6-7-2016

Investigating Parameter Recovery and Item Information for Triplet Multidimensional Forced Choice Measure: An Application of the GGUM-RANK Model

Philseok Lee
University of South Florida, philseok@mail.usf.edu

Scholar Commons Citation

Lee, Philseok, "Investigating Parameter Recovery and Item Information for Triplet Multidimensional Forced Choice Measure: An Application of the GGUM-RANK Model" (2016). *Graduate Theses and Dissertations.*
http://scholarcommons.usf.edu/etd/6298

Investigating Parameter Recovery and Item Information for Triplet Multidimensional Forced Choice Measure: An Application of the GGUM-RANK Model

by

Philseok Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Stephen Stark, Ph.D.
Michael Coovert, Ph.D.
Walter Borman, Ph.D.
Joseph Vandello, Ph.D.
Robert Dedrick, Ph.D.
Oleksandr S. Chernyshenko, Ph.D.

Date of Approval:
April 26, 2016

Keywords: Multidimensional Forced Choice Format, Item Response Theory (IRT),
Monte Carlo Simulation, Parameter Recovery, Item Information

DEDICATION

I would like to express the deepest appreciation to my advisor, Dr. Stephen Stark, who has continually and convincingly conveyed a spirit of adventure in regard to research. Without his guidance, patience, and persistent help, this dissertation would not have been possible. I would also like to thank my committee members, Dr. Michael Coovert, Dr. Walter Borman, Dr. Joseph Vandello, Dr. Robert Dedrick, and Dr. Oleksandr Chernyshenko, for their suggestions and advice throughout the dissertation process. I express my deepest gratitude to my mother. Without her sacrifice and prayer, I would never be where I am today. I would like to thank my wife, Joo, who has supported me in every way she could. Without her love and support, this dissertation would not have been completed. I also would like to thank all my family members. Most of all, I dedicate this dissertation to the almighty GOD who gave me the strength and patience throughout the entire project. Without him, my life is nothing.

On my bed I remember you; I think of you through the watches of the night. Because you are my help, I sing in the shadow of your wings. My soul clings to you; your right hand upholds me. Psalm 63:6-8

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

To control various response biases and rater errors in noncognitive assessment, multidimensional forced choice (MFC) measures have been proposed as an alternative to single-statement Likert-type scales. Historically, MFC measures have been criticized because conventional scoring methods can lead to ipsativity problems that render scores unsuitable for inter-individual comparisons. However, with the recent advent of classical test theory and item response theory scoring methods that yield normative information, MFC measures are surging in popularity and becoming important components of personnel and educational assessment systems**.** This dissertation presents developments concerning a GGUM-based MFC model henceforth referred to as the GGUM-RANK. Markov Chain Monte Carlo (MCMC) algorithms were developed to estimate GGUM-RANK statement and person parameters directly from MFC rank responses, and the efficacy of the new estimation algorithm was examined through computer simulations and an empirical construct validity investigation. Recently derived GGUM-RANK item information functions and information indices were also used to evaluate overall item and test quality for the empirical study and to give insights into differences in scoring accuracy between two-alternative (pairwise preference) and three-alternative (triplet) MFC measures for future work. This presentation concludes with a discussion of the research findings and potential applications in workforce and educational setting

CHAPTER ONE:

INTRODUCTION

Over the last two decades, there has been increasing interest in noncognitive constructs in educational and occupational settings. This increase stems in part from legal and societal concerns about adverse impact associated with cognitive ability testing (Hough & Oswald, 2008; Sinha, Oswald, Imus, & Schmitt, 2011) and evidence that noncognitive constructs provide incremental validity for many outcomes (Lievens, Buyse, & Sackett, 2005; Ployhart & Holtz, 2008; Schmidt & Hunter, 1998). In industrial-organizational (I-O) psychology, for example, personality has been shown to predict citizenship and counterproductive work performance (Barrick, Mount, & Judge, 2001; Berry, Ones, & Sackett, 2007; Borman et al., 2001b; Hurtz & Donova, 2000), leadership (Bono & Judge, 2004), career success (Judge & Hurst, 2007), and creativity (Bartram, 2005). In addition, other noncognitive constructs, such as emotional intelligence (Van Rooy & Viswesvaran, 2004), vocational interests (Morris, 2003), values (Schwartz, 2012), and social skills (Semadar, Robins, & Ferris, 2006) have been linked to important job criteria.

Historically, noncognitive constructs have been measured predominantly using Likert-type scales, which require respondents to indicate their level of agreement with a set of statements using, for example, a 1 (Strongly Disagree) to 5 (Strongly Agree) format. However, this methodology has been criticized due to its susceptibility to various types of response biases. In particular, socially desirable responding, which is also known as impression management or faking good, tends to inflate scale means and intercorrelations, and it can reduce the validity and utility of measures used for high-stakes decision making (e.g., Griffith, Chmielowski, & Yoshita, 2007; Hough, Eaton, Dunnette, Camp, & McCloy, 1990; Schmitt & Oswald, 2006; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001; White & Young, 1998). Likert-type scales are also susceptible to rater errors (e.g., leniency, halo) and cultural-specific response biases (e.g., central tendency, extremity or acquiescence), which may inflate cross-dimension correlations (Baron, 1996; Borman et al., 2001a; Brown & Maydeu-Olivares, 2014; Meade, 2004; Stark & Drasgow, 2002) or attenuate relationships with outcomes in cross-cultural research contexts (e.g., Ferrando, Anguiano-Carrasco, & Chico, 2011; He & van de Vijver, 2013).

To control response biases and rater errors, multidimensional forced choice (MFC) measures have been proposed as an alternative to Likert scales for noncognitive assessment. MFC measures commonly present statements in blocks of two (pair), three (triplet), or four (tetrad). Within the blocks, statements representing different constructs may be matched on social desirability and/or extremity. The respondent's task is to choose the statement in each block that is "most like me", or to rank the statements in each block from "most like me" to "least like me". In theory, matching on social desirability and/or extremity makes the "best" answers difficult to discern, and by forcing respondents to choose between alternatives, rather than indicating their level of agreement with each statement, response biases and rater errors may be reduced (see Bowen, Martin, & Hunt, 2002; Cheung & Chan, 2002; Christiansen, Burns, & Montgomery, 2005; Ferrando et al., 2011; He, Bartram, Inceoglu, & van de Vijver, 2014; Jackson, Wroblewski, & Ashton, 2000; Vasilopoulos, Cucina, Dyomina, Morewitz, & Reilly, 2006).

Like Likert scales, MFC measures have been criticized, but primarily because conventional MFC scoring methods can lead to ipsativity problems that render scores unsuitable for inter-individual comparisons (i.e., ipsativity problems; see Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988; Meade, 2004). However, with the advent of classical test theory (CTT; White & Young, 1998) and item response theory (IRT) scoring methods (e.g., Brown & Maydeu-Olivares, 2011; de la Torre et al., 2012; Stark, 2002; Stark, Chernyshenko, & Drasgow, 2005) that yield normative information and studies showing MFC scores based on these methods predict important criteria (e.g., Brown & Bartram, 2009a; Salgado & Táuriz, 2014), MFC measures are surging in popularity and becoming important components of personnel and educational assessment systems.

At present, research is needed to answer questions concerning the optimal configuration of MFC tests for reducing response biases and rater errors. There is also interest in computerized adaptive testing to increase efficiency, parallel test forms to prevent overexposure of items and provide backup in the event of a test compromise, and measurement invariance methods for fairness evaluations and cross-national comparisons. For these purposes, IRT methods are more desirable than CTT methods because they provide test constructors with more statistical information and can support a wider range of testing applications.

Within the IRT framework, only a few MFC psychometric models have been proposed. Stark et al. (2005) proposed the Multi-Unidimensional Pairwise Preference (MUPP; Stark, 2002) model for constructing and scoring multidimensional pairwise preference items. They estimated parameters for individual personality statements using the Generalized Graded Unfolding Model (GGUM, Roberts, Donoghue, & Laughlin, 2000) and used those statement parameters, in conjunction with social desirability ratings, to construct and score Multidimensional Pairwise Preference (MDPP) test forms for assessment purposes. (This is referred to as a two-step process, because a statement pool is calibrated in step 1, and MFC tests are built and administered to examinees for assessment in step two).

As an alternative, Brown (2010) and Brown and Maydeu-Olivares (2011) proposed the Thurstonian IRT (TIRT) MFC model, which applies not only to “more like me” judgments associated with pairwise preference items, but also to rank-order judgments involving blocks of three or more statements. This approach transforms rank-order judgments among a set of stimuli into a set of binary judgments that are scaled and scored using a multidimensional normal ogive IRT model via the Mplus (Muthén & Muthén 1998–2015) statistical program. (Because item parameter estimation and scoring can be accomplished with a single data collection, this is referred to as a one-step, or *direct*, estimation process (Seybert, 2013)).

Finally, de la Torre and colleagues (de la Torre et al., 2012; Hontangas et al., 2015) generalized Stark’s (2002) GGUM-based MUPP model to apply to MFC rank responses. This approach views rank-order responses as a sequence of independent “most like” judgments among a set of diminishing alternatives (e.g., most like me among four statements, then among three, then among two). Statement and person parameters are estimated via a combination of Bayesian methods (Markov chain Monte Carlo [MCMC] for statement parameters and MCMC or expected a posteriori [EAP] estimation for person parameters). Seybert (2013) used this approach as a template for developing an alternative MFC rank model based on Andrich and Luo’s (1993) hyperbolic cosine model (HCM), because the HCM produces IRFs similar to the GGUM with a somewhat simpler mathematical form. Henceforth, to differentiate these multi-unidimensional MFC rank models, de la Torre et al.’s (2012) RANK model will be referred to as the GGUM-RANK, which was further developed and tested in this dissertation.

Limitations and Opportunities for Research

The models described above approach MFC testing differently and are not without limitations. Stark (2002) proposed a two-step strategy for MUPP parameter estimation due to the complexities of “direct” marginal maximum likelihood estimation of MFC statement parameters. Although time consuming and perhaps costly to pretest and calibrate statements before an MFC test administration, this approach is advantageous for computerized adaptive testing (CAT), because any number of different tests can be built and scored once a statement pool has been calibrated. In contrast, direct MFC estimation obviates statement pretesting and may provide more accurate statement parameter estimates by taking into account potential interactions among the statements within a block (see Brown, 2010). If such interactions occur, however, statements cannot be shuffled to create new MFC items or test forms without re-estimation. This has interesting and important implications for measurement invariance or, conversely, differential item functioning; if statement parameters depend on context, then it is impossible to classify individual statements as good or bad and designate them for revision or exclusion from future tests; instead, each potential block of statements would need to be examined. Finally, neither de la Torre and colleagues (2012) nor Seybert (2013) presented information functions for MFC blocks involving triplets or tetrads of statements. Information functions are helpful in creating effective MFC items and developing measures that meet reliability goals.

The Present Investigation

This dissertation had three aims. First, it aimed to develop a MCMC algorithm for estimating *statement and person parameters* directly from MFC triplet (three-alternative) rank responses, based on the GGUM-RANK model. In contrast to previous Monte Carlo simulations that focused exclusively on scoring (e.g., Hontangas et al., 2015) using idealized distributions of statement parameters for data generation, MFC measures were constructed using statement parameters that were systematically varied across conditions likely to be encountered in practice. Second, newly derived methods for computing GGUM-RANK item and test information (Joo, Lee, & Stark, 2016) were used to examine how manipulating statement parameters influenced estimation error and test reliability, with the larger goal of developing empirically-based guidelines for future MFC test construction. The third aim of this dissertation was to compare GGUM-RANK parameter estimation and test information across two- and three- alternative MFC formats. This phase of research addressed the potential psychometric benefits of using formats more complex than pairs in response to interest in MFC triplet measures expressed in recent articles (e.g., Anguiano-Carrasco, MacCann, Greiger, Seybert, & Roberts, 2014; Brown & Bartram, 2009b; Brown & Maydeu-Olivares, 2011; Bartram & Burke, 2013). The final goal of this dissertation was to demonstrate the viability of the developed GGUM-RANK MCMC method using empirical data. An overarching goal was to provide practitioners with tools and guidelines for constructing effective MFC measures in applied settings.

This dissertation is divided into seven sections. Section 1 discusses MFC triplet measures and an illustrative classical test theory scoring method. Section 2 reviews the MUPP (Stark et al., 2005) and GGUM-RANK (de la Torre et al., 2012) models. Section 3 focuses specifically on the GGUM-RANK model for MFC triplets. Section 4 presents newly derived GGUM-RANK information functions and describes how test information was calculated in the proposed simulations. Section 5 presents a new MCMC algorithm for estimating GGUM-RANK parameters. Section 6 describes two simulation studies and an empirical study that were conducted. Study 1 examined GGUM-RANK statement and person parameter recovery and item and test information as a function of sample size, test length, intrablock discrimination, and intrablock location parameter variability. Study 2 compared GGUM-RANK parameter recovery and test information for MFC pair and triplet measures for different test lengths and sample sizes. The results informed MFC test construction practices. Study 3 examined convergent and predictive validity evidence for MFC and Likert-type personality measures, which were administered to a large sample of online research participants, and scored using the GGUM-RANK MCMC and classical test theory methods, respectively. Finally, Section 7 of this presentation summarizes key findings, limitations, and implications for future research and practice.

Classical Test Theory Scoring of MFC Responses

MFC measures require respondents to rank or choose statements from multiple alternatives representing different psychological traits within a block. As described by Hontangas et al. (2015), MFC response formats can be categorized into three types: PICK (choose the statement that is *most like* you), MOLE (choose the *most like* and *least like* statements), and RANK (*rank* the statements from most to least like you). Hontangas et al. (2015) found that the RANK response format yielded better latent trait (person parameter) recovery than the MOLE and PICK response formats. This dissertation focuses on RANK responses. An example MFC triplet item for rank responses is shown below.

For each block of statements that follow, rank the statements from most like you (1) to least like you (3).

	Rank Order	Classical Score
(A) I always turn in my assignments on time. (+C)	3	1
(B) I generally perform well under pressure. (+Em)	1	2
(C) I enjoy learning about other cultures. (+O)	2	3

Note: A MFC triplet item for rank responses involving positively (+) keyed statements representing Conscientiousness (C), Emotional Stability (Em), and Openness to Experience (O).

With MFC measures, simple classical scoring methods can produce *ipsative* data that exhibit negative scale-intercorrelations and distorted reliability and validity estimates (e.g., Baron, 1996; Clemans, 1966; Hicks, 1970; Meade, 2004). Ipsative data support only intra-individual comparisons (Hicks, 1970). If, for example, one simply assigns points corresponding to the inverted ranks of statements within MFC blocks, the points for each block would sum to a constant (6 in the example above), and the sum of the scale scores would be the same for every examinee, making inter-individual comparisons problematic (Hicks, 1970; Meade, 2004; Stark, 2002; Stark et al. 2005). However, by taking steps to introduce variation in scale scores (e.g., by including distractor statements that are not scored, or by varying the number of statements representing each dimension in an MFC measure), it is possible to produce *partially ipsative* scores (Hicks, 1970; White & Young, 1998; Stark, 2002; Stark et al., 2005) that can predict important organizational outcomes (Salgado & Táuriz, 2014). Importantly, as discussed by Brown and Maydeu-Olivares (2014), ipsativity is not an inherent property of MFC measures. Classical test theory (White & Young, 1998; McCloy, Heggestad, & Reeve, 2005) and item response theory MFC methods (e.g., Brown & Maydeu-Olivares, 2011; Stark, 2002; Stark et al., 2005; de la Torre et al., 2012) can yield normative scores that are useful for applications, such as personnel screening (Stark et al., 2014; White & Young, 1998).

Item Response Theory Models for MFC Responses

The Multi-Unidimensional Pairwise Preference (MUPP) Model.

Stark (2002) proposed the MUPP model for MFC test construction and scoring. The model assumes that when a respondent is presented with a pair of statements (*j* and *k*) and is asked to choose the statement that is more descriptive of him/her, the respondent considers each statement separately. A preferential decision is equivalent to agreeing with one statement and disagreeing with the other. If the respondent agrees with both statements, then he/she must reevaluate them independently until a preference is reached. This preference is represented mathematically as a joint probability, which depends on the respondent's trait scores and the parameters associated with the statements based on a unidimensional IRT model. The probability of preferring statement *j* over statement *k* is given by,

$$P_{(j>k)_i}\left(\theta_{d_j,}\theta_{d_k}\right) = \frac{P_{jk}\{1,0\}}{P_{jk}\{1,0\}+P_{jk}\{0,1\}} = \frac{P_j\{1\}P_k\{0\}}{P_j\{1\}P_k\{0\}+P_j\{0\}P_k\{1\}} \quad (1)$$

where,

"$>$" means "preferred,"

i = index for items, *i* = 1, 2, …, *I*,

j, k = indices for first and second statements in item (MFC block) *i*,

d = index for dimensions (constructs) represented by the statements, *d* = 1, 2, …., D,

$\theta_{d_j,}\theta_{d_k}$= latent trait scores for a respondent on dimensions d_j, d_k, respectively,

$P_{jk}\{1,0\}$ = joint probability of endorsing statement *j* and not endorsing statement *k* at

$(\theta_{d_j,}\theta_{d_k})$,

$P_{jk}\{0,1\}$ = joint probability of not endorsing statement *j* and endorsing statement *k* at

$(\theta_{d_j,}\theta_{d_k})$,

$P_j\{1\}, P_j\{0\}$ = probability of endorsing / not endorsing statement *j* at $\theta_{d_{j,}}$

$P_k\{1\}, P_k\{0\}$ = probability of endorsing / not endorsing statement *k* at θ_{d_k}, and

$P_{(j>k)}\left(\theta_{d_{j,}}\theta_{d_k}\right)$ = probability of a respondent preferring statement *j* to statement *k* in item (block) *i*, given his or her scores on the respective dimensions.

