not sacrifice too much for work | -7.0539 | 29.6252 | .2289 | .0948 |
| 20. I try to do the minimal amount of work possible to maintain my current status | -3.1852 | 16.6491 | 2.2713 | .2419 |

† An infinite value was estimated by GGUM2004 for that parameter.

## Table 2 Type I Error Rates and Power for DIF Detection by GGUM2004 When Impact = .50 *SD*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Scale length | DIF% | *N* | Three DIF detection methods | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Log-likelihood ratio | | | | | | | | | | | |  | AIC | | | | | | | | | | | |  | Lord’s  Chi-square | |
| Type I | | | | | | Power | | | | | | Type I | | | | | | Power | | | | | | Type I | Power |
| 1a | | 2a | | 4a | | 1a | | 2a | | 4a | | 1a | | 2a | | 4a | | 1a | | 2a | | 4a | | – | – |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 10 | 20%  (*n* = 2) | 250 | .08 | (.30) | .10 | (.18) | .07 | (.10) | .95 | (.99) | .95 | (.97) | .94 | (.95) |  | .09 | (.35) | .11 | (.20) | .09 | (.11) | .95 | (.99) | .95 | (.97) | .94 | (.96) |  | .06 | .24 |
| 500 | .12 | (.65) | .11 | (.27) | .10 | (.15) | .99 | (1.00) | 1.00 | (1.00) | 1.00 | (.99) |  | .14 | (.68) | .13 | (.30) | .12 | (.17) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (.99) |  | .10 | .40 |
| 1,000 | .24 | (.88) | .23 | (.59) | .21 | (.29) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (1.00) |  | .26 | (.89) | .25 | (.62) | .24 | (.31) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (1.00) |  | .15 | .49 |
| 40%  (*n* = 4) | 250 | .33 | (.37) | .26 | (.16) | .29 | (.12) | .84 | (1.00) | .82 | (.98) | .85 | (.96) |  | .35 | (.41) | .29 | (.16) | .31 | (.15) | .85 | (1.00) | .83 | (.98) | .85 | (.96) |  | .14 | .22 |
| 500 | .50 | (.63) | .47 | (.32) | .47 | (.11) | .96 | (1.00) | .95 | (1.00) | .95 | (1.00) |  | .51 | (.66) | .49 | (.36) | .48 | (.13) | .97 | (1.00) | .96 | (1.00 | .96 | (1.00) |  | .27 | .35 |
| 1,000 | .67 | (.84) | .71 | (.54) | .72 | (.27) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (1.00) |  | .70 | (.86) | .73 | (.57) | .74 | (.30) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (1.00) |  | .42 | .46 |
| 20 | 20%  (*n* = 4) | 250 | .11 | (.35) | .10 | (.13) | .11 | (.11) | .94 | (1.00) | .92 | (.90) | .91 | (.93) |  | .12 | (.39) | .12 | (.15) | .13 | (.13) | .94 | (1.00) | .92 | (.90) | .92 | (.93) |  | .06 | .35 |
| 500 | .16 | (.57) | .18 | (.28) | .18 | (.17) | .98 | (1.00) | .98 | (1.00) | 1.00 | (1.00) |  | .18 | (.61) | .20 | (.30) | .21 | (.18) | .98 | (1.00) | .98 | (1.00) | 1.00 | (1.00) |  | .11 | .48 |
| 1,000 | .30 | (.85) | .30 | (.53) | .30 | (.24) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (1.00) |  | .32 | (.87) | .32 | (.57) | .33 | (.25) | 1.00 | (1.00) | 1.00 | (1.00) | 1.00 | (1.00) |  | .16 | .69 |
| 40%  (*n* = 8) | 250 | .37 | (.36) | .34 | (.20) | .37 | (.12) | .79 | (.99) | .77 | (.96) | .81 | (.96) |  | .39 | (.39) | .37 | (.22) | .39 | (.14) | .81 | (.99) | .78 | (.96) | .83 | (.97) |  | .18 | .23 |
| 500 | .63 | (.65) | .61 | (.26) | .63 | (.16) | .95 | (1.00) | .93 | (1.00) | .95 | (.98) |  | .65 | (.68) | .63 | (.29) | .66 | (.18) | .96 | (1.00) | .94 | (1.00) | .95 | (.98) |  | .30 | .41 |
| 1,000 | .84 | (.83) | .86 | (.54) | .85 | (.27) | 1.00 | (1.00) | .99 | (1.00) | 1.00 | (1.00) |  | .86 | (.85) | .87 | (.57) | .86 | (.30) | 1.00 | (1.00) | .99 | (1.00) | 1.00 | (1.00) |  | .46 | .53 |