Based on model-data fit investigations showing that ideal point models and, particularly, the GGUM fit ordered-categorical personality data well (Chernyshenko et al., 2001; Stark et al., 2006), Stark (2002) suggested using the dichotomous version of the GGUM (Roberts et. al., 2000) for computing MUPP statement endorsement probabilities ($P_j\{1\},\ P_j\{0\},\ P_k\{1\},\ P_k\{0\}$ in Equation 1), which are henceforth referred to as *component probabilities*. He proposed a two-step process for MFC testing: 1) Individual personality statements representing different dimensions are administered to a large sample of examinees (N>400) using a four-point ordinal response format. The response data are dichotomized and calibrated for each dimension separately using a program that provides GGUM statement parameters (e.g., GGUM2000; Roberts, 2000). 2) Multidimensional pairwise preference measures are then created by forming MFC items using MUPP information functions and separately obtained social desirability ratings. The MFC measure is administered for assessment purposes and the response data are scored using a multidimensional Bayes modal estimation algorithm. Research since 2002 has shown that this algorithm can adequately recover latent trait scores with multidimensional pairwise preference tests involving as many as 25 dimensions (Stark et al., 2005; 2012). (For reference, the next section provides a short description of the dichotomous GGUM.)

Generalized Graded Unfolding Model (GGUM)

The GGUM is a unidimensional ideal point model that can be applied to dichotomous and polytomous responses. Stark (2002) provided the simplified version for dichotomous data shown below:

$$P(0) = P(Z = 0|\theta) = \frac{1+\exp(\alpha[3(\theta-\delta)])}{\gamma}, \text{ and} \qquad (2a)$$

$$P(1) = (Z = 1|\theta) = \frac{\exp(\alpha[(\theta-\delta)-\tau])+\exp(\alpha[2(\theta-\delta)-\tau])}{\gamma}, \qquad (2b)$$

where,

α = the discrimination parameter for a particular statement,

δ = the location of the statement on the latent trait continuum,

τ = the location of the subjective response category threshold on the latent trait continuum, and

γ= $1 + \exp(\alpha[3(\theta - \delta)]) + \exp(\alpha[(\theta - \delta) - \tau]) + \exp(\alpha[2(\theta - \delta) - \tau])$ is a normalizing factor equal to the sum of the numerators in Equations (2a) and (2b).

The GGUM assumes that when respondents evaluate statements to make endorsement (agree/disagree) decisions, they consider the distance between their location and the location of the statements on the trait continuum (i.e., $|\theta - \delta|$). As $|\theta - \delta|$ increases, the probability of agreement decreases, leading to bell-shaped item response function (IRFs) that peak at $\theta = \delta$. In other words, respondents are most likely to agree with statements that express attitudes, feelings, beliefs, or actions similar to their own, and they tend to disagree as perceived dissimilarity grows.

Figure 1 displays a GGUM IRF for a hypothetical statement having parameters $\alpha = 1.50$, $\delta = 0.00$, and $\tau = -1.00$, respectively. The vertical axis represents the probability of agreeing with the statement, and the horizontal axis indicates the latent trait scores. The figure shows that the probability of agreement peaks at $|\theta - \delta| = 0$ and decreases in both directions, yielding a nonmonotonic symmetric function. The location of the peak on the latent trait continuum is determined by the location (a.k.a., extremity) parameter, δ. The steepness of the IRF is determined by the discrimination parameter (α) and the subjective response category threshold, τ (Roberts et al., 2000). For details concerning the use of the GGUM in connection with the MUPP model, see Stark (2002) and Stark et al. (2005).

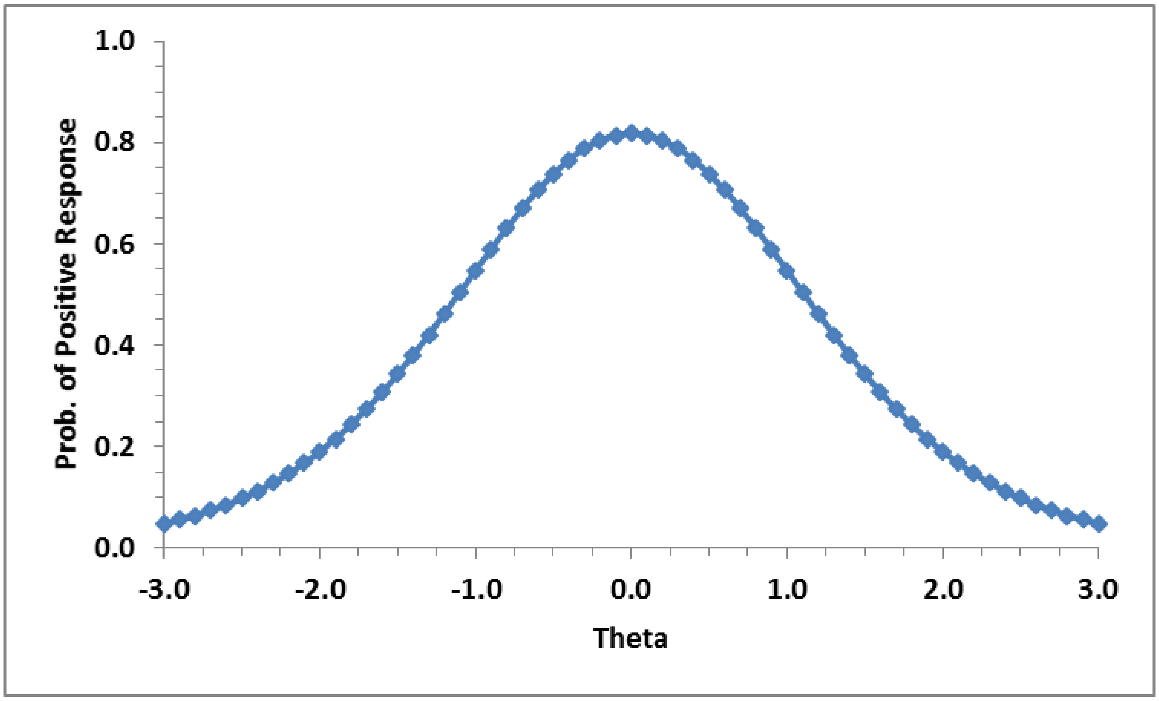


Figure 1. Item response function (IRF) for the dichotomous GGUM.

The PICK, RANK, and MOLE Models.

de la Torre et al. (2012) extended the MUPP model to more complex MFC formats. The PICK model is a generalized version of Stark's (2002) MUPP model for MFC items (blocks) involving 2 to *M* statements per block. For example, if a respondent is presented with a block of four statements labeled A, B, C, and D (a tetrad), and is instructed to select the statement that is "most like you," the model assumes that the respondent evaluates the statements independently until he/she agrees with just one. If a respondent chooses statement A, the probability of the decision would be given by:

$$P_{(\mathrm{A}>\mathrm{B,C,D})} = \frac{P\{1,0,0,0\}}{P\{1,0,0,0\}+P\{0,1,0,0\}+P\{0,0,1,0\}+P\{0,0,0,1\}}\,, \quad (3)$$

where *P{1,0,0,0}* represents the joint probability of agreeing with statement A and disagreeing with statements B, C, and D. If the respondent were to choose statement B, the numerator would become *P{0,1,0,0}*, and similar logic would apply for choosing statements C or D as "most like." Note that the denominator is the same in each case – representing the sum over all possible outcomes. As with the MUPP, the independence assumption allows the joint probability terms in the numerator and denominator to be separated into their component probabilities and computed using a unidimensional model for dichotomous responses, such as the GGUM (Roberts et al., 2000).

Next, following Luce (1959), who proposed that the probability of a set of ranks can be viewed as sequence of independent "most like" (PICK) decisions among a set of diminishing alternatives (M, M-1, …, 2), de la Torre et al. (2012) developed the RANK model for MFC rank responses. For the tetrad example above, the probability of the hypothetical ranking A>D>B>C would be given by the following sequence of PICK decisions:

$$P_{(\mathrm{A}>\mathrm{D}>\mathrm{B}>\mathrm{C})} = P_{(\mathrm{A}>\mathrm{B,C,D})} * P_{(\mathrm{D}>\mathrm{B,C})} * P_{(\mathrm{B}>\mathrm{C})}\,. \quad (4)$$

Finally, for MFC formats involving partial ranks, corresponding to instructions such as, "in each block, choose one statement that is 'most like you' and one statement in each block that is 'least like you' (e.g., White & Young, 1998)", the probability of choosing statement A as 'most like' and statement C as 'least like' would be given by adding the RANK probabilities,

$$P_{(\mathrm{A}**\mathrm{C})} = P_{(\mathrm{A}>\mathrm{B}>\mathrm{D}>\mathrm{C})} + P_{(\mathrm{A}>\mathrm{D}>\mathrm{B}>\mathrm{C})}\,, \quad (5)$$

which takes into account the unknown preference regarding statements B and D.

The GGUM-RANK Model for MFC Triplet Measures

This section provides a detailed description of a GGUM-based RANK model for MFC triplets, referred to henceforth as the GGUM-RANK, for which parameter estimation algorithms were developed and tested in this research. For a block of three statements labeled (A, B, C), the respective PICK probabilities are:

$$P_{(\mathrm{A}>\mathrm{B,C})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = \frac{P_\mathrm{A}(1)P_\mathrm{B}(0)P_\mathrm{C}(0)}{P_\mathrm{A}(1)P_\mathrm{B}(0)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(1)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(0)P_\mathrm{C}(1)} \tag{6a}$$

$$P_{(\mathrm{B}>\mathrm{A,C})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = \frac{P_\mathrm{A}(0)P_\mathrm{B}(1)P_\mathrm{C}(0)}{P_\mathrm{A}(1)P_\mathrm{B}(0)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(1)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(0)P_\mathrm{C}(1)} \tag{6b}$$

$$P_{(\mathrm{C}>\mathrm{A,B})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = \frac{P_\mathrm{A}(0)P_\mathrm{B}(0)P_\mathrm{C}(1)}{P_\mathrm{A}(1)P_\mathrm{B}(0)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(1)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(0)P_\mathrm{C}(1)} \tag{6c}$$

where,

“>” means “preferred,”

i = the index for each item (block of three statements; triplet), *i* = 1 to *I*,

A, B, C = the labels for the statements in block *i*,

d = the index for dimensions represented by the statements, *d* = 1, … , *D*,

$\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}$ = the respondent’s latent trait scores on the dimensions represented by the statements A, B, and C in block *i*,

$P_\mathrm{A}(1), P_\mathrm{B}(1), P_\mathrm{C}(1)$ = the probabilities of endorsing statements A, B, and C,

$P_\mathrm{A}(0), P_\mathrm{B}(0), P_\mathrm{C}(0)$ = the probabilities of not endorsing statements A, B, and C,

$P_{(\mathrm{A}>\mathrm{B,C})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent preferring statement A over statements B and C in block *i*,

$P_{(\mathrm{B}>\mathrm{A,C})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent preferring statement B over statements A and C in block *i*, and

$P_{(\mathrm{C}>\mathrm{A,B})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent preferring statement C over statements A and B in block *i*.

With MFC blocks involving three or more dimensions, PICK probabilities cannot be displayed using a single three-dimensional surface. However, for the special case in which all statements represent the same underlying dimension, the PICK probabilities can be displayed for all statements simultaneously using plots like Figure 2. In this case, statement A is most likely to be chosen by respondents having trait scores, between -4 and -1, statement B is most likely to be chosen by respondents having trait scores between -1 and +1, and statement C is most likely to be chosen by respondents having higher trait scores.

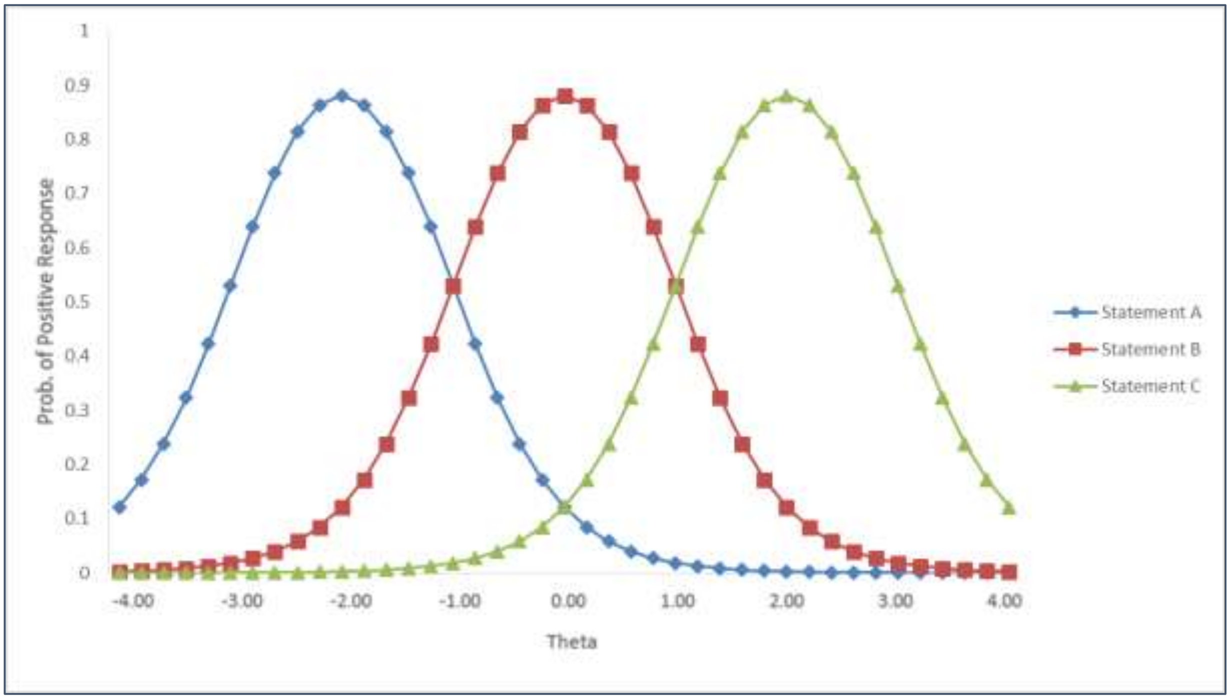


Figure 2. GGUM-PICK response functions for a block involving three statements (A, B, and C) measuring the same dimension. In the panel, α=2, δ= -2, τ= -1 for statement A; α=2, δ= 0, τ= -1 for statement B; α=2, δ= 2, τ= -1 for statement C.

With MFC triplet items, there are 6 possible ways a respondent can rank the statements: 1) A>B>C, 2) A>C>B, 3) B>A>C, 4) B>C>A, 5) C>A>B, 6) C>B>A. The GGUM-RANK probabilities are computed as follows:

$$P_{(\mathrm{A>B>C})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = P_{(\mathrm{A>B,C})} * P_{(\mathrm{B>C})} =$$

$$\frac{P_\mathrm{A}(1)P_\mathrm{B}(0)P_\mathrm{C}(0)}{P_\mathrm{A}(1)P_\mathrm{B}(0)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(1)P_\mathrm{C}(0)+P_\mathrm{A}(0)P_\mathrm{B}(0)P_\mathrm{C}(1)} * \frac{P_\mathrm{B}(1)P_\mathrm{C}(0)}{P_\mathrm{B}(1)P_\mathrm{C}(0)+P_\mathrm{B}(0)P_\mathrm{C}(1)} \tag{7a}$$

$$P_{(\mathrm{A>C>B})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = P_{(\mathrm{A>C,B})} * P_{(\mathrm{C>B})} =$$

$$\frac{P_A(1)P_B(0)P_C(0)}{P_A(1)P_B(0)P_C(0)+P_A(0)P_B(1)P_C(0)+P_A(0)P_B(0)P_C(1)} * \frac{P_B(0)P_C(1)}{P_B(1)P_C(0)+P_B(0)P_C(1)} \tag{7b}$$

$$P_{(\mathrm{B>A>C})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = P_{(\mathrm{B>A,C})} * P_{(\mathrm{A>C})} =$$

$$\frac{P_A(0)P_B(1)P_C(0)}{P_A(1)P_B(0)P_C(0)+P_A(0)P_B(1)P_C(0)+P_A(0)P_B(0)P_C(1)} * \frac{P_A(1)P_C(0)}{P_A(1)P_C(0)+P_A(0)P_C(1)} \tag{7c}$$

$$P_{(\mathrm{B>C>A})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = P_{(\mathrm{B>C,A})} * P_{(\mathrm{C>A})} =$$

$$\frac{P_A(0)P_B(1)P_C(0)}{P_A(1)P_B(0)P_C(0)+P_A(0)P_B(1)P_C(0)+P_A(0)P_B(0)P_C(1)} * \frac{P_A(0)P_C(1)}{P_A(1)P_C(0)+P_A(0)P_C(1)} \tag{7d}$$

$$P_{(\mathrm{C>A>B})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = P_{(\mathrm{C>A,B})} * P_{(\mathrm{A>B})} =$$

$$\frac{P_A(0)P_B(0)P_C(1)}{P_A(1)P_B(0)P_C(0)+P_A(0)P_B(1)P_C(0)+P_A(0)P_B(0)P_C(1)} * \frac{P_A(1)P_B(0)}{P_A(1)P_B(0)+P_A(0)P_C(1)} \tag{7e}$$

$$P_{(\mathrm{C>B>A})_i}\left(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}\right) = P_{(\mathrm{C>B,A})} * P_{(\mathrm{B>A})} =$$

$$\frac{P_A(0)P_B(0)P_C(1)}{P_A(1)P_B(0)P_C(0)+P_A(0)P_B(1)P_C(0)+P_A(0)P_B(0)P_C(1)} * \frac{P_A(0)P_B(1)}{P_A(1)P_B(0)+P_A(0)P_B(1)} \tag{7f}$$

where

“>” means “preferred,”

i = the index for each item (block of three statements; triplet), *i* = 1 to *I*,

A, B, C = the labels for the statements in block *i*,

d = the index for dimensions represented by the statements, *d* = 1, … , *D*,

$\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}}$ = the respondent's latent trait scores on the dimensions represented by the statements A, B, and C in block *i*,

$P_\mathrm{A}(1), P_\mathrm{B}(1)\ , P_\mathrm{C}(1)$ = the probabilities of endorsing statements A, B, and C,

$P_\mathrm{A}(0), P_\mathrm{B}(0), P_\mathrm{C}(0)$ = the probabilities of not endorsing statements A, B, and C,

$P_{(\mathrm{A>B>C})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent ranking statements A, B, and C as their 1st, 2nd, and 3rd choices, respectively, in block *i*,

$P_{(\mathrm{A>C>B})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent ranking statements A, C, and B as their 1st, 2nd, and 3rd choices, respectively, in block *i*,

$P_{(\mathrm{B>A>C})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent ranking statements B, A, and C as their 1st, 2nd, and 3rd choices, respectively, in block *i*,

$P_{(\mathrm{B>C>A})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent ranking statements B, C, and A as their 1st, 2nd, and 3rd choices, respectively, in block *i*,

$P_{(\mathrm{C>A>B})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent ranking statements C, A, and B as their 1st, 2nd, and 3rd choices, respectively, in block *i*,

$P_{(\mathrm{C>B>A})_i}(\theta_{d_\mathrm{A}}, \theta_{d_\mathrm{B}}, \theta_{d_\mathrm{C}})$ = the probability of a respondent ranking statements C, B, and A as their 1st, 2nd, and 3rd choices, respectively, in block *i*.

For the special case of MFC triplets with statements representing the same dimension, the probabilities of the six possible ranks can be displayed like ordinary option response functions. In Figure 3, the vertical axis represents the probability of a ranking given a respondent's trait scores (θ) and the statements' parameters. Because there are six possible ranks, there are six response functions, labeled as shown to the right of the graph. Note that at low trait scores, the ranking A>B>C is most probable, and the probabilities peak for rankings B>A>C, B>C>A and

C>B>A, respectively, as θ increases. Note also that A>C>B and C>A>B have very low probabilities of being observed throughout the range of trait scores.

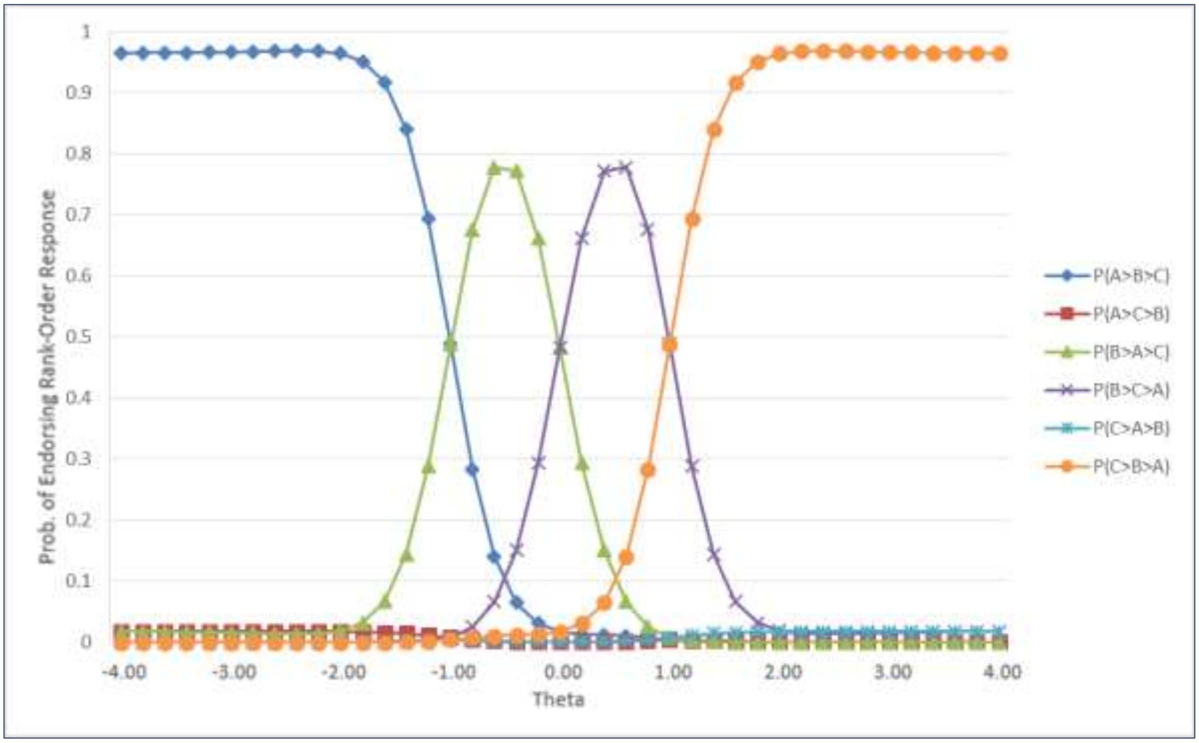


Figure 3. GGUM-RANK option response function selecting each possible rank response in a triplet. In the panel, α=2, δ= -2, τ= -1 for statement A;α=2, δ= 0, τ= -1 for statement B; α=2, δ= 2, τ= -1 for statement C.

GGUM-RANK Item and Test Information Indices

This section describes an analytical solution and numerical approximation methods for GGUM-RANK item information and test information index developed by Joo, Lee, and Stark (2016). Also presented are new scalar (unconditional) information indices to facilitate comparisons across blocks involving different numbers of statements as well as mixed format measures (i.e., a mix of pairs, triplets, and tetrads). Due to the complexity of these functions, the numerical approximations may be preferred in practice.

GGUM-RANK information functions were derived starting with the general definition of *item information* $I_i(\theta)$ for unidimensional polytomous models provided by Samejima (1974):

$$I_i(\theta) = \sum_{m=1}^{M} I_{im}(\theta) P_{im}(\theta), \tag{8a}$$

where

i is the index for items (*i*=1, 2, … , *I*),

m is the index for response categories (*m*=1, 2, …, *M*), and

$I_{im}(\theta) = -\frac{\partial^2 \ln P_{im}(\theta)}{\partial \theta^2} = -\frac{\partial}{\partial \theta} \frac{\left(\frac{\partial P_{im}(\theta)}{\partial \theta}\right)^2}{P_{im}(\theta)}$ is the *item category information* function.

By substitution, it follows that:

$$I_i(\theta) = \sum_{m=1}^{M} \left(-\frac{\partial}{\partial \theta} \frac{\left(\frac{\partial P_{im}(\theta)}{\partial \theta}\right)^2}{P_{im}(\theta)} \right) P_{im}(\theta), \tag{8b}$$

$$I_i(\theta) = \sum_{m=1}^{M} \left(\frac{\left(\frac{\partial P_{im}(\theta)}{\partial \theta}\right)^2 - P_{im}(\theta) \frac{\partial^2 P_{im}(\theta)}{\partial \theta^2}}{[P_{im}(\theta)]^2} \right) P_{im}(\theta), \tag{8c}$$

and,

$$I_i(\theta) = \sum_{m=1}^{M} \left[\frac{\left(\frac{\partial P_{im}(\theta)}{\partial \theta}\right)^2}{P_{im}(\theta)} - \frac{\partial^2 P_{im}(\theta)}{\partial \theta^2} \right]. \tag{8d}$$

Applied to the GGUM-RANK model with three statements per block (triplets), there are M=6 possible rankings or response categories. Consequently, Equation 8d can be rewritten as:

$$I_i(\boldsymbol{\theta}) = \sum_{m=1}^{M=6} \left[\frac{\left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2}{P_{im}(\boldsymbol{\theta})} - \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{T}}} \right], \tag{9a}$$

$$= \frac{\left(\frac{\partial P_{i1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2}{P_{i1}(\boldsymbol{\theta})} - \frac{\partial^2 P_{i1}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{T}}} + \frac{\left(\frac{\partial P_{i2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2}{P_{i2}(\boldsymbol{\theta})} - \frac{\partial^2 P_{i2}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{T}}} + \cdots + \frac{\left(\frac{\partial P_{i6}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2}{P_{i6}(\boldsymbol{\theta})} - \frac{\partial^2 P_{i6}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{T}}}\ , \tag{9b}$$

where

$\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_C)$, the vector of trait scores for the dimensions measured in block *i*,

$P_{i1}(\boldsymbol{\theta}) = P_{A>B>C}(\boldsymbol{\theta}),$

$P_{i2}(\boldsymbol{\theta}) = P_{A>C>B}(\boldsymbol{\theta})\ ,$

$P_{i3}(\boldsymbol{\theta}) = P_{B>A>C}(\boldsymbol{\theta}),$

$P_{i4}(\boldsymbol{\theta}) = P_{B>C>A}(\boldsymbol{\theta}),$

$P_{i5}(\boldsymbol{\theta}) = P_{C>A>B}(\boldsymbol{\theta}),$

$P_{i6}(\boldsymbol{\theta}) = P_{C>B>A}(\boldsymbol{\theta}),$

$\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the Gradient vector of first partial derivatives, and

$\frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{T}}}$ is the Hessian matrix of second partial derivatives.

The Gradient vector mentioned above is obtained by taking partial derivatives of the GGUM-RANK probability function with respect to $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_C)$ as shown:

$$\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial P_{im}(\theta_A,\theta_B,\theta_C)}{\partial \theta_A} & \frac{\partial P_{im}(\theta_A,\theta_B,\theta_C)}{\partial \theta_B} & \frac{\partial P_{im}(\theta_A,\theta_B,\theta_C)}{\partial \theta_C} \end{pmatrix} \quad . \qquad \text{(10a)}$$

To compute GGUM-RANK item information using Equation 9a, we need the "squares" of the first partial derivatives, given by the inner product of the vector expression in Equation 10:

$$\left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^2 = \left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \theta_A}\right)^2 + \left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \theta_B}\right)^2 + \left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \theta_C}\right)^2 \quad . \qquad \text{(10b)}$$

Next, we need the Hessian matrix of second partial derivatives of the GGUM-RANK probability function with respect to $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_C)$:

$$\frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\boldsymbol{T}}} = \frac{\partial}{\partial \boldsymbol{\theta}^{\boldsymbol{T}}} \begin{pmatrix} \frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \theta_A} & \frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \theta_B} & \frac{\partial P_{im}(\boldsymbol{\theta})}{\partial \theta_C} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_A^2} & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_A \partial \theta_B} & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_A \partial \theta_C} \\ \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_B \partial \theta_A} & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_B^2} & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_B \partial \theta_C} \\ \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_C \partial \theta_A} & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_C \partial \theta_B} & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial \theta_C^2} \end{pmatrix} \text{(11a)}$$

Because it is assumed that the statements measuring each dimension are evaluated independently by a respondent, the mixed partial derivatives can be set to zero, as shown in Equation 11b:

$$\frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\boldsymbol{T}}} = \begin{pmatrix} \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_A^2} & 0 & 0 \\ 0 & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_B^2} & 0 \\ 0 & 0 & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_C^2} \end{pmatrix}, \tag{11b}$$

To combine this result with the numerator term on the left side of Equation 9a for GGUM-RANK item information, the trace of the Hessian is needed:

$$\frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\boldsymbol{T}}} = tr\begin{pmatrix} \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_A^2} & 0 & 0 \\ 0 & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_B^2} & 0 \\ 0 & 0 & \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_C^2} \end{pmatrix} = \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_A^2} + \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_B^2} + \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_C^2}\,. \tag{12}$$

By substituting Equations 10b and 12 into 9a, we obtain *item information for GGUM-RANK triplets*:

$$I_i(\boldsymbol{\theta}) = \sum_{m=1}^{M=6} \frac{\left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial\theta_A}\right)^2 + \left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial\theta_B}\right)^2 + \left(\frac{\partial P_{im}(\boldsymbol{\theta})}{\partial\theta_C}\right)^2}{P_{im}(\boldsymbol{\theta})} - \left(\frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_A^2} + \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_B^2} + \frac{\partial^2 P_{im}(\boldsymbol{\theta})}{\partial\theta_C^2}\right)\,. \tag{13}$$

GGUM-RANK test information, $I(\boldsymbol{\theta})$, is then obtained by summing the item information functions:

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)\,. \tag{14}$$

Numerical Solution for GGUM-RANK Overall Item Information

Note that examples of first and second partial derivatives of the GGUM-RANK probability function with respect to $\boldsymbol{\theta} = (\theta_A, \theta_B, \theta_C)$ are provided in Appendix A. Because the analytical solution for item information is complex, an Ox computer program (Doornik, 2009) that numerically approximates the Gradient (Equation 10) and Hessian (Equation 11a) was developed and tested by Joo et al. (2016) for special cases that could be vetted. An *overall item information (OII) index* was also developed to allow comparisons of item quality with MFC blocks involving different numbers of statements and/or dimensions. To compute OII, 10,000 random vectors of trait scores are drawn from a multivariate standard normal distribution, item information conditional on the trait scores is computed, and the information values are averaged over samples and dimensions to obtain a scalar index of item quality that can be used to judge the benefit of administering one item relative to another in the examinee population. Because OII is a scalar quantity, the values can then be summed across items to compute an *overall test information (OTI)* index that varies in association with the precision of estimated trait scores. More specifically, items and tests having larger OII and OTI values should allow, on average, more accurate assessment of examinees.

Estimating GGUM-RANK Item Parameters

To date, multiple methods have been developed to estimate item and person parameters in the context of IRT. Joint Maximum Likelihood (JML) and Marginal Maximum Likelihood (MML) estimation are well known examples. In JML estimation, the likelihood is jointly maximized in terms of item and person parameters through an alternating sequence of steps. More specifically, JML starts by assuming initial values for the item parameters. These “provisional” item parameters are used with observed data to estimate person parameters via maximum likelihood; then these person parameter estimates are used to obtain better item parameter estimates. This alternating sequence of maximization steps continues until the changes across iterations fall below some predetermined threshold, indicating that the algorithm has converged.

In contrast to this back and forth process for estimating parameters, which sometimes leads to convergence problems, MML estimation uses an Expectation-Maximization (EM) algorithm (Bock & Aitken, 1981), which begins by assuming a prior distribution of trait scores (e.g., standard normal) that serves to identify the scale of measurement. The trait continuum is divided into a series of discrete points (quadrature nodes). At each node, the expected number of responses and the expected number of correct responses are computed, and these “pseudocounts” are used to compute the likelihood of the observed data matrix (E-step). Next, the likelihood of the data is maximized by using a numerical method, such as Newton-Raphson iterations, which requires starting values for the item parameters and first and second derivatives of the likelihood function (M-step). The values that maximize the likelihood are used as the item parameters for the next E-step, and this process continues until the changes across cycles are sufficiently small, indicating convergence. Research has shown that this approach is effective and produces item parameter estimates that are consistent. However, as with JML, the need for derivatives in the M-Step makes this approach difficult to implement with complex IRT models.

In the last decade, Markov Chain Monte Carlo (MCMC) estimation emerged as a viable alternative to JML and MML methods that is well suited for complex IRT models, such as the GGUM-RANK (e.g., Bolt & Lall, 2003; de la Torre, Stark, & Chernyshenko, 2006; Sinharay, Johnson, & Stern, 2006; Johnson & Junker, 2003; Patz & Junker, 1999). Like JML, MCMC methods estimate item and person parameters in tandem; however, MCMC methods do not require derivatives. Instead, prior distributions for the item and person parameters are assumed, the likelihood of the data are computed at “candidate” values sampled from those distributions, and probabilistic decisions are made concerning the acceptability of the candidate values over many (e.g., 50,000) iterations. After a “burn in” period of several hundred to several thousand iterations, which are needed for the algorithm to reach a steady state, the item and person parameters accepted/retained on each iteration are averaged to obtain the MCMC parameter estimates, and the standard deviation of these post-burn-in values gives their standard error. At this point, various diagnostics can be examined to determine whether the algorithm converged, or whether the starting values, priors, and number of cycles need to be adjusted for a subsequent run.

Several different MCMC methods have been explored (e.g., *Gibbs sampling* (Geman & Geman, 1984; *Metropolis-Hastings* (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), and *Metropolis-Hasting within Gibbs* (MHWG; Tierney, 1994)). They differ in how a new draw in the Markov chain is sampled based on the previous draw. As discussed by Patz and Junker (1999), MHWG is a flexible approach combining MH and Gibbs sampling that is amenable to IRT applications. For that reason, the MHWG method was chosen for estimating GGUM-RANK triplet item and person parameters in this research.

The MHWG Algorithm

To implement the MHWG algorithm for the GGUM-RANK model, the likelihood of the rank response data given all model parameters $(\boldsymbol{\theta}, \alpha, \delta, \tau)$ must be specified:

$$\mathrm{P}(\boldsymbol{X} \,|\boldsymbol{\theta}, \alpha, \delta, \tau) = \prod_{n}^{N} \prod_{i}^{I} P_{ni}(\boldsymbol{\theta}_n), \tag{15}$$

where

n = índex for respondents, *n* =1, 2, …, *N*,

i = index for items (blocks), *i* = 1, 2, …, *I*,

$\boldsymbol{X} =$ (NxI) matrix of rank responses, and

$P_{ni}(\boldsymbol{\theta}_n)$ = the GGUM-RANK probability for person *n*’s ranking of the statements in item *i*, computed using Equations (7a) through (7f).

In the MHWG algorithm, all model parameters are updated individually on each iteration. A step-by-step description is provided below.

- For each parameter, initial values $(\theta^0, \alpha^0,\ \delta^0, \tau^0)$ are determined based on prior information.
- Each model parameter is updated sequentially on each iteration t:
    - Proposed *candidate* values ($\boldsymbol{\theta}^*$), for ten sets of latent trait scores, are drawn from independent normal distributions with mean of value at the *t*-1 state and specified variances to yield adequate acceptance rates (Patz & Junker, 1999a): $\boldsymbol{\theta}^* \sim N(\theta^{t-1}, \sigma^2)$.
        - For each set of $\boldsymbol{\theta}^*$, an acceptance probability is obtained by dividing the posterior probability at the proposed state by the posterior probability at the current state (i.e., $\frac{P(\theta_{Proposed})}{P(\theta_{Current})}$).
        - If $\frac{P(\theta_{Proposed})}{P(\theta_{Current})} > 1$, the proposed values of $\boldsymbol{\theta}^*$ is accepted.
        - If $\frac{P(\theta_{Proposed})}{P(\theta_{Current})} < 1$, the proposed values of $\boldsymbol{\theta}^*$ is accepted probabilistically. That is, if the acceptance probability is greater than the random uniform number, the proposed set of $\boldsymbol{\theta}^*$ is accepted. If not, the proposed value at previous state (*t*-1) is retained.
        - This process can be expressed in Equation 16:

$$\mathrm{P}(\boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^*) = min\left\{\frac{\left(X\middle|\boldsymbol{\theta}^*, \alpha^{(t-1)}, \delta^{(t-1)}, \tau^{(t-1)}\right)_{P(\boldsymbol{\theta}^*)}}{P\left(X\middle|\boldsymbol{\theta}^{(t-1)}, \alpha^{(t-1)}, \delta^{(t-1)}, \tau^{(t-1)}\right)_{P(\boldsymbol{\theta}^{(t-1)})}}, 1\right\}. \tag{16}$$

    - Proposed candidate values (α^*) for statement discrimination parameters are drawn from $N(\alpha^{t-1}, \sigma^2)$.