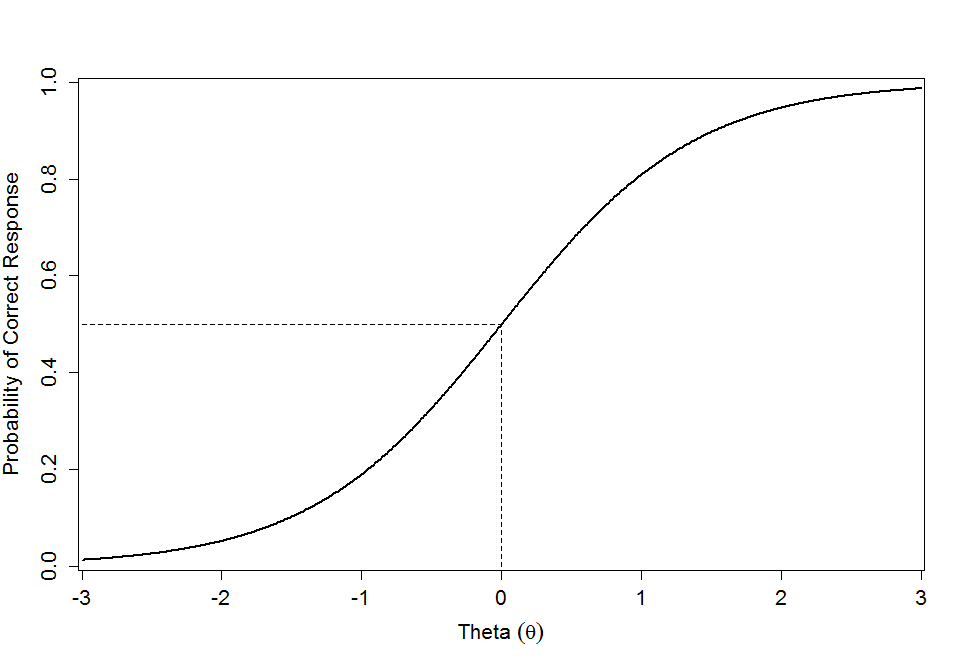
*Note.* a indicate the number of items linking the focal and reference groups. Results were from the condition of uniform DIF type with an impact of .50 *SD*.

## Table 3 Sample characteristics of the two subject groups

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Undergraduate group | |  | MTurk group | |
|  | N | % |  | N | % |
| Age |  |  |  |  |  |
| Under 18 | 0 | 0.00 |  | 1 | 0.20 |
| 18-25 | 727 | 99.18 |  | 166 | 31.38 |
| 26-35 | 6 | 0.82 |  | 184 | 34.78 |
| 36-45 | 0 | 0.00 |  | 92 | 17.39 |
| 46-55 | 0 | 0.00 |  | 54 | 10.21 |
| Above 55 | 0 | 0.00 |  | 32 | 6.05 |
| No response | 0 | 0.00 |  | 0 |  |
|  |  |  |  |  |  |
| Gender |  |  |  |  |  |
| Male | 291 | 39.70 |  | 210 | 39.70 |
| Female | 440 | 60.00 |  | 317 | 59.90 |
| No response | 2 | 0.30 |  | 2 | 0.40 |
|  |  |  |  |  |  |
| Ethnicity |  |  |  |  |  |
| White | 349 | 47.70 |  | 430 | 81.30 |
| African | 38 | 5.20 |  | 32 | 6.00 |
| Hispanic | 69 | 9.40 |  | 25 | 4.70 |
| Asian | 250 | 34.10 |  | 31 | 5.90 |
| Other | 23 | 3.10 |  | 10 | 1.90 |
| No response | 4 | 0.50 |  | 1 | 0.20 |
|  |  |  |  |  |  |
| Education |  |  |  |  |  |
| High school diploma or lower | 0 | 0.00 |  | 206 | 38.90 |
| College | 732 | 99.90 |  | 189 | 35.70 |
| Master’s degree or higher | 0 | 0.00 |  | 80 | 15.10 |
| Professional degree | 0 | 0.00 |  | 51 | 9.60 |
| No response | 2 | 0.10 |  | 3 | 0.60 |
|  |  |  |  |  |  |
| Employment |  |  |  |  |  |
| Never | 133 | 18.10 |  | 16 | 3.00 |
| Part-time or full-time before, but not now | 411 | 56.10 |  | 181 | 34.20 |
| Part-time | 185 | 25.30 |  | 117 | 22.10 |
| Full-time | 1 | 0.10 |  | 213 | 40.30 |
| No response | 3 | .40 |  | 2 | 0.40 |