- An acceptance probability for each α^* is obtained by dividing the posterior probability at the proposed state by the posterior probability at the current state (i.e., $\frac{P(\alpha_{Proposed})}{P(\alpha_{Current})}$).
- If $\frac{P(\alpha_{Proposed})}{P(\alpha_{Current})} > 1$, the proposed values of α^* is accepted.
- If $\frac{P(\alpha_{Proposed})}{P(\alpha_{Current})} < 1$, the proposed values of α^* is accepted probabilistically.
- This process can be expressed in Equation 17:

$$\mathrm{P}(\alpha^{t-1}, \alpha^*) = min\left\{\frac{P\left(X\middle|\theta^t, \alpha^*, \delta^{(t-1)}, \tau^{(t-1)}\right)P(\alpha^*)}{P\left(X\middle|\theta^t, \alpha^{(t-1)}, \delta^{(t-1)}, \tau^{(t-1)}\right)P(\alpha^{(t-1)})}, 1\right\}. \quad (17)$$

- Proposed candidate values (δ^*) for statement location parameters are drawn from N(δ^{t-1}, σ^2).
  - An acceptance probability for each δ^* is obtained by dividing the posterior probability at the proposed state by the posterior probability at the current state (i.e., $\frac{P(\delta_{Proposed})}{P(\ \delta_{Current})}$).
  - If $\frac{P(\delta_{Proposed})}{P(\ \delta_{Current})} > 1$, the proposed values of δ^* is accepted.
  - If $\frac{P(\delta_{Proposed})}{P(\ \delta_{Current})} < 1$, the proposed values of δ^* is accepted probabilistically.
  - This process can be expressed in Equation 18:

$$\mathrm{P}(\delta^{t-1}, \delta^*) = min\left\{\frac{P\left(X\middle|\theta^t, \alpha^t, \delta^*, \tau^{(t-1)}\right)P(\delta^*)}{P\left(X\middle|\theta^t, \alpha^t, \delta^{(t-1)}, \tau^{(t-1)}\right)P(\delta^{(t-1)})}, 1\right\}. \quad (18)$$

- Proposed candidate values(τ^*) for statement location parameters are drawn from N(τ^{t-1}, σ^2).

  - An acceptance probability for each τ* is obtained by dividing the posterior probability at the proposed state by the posterior probability at the current state (i.e., $\frac{P(\tau_{Proposed})}{P(\ \tau_{Current})}$).
  - If $\frac{P(\tau_{Proposed})}{P(\ \tau_{Current})} > 1$, the proposed values of τ* is accepted.
  - If $\frac{P(\tau_{Proposed})}{P(\ \tau_{Current})} < 1$, the proposed values of τ* is accepted probabilistically.
  - This process can be expressed in Equation 19:

$$\mathrm{P}(\tau^{t-1}, \tau^{*}) = min\left\{\frac{P\left(X\middle|\theta^{t}, \alpha^{t}, \delta^{t}, \tau^{*}\right)P(\tau^{*})}{P\left(X\middle|\theta^{t}, \alpha^{t}, \delta^{t}, \tau^{(t-1)}\right)P(\tau^{(t-1)})}, 1\right\}. \tag{19}$$

- This procedure continues until a specified number of cycles is reached and the estimated model parameters are recorded on each cycle. The values before the burn-in period are typically excluded. The parameter estimates, standard errors, and covariances were obtained using means, variances, and covariances of model parameters after burn-in period.

An Ox (Doornik, 2009) computer program was created to estimate GGUM-RANK parameters using this MCMC algorithm. Prior distributions for the model parameters $(\boldsymbol{\theta}, \alpha, \delta, \tau)$ and proposal variances for the specific simulation conditions in this research were chosen by following the recommendations of Patz and Junker (1999) and examining some pilot testing results. The next section describes the simulation design and process to evaluate the recovery of GGUM-RANK item and person parameters using this MHWG approach.

CHAPTER TWO:

METHOD

Study1

Study 1 investigated the accuracy of MCMC parameter recovery using an MHWG algorithm developed for GGUM-RANK triplet responses. Simulation conditions were chosen to reflect choices made in practice concerning MFC test construction, and generating parameters were obtained in accordance with other MFC scoring and GGUM estimation studies (e.g., Hontangas et al., 2015; Koenig & Roberts, 2007; Roberts et al., 2000).

For this study, MFC test dimensionality was set at 10 dimensions, because tests in the field usually involve 10 or more dimensions (e.g., Stark et al., 2014), and tests of higher dimensionality would require more items and lead to excessive run-times. Four independent variables, shown below, were fully crossed to produce 16 experimental conditions. Because pilot testing revealed that one replication would take 11-47 hours, depending on the condition, the number of replications in each condition was set at 20.

Simulation Study Design

1) Sample size (2):
    a) N = 250, and
    b) N = 500.
2) Test length (2):

a) 30-Triplet, and

b) 60-Triplet (the first 30 triplets were duplicated).

3) Intrablock discrimination (2):

a) Low: α sampled randomly from uniform distribution, U(0.75, 1.25), and

b) High : α sampled randomly from uniform distribution, U(1.75, 2.25).

4) Intrablock location parameter variability (2):

a) Low: δ standard deviation (SD) ≈ 0.3, and

b) High: δ standard deviation (SD) ≈ 1.3.

The same threshold (τ) parameters were used in each experimental condition; these were sampled from a uniform distribution, *U*(-1.4, -0.4). In the *intrablock location parameter variability conditions*, half of the δ parameters were sampled from a *U*(-2, 0), and the other half were sampled from a *U*(0, 2). Then the large variability (δ SD = 1.3) and small variability (δ SD = 0.3) conditions were fulfilled by mixing the generated δ parameters.

MFC Tests Constructed for this Simulation. To prepare for this study, 10-D MFC tests were built in accordance with the study design. Appendices B1 through B4 present the specifications for the 30-Triplet measures. On the left side of each table, column 1 shows the triplet (item) number, column2 shows the statement number, column 3 shows the dimension number represented by the statement, columns 4 through 6 show the generating statement discrimination, location, and threshold parameters, respectively (α, δ, and τ), and so on for the remaining columns.

With 10-D MFC tests, there are 120 possible combinations of three dimensions that could be chosen for test construction (excluding repeats within blocks). 30-Triplet measures were created by choosing from these combinations, so that each dimension appeared in 9 different blocks. Item parameters were selected to satisfy the design specifications in the particular conditions. Next, 60-Triplet measures were formed by duplicating the item parameters in the 30-Triplet measures, while assigning different dimension numbers to make well-balanced tests. The dimensionality specifications for the 30-Triplet and 60-Triplet MFC tests are shown in Table 1.

Table 1. Dimension Specification of 30-Triplet and 60-Triplet Tests.

30-Triplet				60-Triplet							
	Statements				Statements				Statements		
Block #	A	B	C	Block #	A	B	C	Block #	A	B	C
1	1	2	3	1	1	2	3	31	1	2	4
2	1	6	9	2	1	6	9	32	1	2	9
3	4	5	8	3	4	5	8	33	1	3	7
4	6	7	8	4	6	7	8	34	1	8	9
5	2	6	7	5	2	6	7	35	2	3	6
6	8	9	10	6	8	9	10	36	2	5	8
7	2	3	5	7	2	3	5	37	3	4	10
8	1	4	9	8	1	4	9	38	3	9	10
9	1	2	10	9	1	2	10	39	4	7	8
10	3	6	8	10	3	6	8	40	4	9	10
11	4	8	9	11	4	8	9	41	5	7	8
12	6	7	9	12	6	7	9	42	6	8	9
13	2	7	10	13	2	7	10	43	7	8	10
14	1	9	10	14	1	9	10	44	1	2	6
15	1	2	7	15	1	2	7	45	1	4	7
16	4	6	10	16	4	6	10	46	1	6	10
17	5	8	10	17	5	8	10	47	2	3	7
18	3	4	5	18	3	4	5	48	2	4	5
19	1	7	10	19	1	7	10	49	3	5	8
20	1	3	5	20	1	3	5	50	4	6	8
21	5	6	7	21	5	6	7	51	4	6	7
22	3	4	6	22	3	4	6	52	5	6	8
23	7	8	9	23	7	8	9	53	6	9	10
24	3	8	10	24	3	8	10	54	7	9	10
25	2	3	9	25	2	3	9	55	1	2	5
26	1	5	10	26	1	5	10	56	1	3	10
27	4	5	6	27	4	5	6	57	2	5	6
28	4	5	9	28	4	5	9	58	3	5	7
29	2	3	4	29	2	3	4	59	3	4	9
30	2	7	8	30	2	7	8	60	5	9	10

In Table 1, note that triplets 1-30 represent the same dimensions in the 30- and 60- triplet tests. However, triplets 31-60 in the longer tests involve different combinations of dimensions. Also, in the 60-Triplet tests, each dimension is represented 18 times, as compared to 9 times in the 30-Triplet tests.

In sum, 8 MFC tests were created for this study with characteristics shown:

1) 30-Triplet Tests (4):

a) low α / low δ SD,

b) low α / high δ SD,

c) high α / low δ SD, and

d) high α / high δ SD.

2) 60-Triplet Tests (4):

a) low α / low δ SD,

b) low α / high δ SD,

c) high α / low δ SD, and

d) high α / high δ SD.

Table 2 shows the average generating parameters by dimension for the 8 test designs to illustrate that the psychometric properties of the statements representing the dimensions are similar.

Table 2. Average Generating Parameter by Each Dimension.

			Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Total
30-Triplet	Low α / Low δ SD	avg.α	0.98	0.97	0.97	0.99	0.94	1.01	0.98	1.03	0.96	0.88	0.97
		avg.δ	0.02	0.08	0.03	0.08	0.07	0.06	-0.06	-0.04	-0.09	0.09	0.02
		avg.τ	-0.89	-0.96	-0.83	-0.89	-0.82	-0.87	-1.04	-0.80	-0.97	-0.78	-0.88
	Low α / High δ SD	avg.α	0.98	0.97	0.97	0.99	0.94	1.01	0.98	1.03	0.96	0.88	0.97
		avg.δ	-0.09	0.08	0.04	-0.01	0.05	0.09	-0.08	-0.05	0.09	0.01	0.01
		avg.τ	-0.89	-0.96	-0.83	-0.89	-0.82	-0.87	-1.04	-0.80	-0.97	-0.78	-0.88
	High α / Low δ SD	avg.α	2.00	1.99	1.95	1.98	2.01	2.05	1.97	1.98	1.92	2.03	1.99
		avg.δ	0.02	0.08	0.03	0.08	0.07	0.06	-0.06	-0.04	-0.09	0.09	0.02
		avg.τ	-0.89	-0.96	-0.83	-0.89	-0.82	-0.87	-1.04	-0.80	-0.97	-0.78	-0.88
	High α / High δ SD	avg.α	2.00	1.99	1.95	1.98	2.01	2.05	1.97	1.98	1.92	2.03	1.99
		avg.δ	-0.09	0.08	0.04	-0.01	0.05	0.09	-0.08	-0.05	0.09	0.01	0.01
		avg.τ	-0.89	-0.96	-0.83	-0.89	-0.82	-0.87	-1.04	-0.80	-0.97	-0.78	-0.88
60-Triplet	Low α / Low δ SD	avg.α	0.97	0.95	0.99	0.99	0.98	0.95	0.91	1.00	1.01	0.93	0.97
		avg.δ	-0.22	0.11	0.23	-0.02	0.05	0.12	-0.07	-0.15	-0.09	-0.05	-0.01
		avg.τ	-0.92	-0.92	-0.86	-0.91	-0.88	-0.88	-0.99	-0.77	-0.99	-0.73	-0.89
	Low α / High δ SD	avg.α	0.97	0.95	0.99	0.99	0.98	0.95	0.91	1.00	1.01	0.93	0.97
		avg.δ	-0.22	0.12	0.19	-0.28	0.12	0.03	-0.17	0.02	-0.07	0.04	-0.02
		avg.τ	-0.92	-0.92	-0.86	-0.91	-0.88	-0.88	-0.99	-0.77	-0.99	-0.73	-0.89
	High α / Low δ SD	avg.α	1.99	2.00	2.00	1.98	1.98	2.02	1.97	1.98	1.93	2.04	1.99
		avg.δ	-0.22	0.11	0.23	-0.02	0.05	0.12	-0.07	-0.15	-0.09	-0.05	-0.01
		avg.τ	-0.92	-0.92	-0.86	-0.91	-0.88	-0.88	-0.99	-0.77	-0.99	-0.73	-0.89
	High α / High δ SD	avg.α	1.99	2.00	2.00	1.98	1.98	2.02	1.97	1.98	1.93	2.04	1.99
		avg.δ	-0.22	0.12	0.19	-0.28	0.12	0.03	-0.17	0.02	-0.07	0.04	-0.02
		avg.τ	-0.92	-0.92	-0.86	-0.91	-0.88	-0.88	-0.99	-0.77	-0.99	-0.73	-0.89

Simulation Details

Data Generation. GGUM-RANK response data were generated using an Ox (Doornik, 2009) computer program. For each respondent, a vectors of 10 latent trait scores ($\boldsymbol{\theta}$) were sampled from a multivariate standard normal distribution with the covariance among dimensions set to zero for this initial study. Using these trait scores and statement parameters for each experimental condition, GGUM-RANK probabilities were computed using Equations 7a-7f. These response probabilities were compared to random uniform numbers, and ranks were assigned using a decision scheme analogous to data generation with polytomous IRT models.

1. Order the six possible rankings for MFC triplets as:

   r_1 = Ranking1 = A>B>C with numerical code 123

   r_2 = Ranking2 = A>C>B with numerical code 132

   r_3 = Ranking3 = B>A>C with numerical code 213

   r_4 = Ranking4 = B>C>A with numerical code 231

   r_5 = Ranking5 = C>A>B with numerical code 312

   r_6 = Ranking6 = C>B>A with numerical code 321

2. Compare the probabilities of Rankings 1-6 to a randomly sampled uniform number (“rand”) and assign response codes as follows:

   a. If $\{1-P(r_1)\}$ < rand, then code response as A>B>C or 123.

   b. If $\{1-(P(r_1)+ P(r_2))\} < r$, then code response as A>C>B or 132.

   c. If $\{1-(P(r_1)+P(r_2)+P(r_3))\} < r$, then code response as B>A>C or 213.

   d. If $\{1-(P(r_1)+ P(r_2)+P(r_3)+P(r_4))\} < r$, then code response as B>C>A or 231.

   e. If $\{1-(P(r_1)+ P(r_2)+P(r_3)+P(r_4)+P(r_5))\} < r$, then code response as C>A>B or 312.

   f. Otherwise, code response as C>B>A or 321.

MCMC Convergence Checks. Convergence implies that a chain has reached a stationary state so that samples are being drawn from the desired posterior distributions. Before examining parameter recovery, the convergence of the MCMC algorithm was assessed using the Gelman-Rubin diagnostic index ($\hat{R}$) (for details, see Gelman & Rubin, 1992). This index uses the samples from multiple (e.g., three) independent chains after the burn-in periods. For any given parameter, the $\hat{R}$ statistic assesses the ratio of the between-chain variation to the within-chain variation. If the chains have converged, the between-chain and within-chain variation will be near 1; otherwise, larger ratios will be observed. Strict convergence is met when $\hat{R} < 1.20$ for all parameters (de la Torre et al., 2012).

In preparation for this study, pilot simulations were conducted to examine the convergence of the new MHWG algorithm in various conditions. It was found that convergence occurred at approximately 30,000 iterations. Thus, 30,000 iterations with three Markov chains were performed in the proposed study, and the first 15,000 iterations from each chain were discarded as a burn-in period. Post-burn-in samples were used to compute the mean and standard deviation of the posterior distributions.

Indices of Estimation Accuracy. To evaluate the efficacy of GGUM-RANK parameter estimation, several indices were used. First, Pearson correlations were computed between the true parameters (α, δ, τ, and $\boldsymbol{\theta}$) and estimated parameters ($\hat{\alpha}, \hat{\delta}, \hat{\tau},$ and $\widehat{\boldsymbol{\theta}}$) on each replication. These correlations were then averaged across replications, recorded, and averaged across dimensions to obtain a single indicator of parameter recovery.

Second, absolute biases (ABS) for parameters (α, δ, τ, and $\boldsymbol{\theta}$) were computed as follows:

$$ABS\ (\hat{\alpha}) = \frac{\sum_j |\hat{\alpha} - \alpha|}{n},$$

$$ABS\ (\hat{\delta}) = \frac{\sum_j |\hat{\delta} - \delta|}{n},$$

$$ABS\ (\hat{\tau}) = \frac{\sum_j |\hat{\tau} - \tau|}{n}, \text{ and}$$

$$ABS\ (\hat{\theta}_d) = \frac{\sum_j |\hat{\theta}_d - \theta_d|}{n},$$

where

j is the statement number,

n is the total number of statements, and

d represents the dimension associated with trait score θ.

The absolute biases of the statement parameters were averaged across replications, recorded, and then averaged across dimensions. Smaller absolute biases indicate better parameter recovery.

Third, root mean square errors (RMSE) was computed for statement and person parameter estimates as follows:

$$RMSE\ (\hat{\alpha}) = \sqrt{\frac{\sum_j (\hat{\alpha} - \alpha)^2}{n}},$$

$$RMSE\ (\hat{\delta}) = \sqrt{\frac{\sum_j (\hat{\delta} - \delta)^2}{n}}\ ,$$

$$RMSE\ (\hat{\tau}) = \sqrt{\frac{\sum_j (\hat{\tau} - \tau)^2}{n}}\ , \text{ and}$$

$$RMSE\ (\hat{\theta}_d) = \sqrt{\frac{\sum_j (\hat{\theta}_d - \theta_d)^2}{n}}.$$

The computed RMSEs for parameters were averaged across replications. Particularly, RMSEs for trait scores involving the 10 dimensions were reported individually and then averaged across dimensions. Smaller RMSEs indicate better parameter recovery.

Fourth, posterior standard deviations (PSDs) were computed for the parameter estimates. PSDs were obtained by taking the square root of the variance of the posterior samples after burn-in. Smaller PSDs indicate better parameter recovery.

MCMC Estimation Prior Distribution and Initial Parameter Values

Prior Distributions for MCMC Estimation. Prior distributions must be specified for all item and person parameters. For this investigation, the following prior distributions were chosen:

$$P(\theta_d)\sim N(0,1),$$

$$P(\alpha)\sim Beta(v_\alpha,\omega_\alpha,a_\alpha,b_\alpha),$$

$$P(\delta)\sim Beta(v_\delta,\omega_\delta,a_\delta,b_\delta),\text{ and}$$

$$P(\tau)\sim Beta(v_\tau,\omega_\tau,a_\tau,b_\tau),$$

where $Beta(v,\omega,a,b)$ is the four-parameter beta distribution with shape parameters (v,ω) and support parameters (a,b), which define the lower and upper bounds of the probability region. One attractive feature of the four-parameter beta distribution is its flexibility; by changing the shape and support parameters, probability functions of many different forms can be produced, making it an excellent choice for MCMC applications. In this study, the four-parameter beta priors {1.5, 1.5, .25, 4}, {2, 2, -4, 4} and {2, 2, -3, 1} were used for α, δ, and τ estimation, respectively. For the person parameters associated with each dimension (d), i.e., (θ_d), a N(0, 1) prior was used.

Initial Parameter Values. To start MCMC estimation, initial values must be set for all parameters. In this study, all θ parameters were initialized to zero. δ parameters were initialized to -1 or 1, in accordance with the sign of the true parameter. (This follows other simulation studies, which have assumed that subject matter experts can readily determine whether a statement is positive or negative by looking at its content (e.g., Seybert, 2013). All α parameters were initialized to 1, and all τ parameters were initialized to -1.

Overview of the Simulation Process

1) GGUM-RANK triplet responses were generated for samples of 250 and 500 respondents, for 10-D 30-Triplet and 60-Triplet tests, using the statement parameters in Appendix B Tables 1-4 and person parameters randomly sampled from a multivariate standard normal distribution.
2) Statement and person parameters were estimated directly from the GGUM-RANK triplet responses, using the MHWG algorithm developed for this dissertation, and the results were saved.
3) Steps 1 and 2 were performed 20 times.
4) Upon completion of the 20 replications, a convergence assessment was performed. Individual parameters that did not show convergence (e.g., $\hat{R}$>1.2) were excluded from subsequent parameter recovery evaluation.
5) Indices of parameter recovery (correlations between estimated and true values, RMSEs, absolute biases, posterior standard deviations) were computed and evaluated for each experimental condition. Also the overall item and test information (OII and OTI) values were averaged over replications and reported.

Hypotheses

For this study, the following hypotheses were proposed.

1) More accurate parameter recovery will be obtained in the larger sample (N = 500) conditions than in the small sample (N = 250) conditions, as indicated by larger Pearson correlations with true parameters and lower absolute bias, RMSE and PSD statistics.
2) More accurate parameter recovery will be observed with the longer (60-Triplet) tests than with the shorter (30-Triplet) tests, because longer tests provide more information. (Each dimension is represented 18 times in the long tests vs. 9 times in the short tests.)
3) More accurate parameter recovery will be observed with tests having high intrablock discrimination than with tests having low intrablock discrimination, because higher discrimination is associated with higher item and test information.
4) More accurate parameter recovery will be observed with tests having large intrablock location variability than with tests having small intrablock location variability.

Using SPSS version 22, these hypotheses were tested using MANOVA with sample size, test length, intrablock discrimination, and intrablock location variability as the between subject factors, and the correlations, absolute biases, RMSEs, and PSDs as dependent variables. Partial eta squared (η_p^2) was used to evaluate effect size. (Note that partial eta squared (η_p^2) equals eta squared (η^2) in one-way ANOVA as there is only one factor (Lakens, 2013)). Values of .01, .06, and .14 represent small, medium, and large effects, respectively (Cohen, 1998). Additionally, parameter recovery and information plots were visually inspected.

Study2

In applied settings, there is growing interest in the benefits of MFC triplet measures with respect to pairwise preference tests that have been used for noncognitive assessment (Stark et al., 2014). Triplet measures may be more cognitively demanding and will almost certainly take longer to complete, so there must be some demonstrated psychometric benefits to justify the more complex format. Study 2 explored these potential psychometric benefits by comparing GGUM-RANK test information and parameter recovery with 10-dimension MFC triplet and pairwise preference measures in a select set of conditions.

Simulation Study Design

1) Sample size (2):
   a. N = 250, and
   b. N = 500.
2) MFC "test type" (3):
   a. 30-Pair
   b. 90-Pair
   c. 30-Triplet

Fully crossing these two independent variables led to the 6 conditions that were explored.

MFC Test Design. From Study 1, the 30-Triplet test in the high intrablock discrimination, high intrablock location variability condition was selected. The 30 triplets were decomposed into the 90 possible pairs to create a 90-Pair MFC test for this study. A 30-Pair MFC test was then created by selecting two statements in each of the 30 triplet blocks. Table 3 displays the average *α, δ,* and τ parameters by dimension for the 30-Pair and 90-Pair tests to illustrate that the psychometric properties of the statements representing the dimensions are similar.

Table 3. Average Generating Parameters for Each Dimension in the 30-Pair and 90-Pair Test Conditions.

		Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Total
30-Pair	avg.α	2.01	1.99	1.96	1.96	1.97	2.00	1.98	1.95	1.94	2.08	1.98
	avg.δ	0.02	-0.02	0.16	0.36	-0.14	-0.04	-0.19	-0.10	0.39	0.19	0.06
	avg.τ	-0.89	-0.92	-0.92	-0.73	-0.82	-0.81	-1.05	-0.82	-1.07	-0.76	-0.88
90-Pair	avg.α	2.00	1.99	1.95	1.98	2.01	2.06	1.98	1.98	1.92	2.03	1.99
	avg.δ	-0.09	0.08	0.04	-0.01	0.05	0.09	-0.10	-0.05	0.09	0.01	0.01
	avg.τ	-0.89	-0.96	-0.83	-0.89	-0.82	-0.87	-1.04	-0.80	-0.97	-0.78	-0.88

Table 4 displays the dimension specifications for the 30-Triplet, 30-Pair, and 90-Pair MFC tests. As described above, the first two statements in triplet 1, representing dimensions {1, 2, 3}, were selected to create the first item in the 30-Pair test, representing dimensions {1, 2}. A similar process was used to create the remaining pairwise preference items. In contrast, the 90-Pair test was created by breaking each triplet into all possible pairs. For example, triplet 1 involving dimensions {1, 2, 3} was decomposed into three pairs involving dimensions {1,2}, {1,3}, and {2,3}, respectively. The statement parameters for the 30-Pair and 90-Pair tests used in Study 2 are presented in Appendix C.

Table 4. Dimension Specification of 30-Triplet, 30-Pair, and 90-Pair Tests.

30-Triplet				30-Pair			90-Pair								
	Statements				Statements		Statements								
Block #	A	B	C	Block #	A	B	Block #	A	B	Block #	A	B	Block #	A	B
1	1	2	3	1	1	2	1	1	2	31	4	8	61	5	6
2	1	6	9	2	6	9	2	1	3	32	4	9	62	5	7
3	4	5	8	3	5	8	3	2	3	33	8	9	63	6	7
4	6	7	8	4	7	8	4	1	6	34	6	7	64	3	4
5	2	6	7	5	6	7	5	1	9	35	6	9	65	3	6
6	8	9	10	6	9	10	6	6	9	36	7	9	66	4	6
7	2	3	5	7	3	5	7	4	5	37	2	7	67	7	8
8	1	4	9	8	1	9	8	4	8	38	2	10	68	7	9
9	1	2	10	9	1	10	9	5	8	39	7	10	69	8	9
10	3	6	8	10	6	8	10	6	7	40	1	9	70	3	8
11	4	8	9	11	4	9	11	6	8	41	1	10	71	3	10
12	6	7	9	12	6	9	12	7	8	42	9	10	72	8	10
13	2	7	10	13	2	10	13	2	6	43	1	2	73	2	3
14	1	9	10	14	9	10	14	2	7	44	1	7	74	2	9
15	1	2	7	15	2	7	15	6	7	45	2	7	75	3	9
16	4	6	10	16	4	6	16	8	9	46	4	6	76	1	5
17	5	8	10	17	8	10	17	8	10	47	4	10	77	1	10
18	3	4	5	18	4	5	18	9	10	48	6	10	78	5	10
19	1	7	10	19	1	7	19	2	3	49	5	8	79	4	5
20	1	3	5	20	1	3	20	2	5	50	5	10	80	4	6
21	5	6	7	21	5	6	21	3	5	51	8	10	81	5	6
22	3	4	6	22	3	4	22	1	4	52	3	4	82	4	5
23	7	8	9	23	7	8	23	1	9	53	3	5	83	4	9
24	3	8	10	24	3	8	24	4	9	54	4	5	84	5	9
25	2	3	9	25	2	3	25	1	2	55	1	7	85	2	3
26	1	5	10	26	1	10	26	1	10	56	1	10	86	2	4
27	4	5	6	27	4	5	27	2	10	57	7	10	87	3	4
28	4	5	9	28	4	5	28	3	6	58	1	3	88	2	7
29	2	3	4	29	2	3	29	3	8	59	1	5	89	2	8
30	2	7	8	30	2	7	30	6	8	60	3	5	90	7	8

Overview of the Simulation Process

1) GGUM-RANK pair responses were generated for samples of 250 and 500 respondents, for 10-D 30-Pair and 10-D 90-Pair tests, using the statement parameters in Appendix C and person parameters randomly sampled from a multivariate standard normal distribution.

2) Statement and person parameters were estimated directly from the GGUM-RANK pair responses, using the MHWG algorithm developed for this dissertation, and the results were saved.
3) Steps 1 and 2 were performed 20 times.
4) Upon completion of the 20 replications, a convergence assessment was performed. Individual parameters that did not show convergence (e.g., $\hat{R}$>1.2) were excluded from subsequent parameter recovery evaluation.
5) Indices of parameter recovery (correlations between estimated and true values, RMSEs, absolute biases, posterior standard deviations) were computed and evaluated for each experimental condition. Overall item and test information (OII and OTI) values were averaged across replications and reported. The results for the 30-Pair and 90-Pair tests were compared to the results for the 30-Triplet test.

Hypotheses

The following hypotheses were proposed.

1) Parameters will be estimated more accurately in the large sample (N=500) conditions than in the small sample (N=250) conditions across MFC pair tests, as indicated by lower absolute bias, RMSE and PSD statistics.
2) Parameters will be estimated more accurately with the 90-Pair test than with the 30-Pair test, because each dimension appears more times in the longer measure (6 times vs. 18 times, respectively). Accuracy will be comparable (no significant difference) for the 30-Triplet and 90-Pair tests.

3) Test information will be higher for the 90-Pair test than the 30-Pair test. Information will be comparable (no significant difference) for the 30-Triplet and 90-Pair tests.

Using SPSS version 22, these hypotheses were tested using MANOVA with sample size and MFC test type as the between subject factors, and the absolute bias, RMSEs, PSDs, and test information index as dependent variables. Partial eta squared (η_p^2) was used to evaluate effect size in the MANOVA and eta squared (η^2) in the follow-up one-way ANOVA. Values of .01, .06, and .14 represent small, medium, and large effects, respectively (Cohen, 1998). Additionally, parameter recovery and information plots were visually inspected1.

Study 3

Study1 and Study 2 examined statement and person parameter recovery and information values using GGUM-RANK MFC triplet and pair measures. Although these simulation studies provided insights into MFC test construction practices, they provided no evidence concerning the comparability of GGUM-RANK MFC and Likert-type CTT scores with real examinees. To address that limitation, a small validity study was conducted. MFC and Likert-type Big Five personality measures were constructed and administered to online research participants along with some criterion measures to examine convergent and criterion-related validity.

Participants

The sample consisted of 495 college students within the United States. Data were collected via an Amazon’s Mechanical Turk online survey, which included MFC and Likert-type measures of Big Five (Goldberg, 1992) personality factors, as well as measures of life satisfaction, positive and negative affect, aggression, and RIASEC vocational interests (Holland, 1985). Each respondent was paid $.75 for participation. The average age of the participants was 25.3 years (*SD* = 4.7). The sample was 47.4% males. 61.1 % were Caucasians, 16.6 % were African Americans, 10.1% were Hispanics, 5.7% were Asians, and 6.3% selected “Other” as their ethnicity.

Measures

Single-Statement (SS) Personality Measure. A single-statement (Likert-type, ordinal response) Big Five personality measure was created by sampling 60 statements from Goldberg’s (1992) International Personality Item Pool (IPIP; Goldberg, 1992). 12 statements were selected to measure each of the five factors (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism). In total, 18 of the statements were negatively worded and 42 were positively worded. The measure was administered using a 5-point Likert-type format (i.e., 1 = *strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree*). Negatively worded statements were reverse scored, and total scores for each personality factor were obtained by summing the respective item scores. The 60-item SS personality measure is displayed in Appendix D1.

MFC Personality Measure. A 20-triplet MFC measure was created by organizing the statements composing the 60-item Likert-type personality measure into blocks of three, with each statement in a block representing a different personality trait and each block containing a mix of positively and negatively worded statements. Respondents were instructed to rank the statements in each block from 1= most like you to 3 = least like you. This MFC Big Five measure is shown in Appendix D2.

Criterion Measures. Several measures were used to examine convergent and criterion validity of the personality measures. The 5-item, 7-point *Satisfaction with Life Scale* (SWLS, Diener, Emmons, Larsen, & Griffin, 1985) was used to assess satisfaction with life. Positive and negative affect were measured using the 12-item, 7-point *Scale of Positive and Negative Experience* (SAPNE, Diener et al., 2010). Aggression was measured using the 12-item, 5-point Buss-Perry Aggression Questionnaire (Bryant & Smith, 2001). Lastly, RIASEC (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional) vocational interests were measured with 10-item 5-point subscales of the O*NET Interest Profiler (Rounds, Su, Lewis, & Rivkin, 2010). The criterion measures for this study are shown in Appendices D3 to D6.

Analytic Approach

The SS personality measure was scored using the conventional summative approach. To score the MFC triplet measure, the GGUM-RANK model was fitted to triplet rank response data using the Ox program developed for this dissertation. Initial parameter values and four-parameter beta priors {1.5, 1.5, .25, 4}, {2, 2, -4, 4} and {2, 2, -3, 1} were specified for α, δ, and τ estimation. 40,000 iterations with three Markov chains were performed, and the first 20,000 iterations from each chain were discarded as burn-in. Convergence of the MCMC algorithm was evaluated using the Gelman-Rubin diagnostic index ($\hat{R}$). Estimated item parameter and item information and test information indices were recorded.

The reliabilities of the SS personality and criterion measures were computed using coefficient alpha. The reliability of the MFC triplet measure was calculated using the marginal reliability equation provided by Brown & Croudace (2015):

$$\bar{\rho} = 1 - \frac{\bar{\sigma}_e^2}{\sigma_{\hat{\theta}}^2} ,$$

where $\sigma_{\hat{\theta}}^{2}$ = variance of estimated trait scores = $\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_J - \bar{\hat{\theta}}_J)^2$,

$\bar{\sigma}_e^2$ = average squared posterior standard deviation of estimate trait scores =

$\frac{1}{N}\sum_{j=1}^{N} PSD^2(\hat{\theta}_J)$.

The correspondence between SS and MFC personality scores was examined via a multi-trait multi-method (MTMM) analysis and by comparing criterion-related validity coefficients. To assess convergent validity, monotrait-heteromethod correlations were examined. To assess discriminant validity, heterotrait-monomethod correlations were examined. Predictive validity was assessed by comparing the pattern of correlations of the SS and MFC personality scores with the criterion variables.

CHAPTER THREE:

RESULTS

Study1

Table 5 presents average convergence rates across the experimental conditions. In general, convergence occurred within 30,000 iterations, with convergence rates ranging from .93 to 1.00. (Individual parameters that did not achieve convergence (e.g., $\hat{R}$ > 1.2) were excluded from parameter recovery index calculations.)

Table 5. Average Convergence Rates across the Experimental Conditions.

Test Length	Sample Size	Intrablock Discrimination	Intrablock Location SD	Alpha	Delta	Tau	Average
30-Triplet	250	High	Large	1.00	.97	.98	.98
			Small	1.00	.99	.97	.98
		Small	Large	.99	.93	.97	.96
			Small	1.00	.98	.99	.99
	500	High	Large	1.00	.98	.87	.95
			Small	1.00	.97	.91	.96
		Small	Large	.91	.89	.97	.93
			Small	.87	.90	.97	.91
60-Triplet	250	High	Large	1.00	.98	.98	.99
			Small	1.00	.98	.97	.98
		Small	Large	1.00	1.00	1.00	1.00
			Small	1.00	1.00	1.00	1.00
	500	High	Large	1.00	.98	.88	.95
			Small	1.00	.98	.88	.95
		Small	Large	1.00	1.00	.98	.99
			Small	1.00	1.00	.98	.99

Table 6 presents the parameter recovery results for GGUM-RANK statement parameter estimation, averaged over replications. Across all conditions, absolute bias (ABS) ranged from .13 to .23, .11 to .27, and .16 to .21 for α, δ, and τ, respectively. The corresponding root mean squared errors (RMSE) ranged from .16 to .31, .14 to .35, and .20 to .26. The correlations (CORR) between true and estimated δ parameters ranged from .96 to .99, but were lower for τ and quite low for α in some conditions. As expected, recovery of α and δ parameters improved with test length (60-Triplet better than 30-Triplet), sample size (500 better than 250), and intrablock discrimination level (high better low), although the pattern of results for τ was inconsistent. Surprisingly, intrablock location variability did not seem to influence the recovery results. It can be seen that ABS, RMSE, PSD, and CORR values are nearly the same across the small and large intrablock location SD conditions.