# FIGURES

## Figure 1. A Typical Item Response Function (IRF) for a Dominance Model



## Figure 2. The Pacifism-Militarism Continuum and Six Item Locations on the Continuum

**(Reprinted from Thurstone (1928, p. 537)**



## Figure 3. Item Endorsement Frequency in Attitude Measurement

**(Reprinted from Thurstone (1928, p. 550)**



## Figure 4. Response Probability from Thurstone (1929)

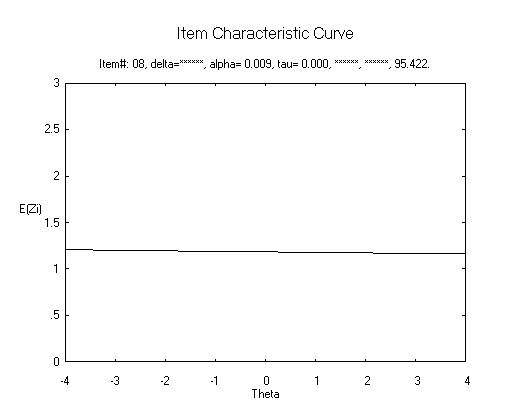
**(Reprinted from Thurstone (1929, p. 229)**



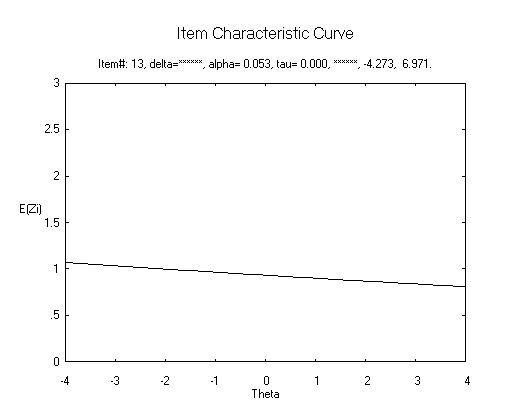
## Figure 5. A Typical Item Response Function (IRF) for an Ideal Point Model with Neutral Location



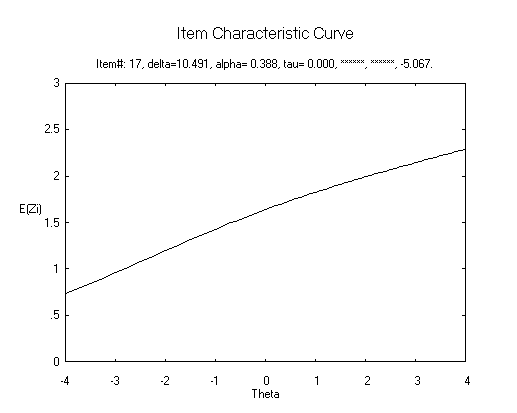
## Figure 6. The Item Curve of Expected Score for Item #8 in Table 1



## Figure 7. The Item Curve of Expected Score for Item #13 in Table 1



## Figure 8. The Item Curve of Expected Score for Item #17 in Table 1



## Figure 9. A Comparison of the Normal Distribution and a Four-Parameter Beta Distribution *Beta*(2, 2, –5, 5)



Note: the thin black line is for the normal distribution and the fat grey line is for the beta distribution *Beta*(2, 2, –5, 5)