Table 6. Statement Parameter Recovery across the Experimental Conditions for Study 1.

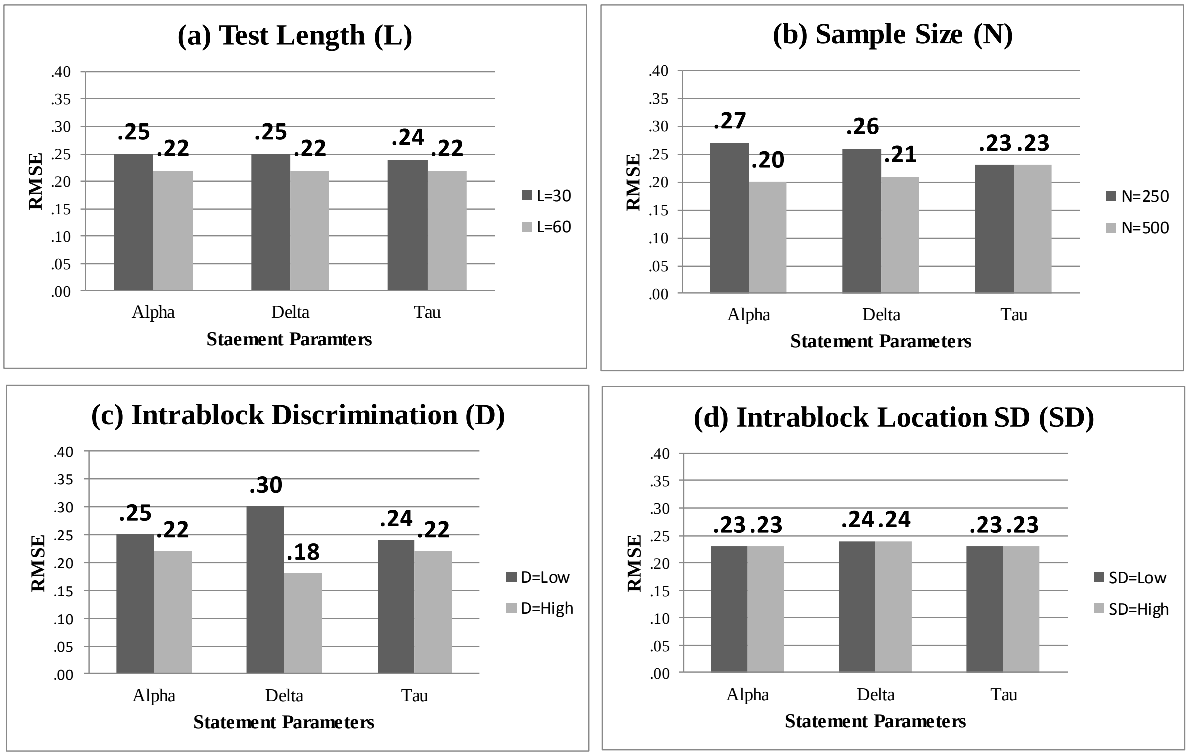
Test Length	Sample Size	Intrabock Discrimination	Intrablock Location SD	Recovery Statistics	Alpha	Delta	Tau
30-Triplet	250	High	Large	ABS	.22	.17	.18
				RMSE	.28	.22	.22
				PSD	.31	.22	.51
				CORR	.37	.99	.80
			Small	ABS	.21	.17	.19
				RMSE	.27	.21	.23
				PSD	.30	.23	.51
				CORR	.41	.99	.78
		Low	Large	ABS	.23	.27	.21
				RMSE	.30	.34	.26
				PSD	.29	.42	.53
				CORR	.46	.96	.69
			Small	ABS	.23	.27	.20
				RMSE	.31	.35	.25
				PSD	.29	.42	.53
				CORR	.44	.96	.69
	500	High	Large	ABS	.16	.12	.19
				RMSE	.20	.16	.23
				PSD	.21	.16	.51
				CORR	.46	.99	.79
			Small	ABS	.16	.13	.20
				RMSE	.20	.16	.23
				PSD	.21	.17	.50
				CORR	.54	.99	.80
		Low	Large	ABS	.18	.23	.20
				RMSE	.23	.29	.25
				PSD	.24	.36	.52
				CORR	.61	.97	.74
			Small	ABS	.18	.23	.19
				RMSE	.22	.30	.24
				PSD	.23	.36	.53
				CORR	.54	.97	.71

Table 6 (Continued)

Test Length	Sample Size	Intrabock Discrimination	Intrablock Location SD	Recovery Statistics	Alpha	Delta	Tau
60-Triplet	250	High	Large	ABS	.19	.15	.17
				RMSE	.24	.19	.20
				PSD	.23	.19	.52
				CORR	.42	.99	.82
			Small	ABS	.19	.16	.16
				RMSE	.25	.20	.20
				PSD	.23	.20	.51
				CORR	.42	.99	.81
		Low	Large	ABS	.21	.24	.19
				RMSE	.26	.30	.24
				PSD	.27	.35	.52
				CORR	.58	.97	.72
			Small	ABS	.20	.23	.20
				RMSE	.25	.30	.24
				PSD	.27	.35	.52
				CORR	.58	.97	.74
	500	High	Large	ABS	.13	.11	.18
				RMSE	.16	.14	.21
				PSD	.16	.13	.51
				CORR	.53	.99	.82
			Small	ABS	.13	.11	.19
				RMSE	.16	.14	.22
				PSD	.16	.14	.51
				CORR	.55	.99	.80
		Low	Large	ABS	.16	.19	.19
				RMSE	.20	.24	.23
				PSD	.20	.26	.52
				CORR	.69	.98	.76
			Small	ABS	.15	.19	.19
				RMSE	.19	.25	.23
				PSD	.20	.26	.52
				CORR	.68	.98	.74

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR = correlation between true and estimated parameters.

To more clearly see how RMSE varied as a function of test length, sample size, intrablock discrimination, and intrablock location SD, these results are presented using bar graphs in Figure 4.


(a) Test Length (L)

(b) Sample Size (N)

(c) Intrablock Discrimination (D)

(d) Intrablock Location SD (SD)

Figure 4. Average RMSEs of statement parameters across simulation conditions.

Table 7 presents parameter recovery statistics for latent trait scores (θ), averaged over replications. The averages across dimensions, Dim1 – Dim10, are shown in the last column. In general, parameter recovery improved as test length and intrablock discrimination increased, as these factors are integrally connected to test information. However, there was no noteworthy improvement with larger samples or location parameter SDs.

Table 7. Person Parameter Recovery Statistics across the Experimental Conditions for MFC Triplet Measure.

Test Length	Sample Size	Intrablock Discrimination	Intrablock Location SD	Recovery Statistics	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Average
30-Triplet	250	High	Large	ABS	.34	.34	.32	.32	.32	.35	.32	.32	.33	.34	.33
				RMSE	.45	.48	.43	.42	.44	.49	.42	.43	.44	.47	.45
				PSD	.43	.42	.41	.41	.41	.43	.40	.41	.42	.42	.43
				CORR	.89	.88	.91	.91	.90	.88	.91	.90	.90	.88	.90
			Small	ABS	.33	.33	.34	.34	.34	.33	.33	.34	.34	.33	.34
				RMSE	.45	.44	.44	.44	.45	.43	.44	.49	.45	.44	.45
				PSD	.42	.41	.42	.43	.42	.41	.42	.43	.43	.42	.42
				CORR	.89	.90	.90	.90	.89	.91	.90	.87	.89	.90	.90
		Low	Large	ABS	.49	.50	.50	.49	.49	.50	.48	.49	.49	.55	.50
				RMSE	.64	.65	.64	.65	.65	.66	.62	.63	.65	.74	.65
				PSD	.58	.58	.58	.58	.59	.60	.56	.56	.58	.65	.59
				CORR	.76	.76	.76	.76	.75	.74	.79	.78	.77	.68	.75
			Small	ABS	.49	.51	.51	.50	.51	.48	.50	.50	.53	.55	.51
				RMSE	.64	.66	.65	.65	.67	.62	.66	.66	.70	.71	.66
				PSD	.60	.57	.57	.57	.61	.61	.58	.54	.56	.81	.60
				CORR	.78	.75	.76	.76	.75	.78	.75	.75	.71	.70	.75
	500	High	Large	ABS	.32	.33	.31	.32	.32	.35	.31	.32	.32	.33	.32
				RMSE	.43	.45	.40	.43	.45	.49	.41	.42	.43	.45	.44
				PSD	.41	.41	.39	.40	.40	.42	.39	.40	.41	.41	.40
				CORR	.90	.90	.92	.90	.90	.87	.91	.91	.90	.90	.90
			Small	ABS	.32	.32	.33	.33	.33	.32	.33	.33	.34	.33	.33
				RMSE	.44	.43	.43	.43	.44	.44	.44	.46	.45	.43	.44
				PSD	.41	.41	.42	.42	.42	.40	.41	.42	.42	.41	.41
				CORR	.90	.90	.90	.90	.90	.90	.90	.89	.89	.90	.90
		Low	Large	ABS	.48	.49	.48	.48	.49	.49	.47	.47	.48	.55	.49
				RMSE	.63	.64	.62	.64	.66	.64	.61	.62	.64	.72	.64
				PSD	.59	.61	.59	.59	.61	.61	.57	.58	.60	.67	.60
				CORR	.77	.77	.78	.77	.75	.76	.79	.79	.76	.68	.76
			Small	ABS	.49	.50	.50	.49	.50	.47	.50	.50	.54	.53	.50
				RMSE	.64	.65	.65	.64	.66	.62	.67	.66	.70	.69	.66
				PSD	.59	.60	.59	.59	.61	.58	.61	.61	.65	.65	.61
				CORR	.77	.76	.77	.77	.75	.78	.74	.76	.71	.72	.75

Table 7 (Continued)

Test Length	Sample Size	Intrablock Discrimination	Intrablock Location SD	Recovery Statistics	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Average
60-Triplet	250	High	Large	ABS	.24	.23	.22	.22	.24	.26	.23	.22	.23	.23	.23
				RMSE	.33	.33	.30	.29	.35	.39	.31	.29	.29	.31	.32
				PSD	.31	.29	.28	.29	.29	.32	.29	.29	.28	.29	.31
				CORR	.94	.94	.96	.96	.94	.92	.95	.96	.96	.95	.95
			Small	ABS	.24	.24	.24	.25	.25	.25	.23	.26	.24	.25	.24
				RMSE	.31	.30	.31	.31	.34	.32	.30	.38	.32	.34	.32
				PSD	.30	.30	.29	.31	.30	.31	.30	.31	.30	.30	.30
				CORR	.95	.96	.95	.95	.94	.95	.95	.92	.95	.94	.95
		Low	Large	ABS	.39	.38	.36	.36	.38	.39	.38	.35	.35	.39	.37
				RMSE	.49	.49	.45	.45	.49	.50	.49	.45	.44	.50	.48
				PSD	.44	.44	.42	.42	.43	.45	.45	.41	.40	.44	.43
				CORR	.87	.87	.88	.89	.87	.85	.87	.89	.89	.86	.87
			Small	ABS	.37	.38	.37	.37	.38	.39	.40	.39	.37	.40	.38
				RMSE	.47	.47	.47	.48	.49	.49	.52	.50	.50	.50	.49
				PSD	.43	.44	.44	.43	.45	.44	.46	.45	.43	.47	.44
				CORR	.88	.87	.88	.88	.87	.87	.85	.85	.86	.85	.87
	500	High	Large	ABS	.24	.23	.22	.22	.23	.25	.22	.22	.22	.23	.23
				RMSE	.32	.34	.30	.29	.33	.42	.30	.29	.28	.30	.32
				PSD	.30	.28	.27	.28	.28	.30	.28	.28	.28	.29	.28
				CORR	.95	.94	.95	.96	.94	.91	.96	.96	.96	.95	.95
			Small	ABS	.23	.23	.24	.24	.23	.24	.23	.24	.24	.23	.24
				RMSE	.31	.30	.32	.31	.32	.32	.30	.33	.31	.30	.31
				PSD	.29	.29	.29	.30	.29	.30	.29	.30	.29	.29	.29
				CORR	.95	.95	.95	.95	.95	.95	.96	.94	.95	.95	.95
		Low	Large	ABS	.37	.37	.36	.35	.36	.39	.38	.35	.34	.38	.36
				RMSE	.49	.48	.46	.45	.47	.50	.50	.46	.43	.51	.47
				PSD	.46	.45	.42	.42	.44	.46	.45	.42	.41	.46	.44
				CORR	.87	.88	.89	.90	.88	.87	.87	.89	.90	.86	.88
			Small	ABS	.37	.38	.37	.36	.36	.38	.39	.38	.37	.39	.37
				RMSE	.48	.49	.47	.47	.47	.49	.51	.50	.48	.51	.49
				PSD	.45	.47	.42	.40	.44	.47	.45	.43	.41	.44	.44
				CORR	.88	.88	.88	.88	.88	.87	.86	.86	.88	.86	.87

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR = correlation between true and estimated parameters.

In the most favorable conditions, for example, the 60-Triplet, High Intrablock Discrimination conditions, ABS ranged from .23 to .24, RMSE ranged from .31 to .32, PSD ranged from .28 to .31, and Average CORRs were .95. In the 30-Triplet High Intrablock Discrimination conditions, Average CORRs were still good (0.9), but the estimation errors were larger: ABS ranged from .32 to .34, RMSE ranged from .44 to .45, and PSD ranged from .40 to .43. In the least favorable conditions, for example, 30-Triplet, Low Intrablock Discrimination conditions, the worst results were observed: ABS ranged from .49 to .51, RMSE ranged from .64 to .66, PSD ranged from .59 to .61, and Average CORR ranged from .75 to .76. For quick comparisons across conditions, the RMSE and Average CORR results are presented graphically in Figures 5 and 6, respectively.

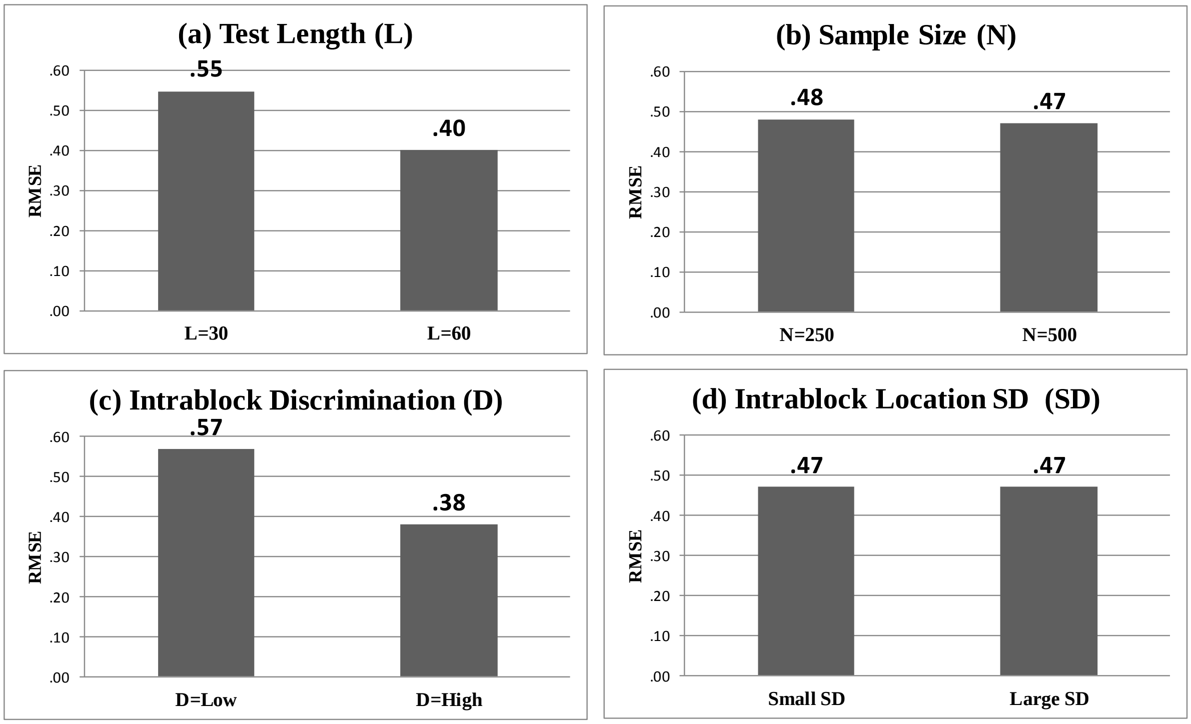


Figure 5. Average RMSEs of person parameters across simulation conditions.

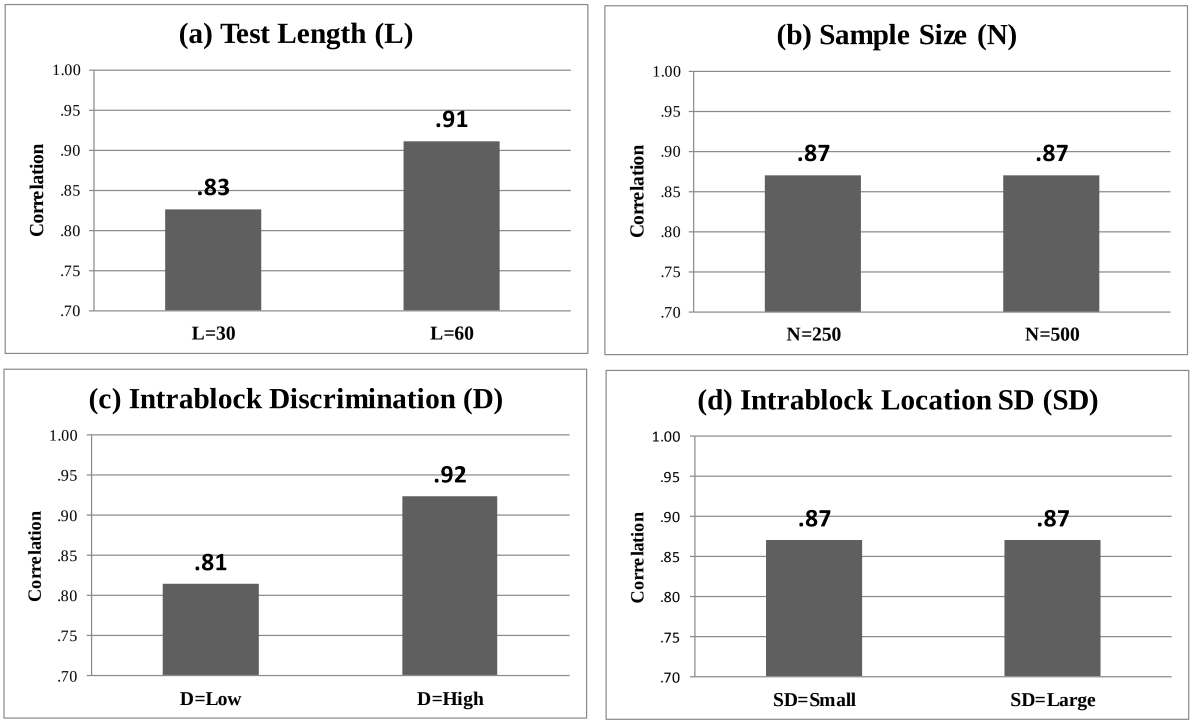


Figure 6. Average correlations between true and estimated person parameters across simulation condition.

To shed additional light on the person parameter recovery results, Table 8 presents the average overall item information (OII) and overall test information (OTI) values for the MFC triplet measure in each experimental condition. Recall that OTI is the sum of the OII values. The results clearly show that the OII was driven by intrablock discrimination, which is consistent with conventional item response theory findings. In the High Intrablock Discrimination conditions, Average OII was approximately 2.5 times greater than in the Low Intrablock Discrimination conditions. Also, consistent with the test design, OTI was about twice as large in the corresponding 60-Triplet vs. 30-Triplet conditions. These findings are echoed in Figures 7 and 8, which present the average OII and OTI values visually to facilitate comparisons with the RMSE and CORR results in Figures 4 – 6.

Table 8. Average OII and OTI across the Experimental Conditions.

Test Length	Sample Size	Intrablock Discrimination	Intrablock Location SD	Average OII	OTI
30-Triplet	250	High	Large	2.50	75.07
			Small	2.64	79.06
		Low	Large	1.05	31.64
			Small	1.09	32.59
	500	High	Large	2.57	77.15
			Small	2.70	80.88
		Low	Large	0.91	27.44
			Small	0.94	28.11
60-Triplet	250	High	Large	2.48	148.77
			Small	2.62	157.05
		Low	Large	1.00	60.09
			Small	1.06	63.33
	500	High	Large	2.54	152.53
			Small	2.66	159.48
		Low	Large	0.92	55.01
			Small	0.94	56.43

Note. OII = Overall Item Information; OTI = Overall Test Information.

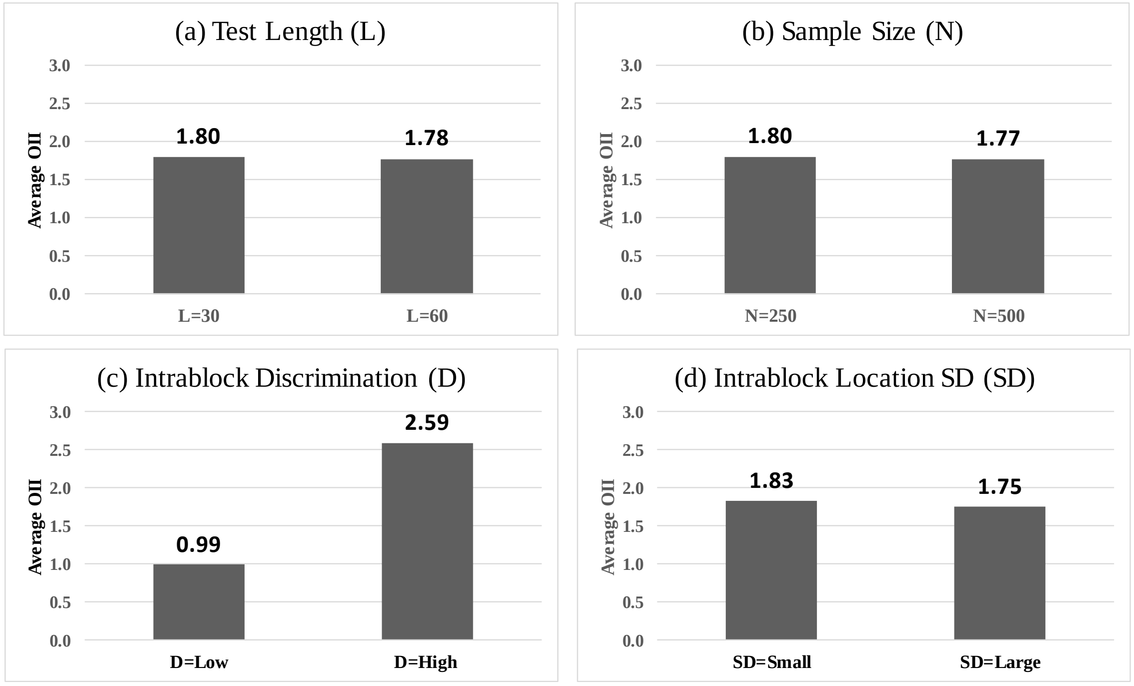


Figure 7. Average overall item information across simulation conditions.

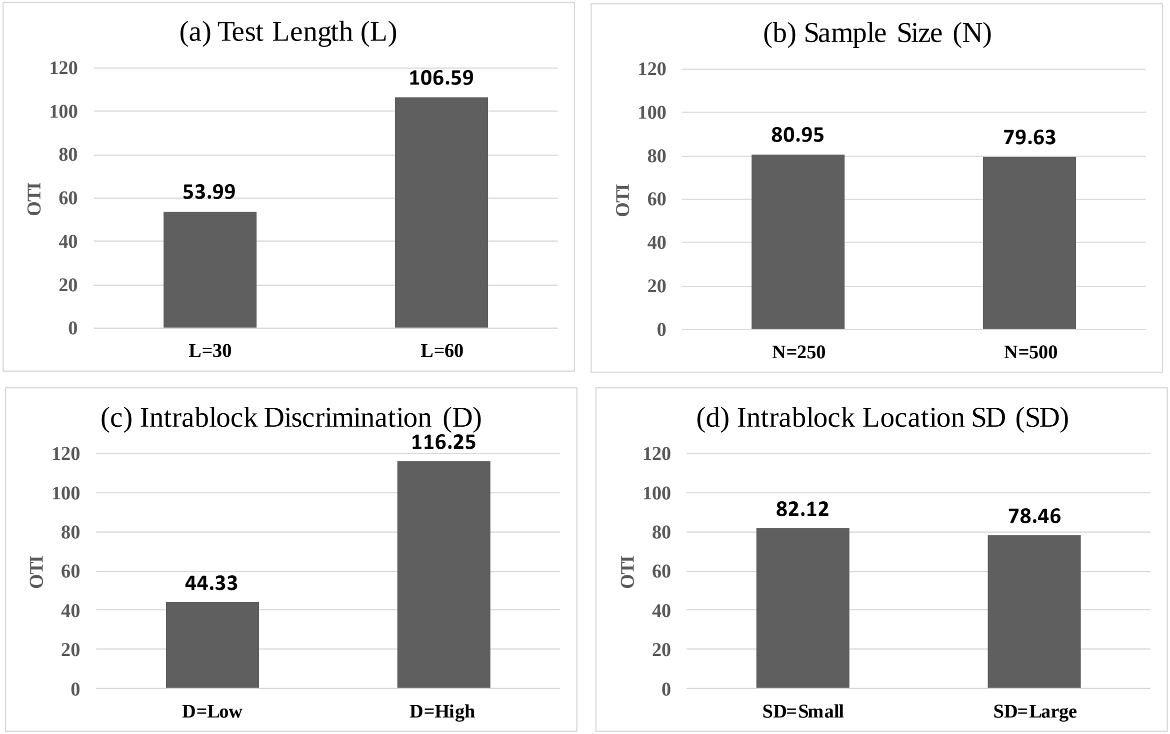


Figure 8. Overall test information across simulation conditions.

Statistical Significance Tests of Study 1 Hypotheses.

To buttress the interpretation of the parameter recovery results shown in Table 6 and 7, and to address the specific hypotheses that were proposed, MANOVAs were conducted separately for statement and person parameters using ABS, RMSE, PSD, and Average CORR as dependent variables. Hypothesis 1 posited that parameters would be estimated more accurately in the larger sample size (N=500) conditions than in the small sample (N=250) conditions. Hypothesis 2 posited that parameters would be estimated more accurately with longer (60-Triplet) tests than with shorter (30-Triplet) tests. Table 9 presents the multivariate test results indicating that Sample Size had a statistically significant effect (p < .05) on statement parameter recovery, and Test Length had a statistically significant effect (p < .05) on person parameter recovery.

Table 9. Multivariate Tests of Between Subjects Effects for Study 1 Hypotheses 1 and 2.

Hypothesis	Effect	Parameter	Wilks' Lambda	F	Hypothesis df	Error df	Sig	η_p^2
Hypothesis 1	Sample Size	Statement Parameter	.09	6.83	9.00	6.00	**.01**	.91
		Person Parameter	.95	.16	4.00	11.00	.95	.06
Hypothesis 2	Test Length	Statement Parameter	.22	2.35	9.00	6.00	.15	.78
		Person Parameter	.35	5.06	4.00	11.00	**.02**	.65

As a follow-up, ANOVAs were conducted on the parameter recovery indices for statement and person parameters to see where the significant differences lie. These results are presented in Table 10, which shows that Sample Size had statistically significant effects on α

ABS, α RMSE, and α PSD, with large effect sizes (η^2) of .73, .72, and .63, respectively. However, Sample Size had no statistically significant effects on parameter recovery indices for δ or τ, most likely because these parameters were estimated fairly well in the N=250 conditions (see Table 6). Also note that Sample Size did not have a statistically significant effect on the indices of person parameter (θ) recovery. This finding is consistent with previous IRT parameter recovery studies (e.g., Reise & Yu, 1990), although the hypotheses were framed broadly to allow for the possibility that factors typically influencing either statement *or* person parameter estimation could have reciprocal effects due to joint estimation in this MCMC algorithm. Thus, Hypothesis 1 was partially supported.

Table10 also shows that Test Length had a statistically significant effect on person parameter (θ) recovery statistics, and effect sizes were quite large: η^2= .35, .39, .42, and .34 for ABS, RMSE, PSD, and Average CORR, respectively. On the other hand, Test Length had little influence on the recovery of statement parameters. Therefore, Hypothesis 2 was also partially supported.

Table 10. Univariate Tests of Between Subjects Effects for Study 1 Hypotheses 1 and 2.

Hypothesis	Effect	Parameter	Dependent Variable	SS	df	Mean Square	F	Sig.	η^2
Hypothesis 1	Sample Size	Statement Parameter	α ABS	.01	1.00	.01	36.88	**.00**	.73
			δ ABS	.01	1.00	.01	2.92	.11	.17
			τ ABS	.00	1.00	.00	.35	.56	.03
			α RMSE	.02	1.00	.02	35.80	**.00**	.72
			δ RMSE	.01	1.00	.01	2.61	.13	.16
			τ RMSE	.00	1.00	.00	.00	1.00	.00
			α PSD	.02	1.00	.02	23.98	**.00**	.63
			δ PSD	.02	1.00	.02	1.97	.18	.12
			τ PSD	.00	1.00	.00	.72	.41	.05
Hypothesis 2	Test Length	Person Parameter	θ ABS	.05	1.00	.05	7.46	**.02**	.35
			θ RMSE	.09	1.00	.09	8.77	**.01**	.39
			θ PSD	.08	1.00	.08	10.13	**.01**	.42
			θ CORR	.03	1.00	.03	7.27	**.02**	.34

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR= correlations between true and estimated parameters.

Table 11 shows multivariate test results for Hypotheses 3 and 4. Hypothesis 3 posited that parameters would be estimated more accurately with tests having high (vs. low) intrablock discrimination, and Hypothesis 4 posited that parameters would be estimated more accurately with tests having large (vs. small) intrablock location SD. As above, these hypotheses were tested separately with statement and person parameter ABS, RMSE, PSD and Average CORR as dependent variables using MANOVAs followed by univariate ANOVAs. The result shows that Intrablock Discrimination had a statistically significant effect on indices of statement and person parameter recovery ($p < .05$), but Intrablock Location SD did not ($p = .90$ and .68, respectively).

Table 11. Multivariate Tests of Between Subjects Effects for Study 1 Hypotheses 3 and 4.

Hypothesis	Effect	Parameter	Wilks' Lambda	F	Hypothesis df	Error df	Sig	$\eta_p{}^2$
Hypothesis 3	Intrablock Discrimination	Statement Parameter	.01	61.65	9.00	6.00	**.00**	.99
		Person Parameter	.23	9.11	4.00	11.00	**.00**	.77
Hypothesis 4	Intrablock Location SD	Statement Parameter	.63	.39	9.00	6.00	.90	.37
		Person Parameter	.83	.58	4.00	11.00	.68	.17

Since the multivariate tests for Hypothesis 3 were significant, follow-up univariate tests were conducted to examine the effects of Intrablock Discrimination on individual parameter recovery statistics. Table 12 shows that Intrablock Discrimination had a statistically significant effect on recovery statistics for δ, τ, and θ, but not α. The effect sizes for $\delta, \tau,$ and θ recovery were large, ranging from .33 to .77, whereas they ranged from just .06 to .09 for α. Thus, Hypothesis 3 was partially supported. Finally, Intrablock Location SD had no statistically significant effects on parameter recovery. Therefore Hypothesis 4 was not supported.

Table 12. Univariate Tests of Between Subjects Effects for Study 1 Hypotheses 3.

Hypothesis	Effect	Parameter	Dependent Variable	SS	df	Mean Square	F	Sig.	η^2
Hypothesis 3	Intrablock Discrimination	Statement Parameter	α ABS	.00	1.00	.00	1.35	.26	.09
			δ ABS	.03	1.00	.03	42.06	**.00**	.75
			τ ABS	.00	1.00	.00	6.89	**.02**	.33
			α RMSE	.00	1.00	.00	1.22	.29	.08
			δ RMSE	.06	1.00	.06	46.08	**.00**	.77
			τ RMSE	.00	1.00	.00	18.42	**.00**	.57
			α PSD	.00	1.00	.00	.91	.36	.06
			δ PSD	.11	1.00	.11	44.45	**.00**	.76
			τ PSD	.00	1.00	.00	27.32	**.00**	.66
		Person Parameter	θ ABS	.09	1.00	.09	24.78	**.00**	.64
			θ RMSE	.14	1.00	.14	21.34	**.00**	.60
			θ PSD	.11	1.00	.11	18.12	**.00**	.56
			θ CORR	.03	1.00	.03	7.27	**.02**	.34

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR= correlations between true and estimated parameters.

Study2

Table 13 presents average convergence rates for Study 2. As in Study 1, overall convergence approached 100% within 30,000 iterations, and individual parameters that did not converged (e.g., $\hat{R}$ > 1.2) were excluded from the computation of parameter recovery statistics.

Table 13. Average Convergence Rates across the Experimental Conditions.

Test Length	Sample	Alpha	Delta	Tau	Average
30-Pair	250	.99	.91	1.00	.97
	500	1.00	1.00	1.00	1.00
90-Pair	250	.91	.91	1.00	.94
	500	1.00	.99	1.00	1.00

Table 14 presents the 30-Pair and 90-Pair parameter recovery results. (Findings for the 30-Triplet measure in Study 1 are presented at the top of the table for easy comparison.) The main finding is that the 30-Triplet measure exhibited better recovery statistics than the 90-Pair measure, so there is a distinct advantage in using a shorter triplet measure over a much longer pairwise preference measure for statement calibration. In the corresponding N=250 and N=500 conditions, the 30-Triplet measure had higher CORR and lower ABS, RMSE, and PSD values.

Next, and of somewhat lesser importance, is the comparison of recovery statistics for the 30-Pair and 90-Pair tests. As was the case with the triplet measures, the best results were found for δ, with CORR near 1 in all conditions and ABS and RMSE below .3 in the N=500 conditions. Results for α and τ were not as good. For τ, CORR ranged from just .56 to .71 and PSD values exceeded .6. For α, CORR ranged from just .13 to .38, due in part to the restricted range of discrimination parameters, but PSD ranged from a low of .29 (90-Pair, N=500) to a high of .48 (30-Pair, N=250).

Table 14. Statement Parameter Recovery across the Experimental Conditions.

Test Length	Sample Size	Recovery Statistics	Alpha	Delta	Tau
30-Triplet in Study1	250	ABS	.22	.17	.18
		RMSE	.28	.22	.22
		PSD	.31	.22	.51
		CORR	.37	.99	.80
	500	ABS	.16	.12	.19
		RMSE	.20	.16	.23
		PSD	.21	.16	.51
		CORR	.46	.99	.79
30-Pair	250	ABS	.31	.30	.24
		RMSE	.39	.36	.28
		PSD	.48	.46	.63
		CORR	.13	.97	.56
	500	ABS	.25	.21	.20
		RMSE	.36	.27	.24
		PSD	.42	.37	.62
		CORR	.29	.98	.62
90-Pair	250	ABS	.32	.28	.17
		RMSE	.38	.33	.20
		PSD	.37	.28	.62
		CORR	.24	.98	.69
	500	ABS	.21	.20	.13
		RMSE	.32	.24	.16
		PSD	.29	.20	.61
		CORR	.38	.99	.71

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR =correlation between true and estimated parameter.

Table 15 presents the θ parameter recovery results averaged over replications. As in Study 1, θ recovery results were highly similar across dimensions, and sample size had little to no effect on estimation accuracy. As expected, θs were estimated better with 90-Pair tests than 30-Pair tests. And, perhaps most importantly, 90 pairs were needed to achieve similar levels of ABS, RMSE, PSD, and CORR to the 30-Triplet test. Specifically, with N=500, ABS, RMSE, PSD and CORR were .32, .44, .40 and .90 for the 30-Triplet test and .30, .40, .38 and .92 for the 90-Pair test, respectively.

Table 15. Person Parameter Recovery Statistics across the Experimental Conditions.

Test Length	Sample Size	Recovery Statistics	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Average
30-Triplet in Study1	250	ABS	.34	.34	.32	.32	.32	.35	.32	.32	.33	.34	.33
		RMSE	.45	.48	.43	.42	.44	.49	.42	.43	.44	.47	.45
		PSD	.43	.42	.41	.41	.41	.43	.40	.41	.42	.42	.43
		CORR	.89	.88	.91	.91	.90	.88	.91	.90	.90	.88	.90
	500	ABS	.32	.33	.31	.32	.32	.35	.31	.32	.32	.33	.32
		RMSE	.43	.45	.40	.43	.45	.49	.41	.42	.43	.45	.44
		PSD	.41	.41	.39	.40	.40	.42	.39	.40	.41	.41	.40
		CORR	.90	.90	.92	.90	.90	.87	.91	.91	.90	.90	.90
30-Pair	250	ABS	.56	.56	.51	.51	.50	.53	.51	.50	.50	.51	.52
		RMSE	.76	.76	.68	.69	.66	.71	.68	.67	.67	.69	.70
		PSD	.71	.69	.66	.65	.63	.66	.65	.63	.64	.63	.65
		CORR	.64	.65	.74	.72	.74	.71	.73	.74	.75	.73	.72
	500	ABS	.55	.54	.51	.51	.50	.51	.51	.50	.49	.50	.51
		RMSE	.74	.74	.67	.67	.66	.68	.71	.66	.66	.70	.69
		PSD	.70	.68	.63	.68	.64	.64	.63	.63	.63	.65	.65
		CORR	.68	.66	.76	.73	.77	.73	.73	.76	.76	.72	.73
90-Pair	250	ABS	.32	.30	.30	.30	.31	.33	.30	.30	.31	.31	.31
		RMSE	.43	.41	.38	.39	.42	.47	.39	.39	.42	.42	.41
		PSD	.41	.39	.39	.38	.40	.40	.39	.38	.39	.41	.39
		CORR	.90	.91	.92	.92	.91	.88	.92	.92	.91	.91	.91
	500	ABS	.31	.30	.29	.29	.30	.30	.30	.30	.30	.31	.30
		RMSE	.42	.40	.37	.38	.41	.42	.39	.39	.41	.45	.40
		PSD	.39	.38	.37	.37	.38	.38	.37	.38	.38	.39	.38
		CORR	.91	.92	.93	.92	.92	.91	.92	.92	.91	.89	.92

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR = correlation between true and estimated parameter.

Table 16 presents the average overall item information (OII) and overall test information (OTI) results, averaged across 20 replications. Not surprisingly, OTI was about three times higher for the 90-Pair test than the 30-Pair test, and MFC triplets provided about 2.8 times the Average OII versus MFC pairs.

Table 16. Average Overall Item Information and Overall Test Information across the Experimental Conditions.

Test Length	Sample Size	Average OII	OTI
30-Triplet in Study1	250	2.50	75.07
	500	2.57	77.15
30-Pair	250	0.86	25.67
	500	0.91	27.32
90-Pair	250	0.92	82.72
	500	0.96	86.37

Note. OII = Overall Item Information; OTI = Overall Test Information.

Finally, to facilitate comparisons across conditions, the parameter recovery results described above are presented graphically in Figures 9 (a-c) and 10. First, the plots show that sample size influenced statement parameter recovery with MFC pair tests, as indicated by the lower ABS, RMSE, and PSD and higher CORR values in the N=500 conditions, but the indirect effect on θ recovery was minimal. Statement parameter recovery was generally worst for α; the results were less consistent for δ and τ. Second, Figure 10 shows a marked difference in θ recovery between the 30-Triplet and 30-Pair tests, regardless of sample size; the most striking difference is the CORR values - .90 for triplets and approximately .70 for pairs. As suggested earlier, about 90 pairs are needed to score as accurately as 30 triplets. This finding has important implications for testing in personnel settings where examinee motivation and fatigue are important concerns.

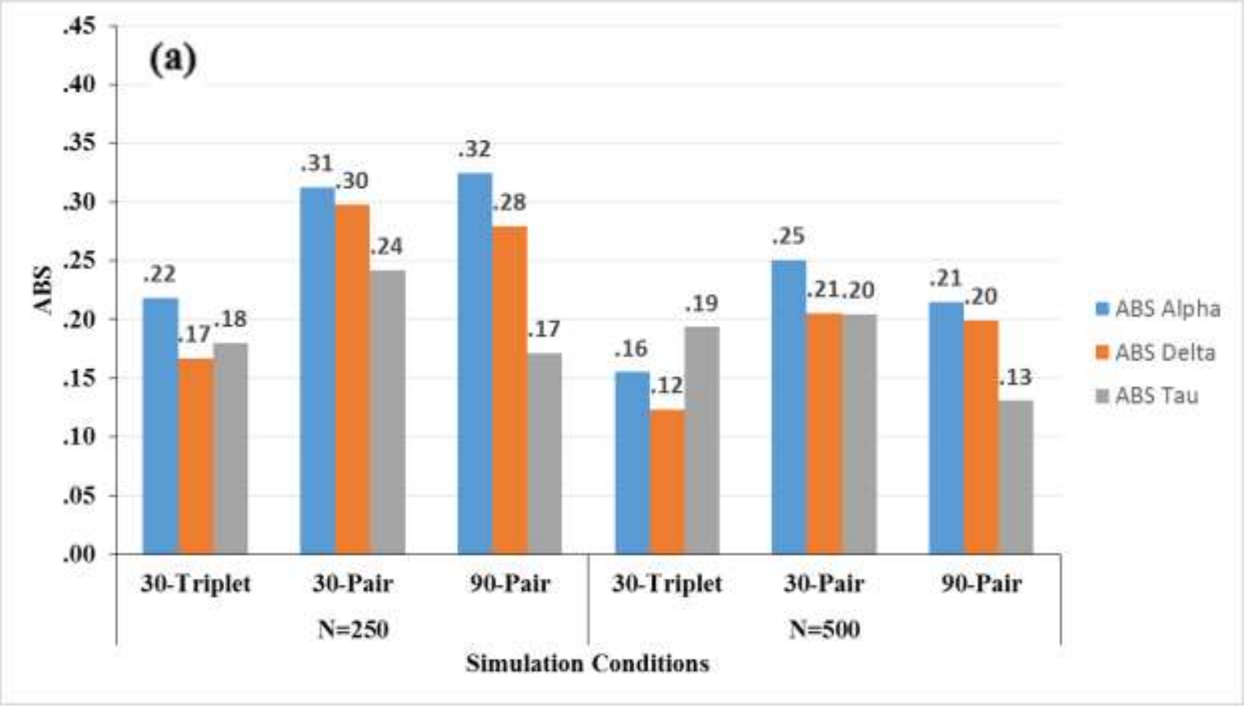


Figure 9a. Absolute biases of item parameters for study 2 hypothesis 1

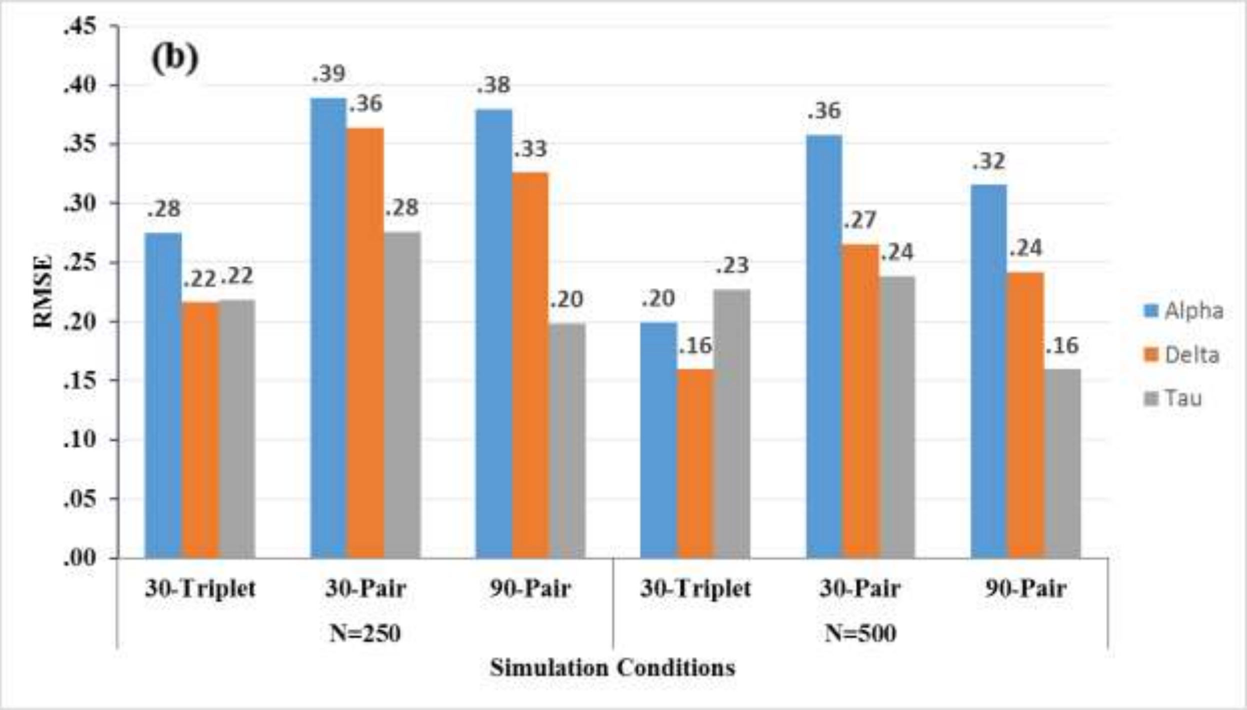


Figure 9b. RMSEs of item parameters for study 2 hypothesis 1

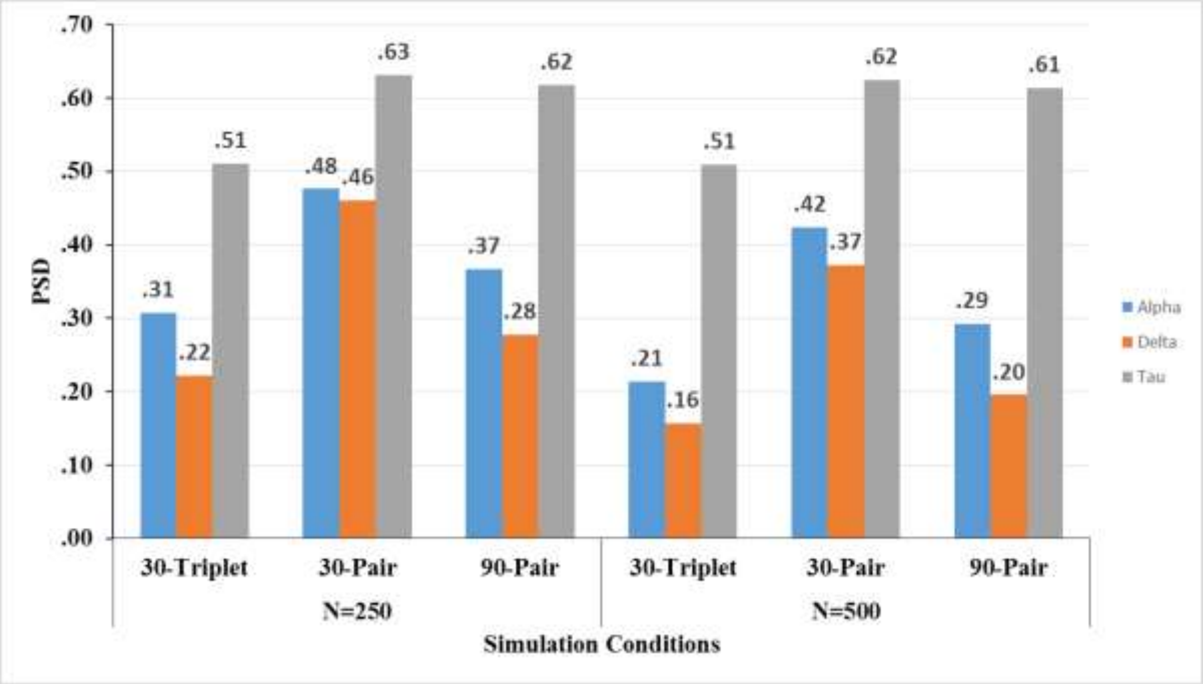


Figure 9c. PSDs of item parameters for study 2 hypothesis 1

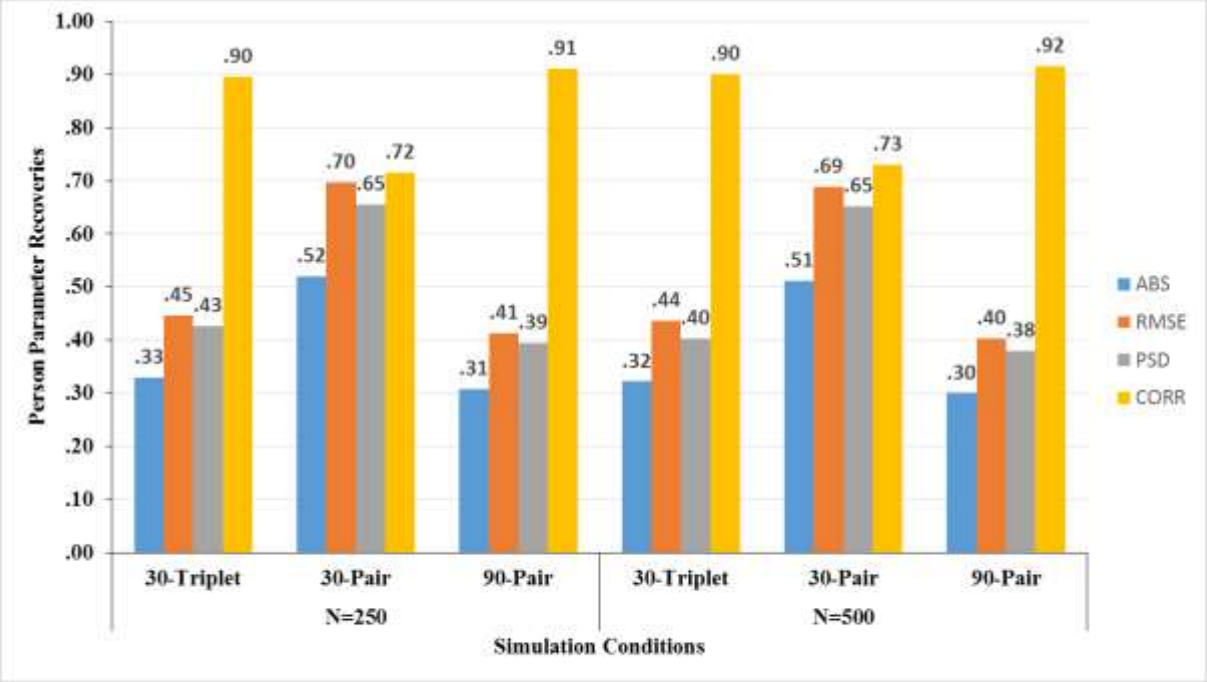


Figure 10. Person parameter recovery for study 2 hypothesis 1

Statistical Significance Tests of Study 2 Hypotheses.

Hypothesis 1 proposed that parameters would be estimated more accurately in the larger sample size (N=500) conditions than in the small sample size (N=250) conditions. This hypothesis was tested by running MANOVAs using the ABS, RMSE, PSD, and CORR values as dependent variables and sample size as between subjects factors for each 30-Pair and 90-Pair tests. Consistent with expectations, the results indicated that Sample Size had a beneficial effect on statement parameter recovery (Wilk's λ =.07, F (9, 30) = 44.06, $p < .05$, $\eta_p{}^2 = .93$ for the 30-Pair test; Wilk's λ =.00, F (9, 30) = 1257.35, $p < .05$, $\eta_p{}^2 = .99$ for the 90-Pair test) but there was no significant effect on person parameter (*θ)* recovery (Wilk's λ =.90, F (4, 35) = .93, $p = .46$, $\eta_p{}^2 = .10$ for the 30-Pair test; Wilk's λ =.95, F (4, 35) = .43, $p = .79$, $\eta_p{}^2 = .05$ for the 90-Pair test). A follow-up univariate test revealed that there was a significant univariate effect of sample size on α ABS ($p < .05$), α RMSE, ($p < .05$), α PSD, ($p < .05$), δ ABS ($p < .05$), δ RMSE ($p < .05$), δ PSD, ($p < .05$) for both 30- and 90- Pair Tests. Therefore, Hypothesis 1 was partially supported.

Hypothesis 2 proposed that parameters would be estimated more accurately with the 90-Pair tests than with the 30-Pair tests, but 30-Triplet and 90-Pair tests would be comparable (no significant difference). Hypothesis 2 was tested using separate MANOVAs for statement and person parameters with Test Type (30-Pair, 90-Pair, 30-Triplet) as a between subject factor and ABS, RMSE, PSD, CORR. The results shown in Table 17 indicate that Test Type had a statistically significant effect ($p < .05$) on person parameter recovery, but not on statement parameter recovery ($p > .05$).

Table 17. Multivariate Tests of Between Subjects Effects for Study 2 Hypothesis 2.

Hypothesis	Effect	Parameter	Wilks' Lambda	F	Hypothesis df	Error df	Sig	η_p^2
Hypothesis 2	Test Type	Statement Parameter	.03	1.76	6.00	2.00	.41	.84
		Person Parameter	.00	12.01	6.00	2.00	**.01**	.97

A follow-up univariate test was conducted for person parameter (θ) recovery to see where the differences lie, and these results are shown in Table 18. There was a significant univariate effect of Test Type on θ ABS ($p < .05$), θ RMSE, ($p < .05$), θ PSD ($p < .05$), and θ CORR ($p < .05$).

Table 18. Univariate Tests of Between Subjects Effects for Study 2 Hypothesis 2.

Hypothesis	Effect	Parameter	Dependent Variable	SS	df	Mean Square	F	Sig.	η^2
Hypothesis 2	Test Type	Person Parameter	θ ABS	.05	2.00	.03	537.33	**.00**	1.00
			θ RMSE	.10	2.00	.05	988.00	**.00**	1.00
			θ PSD	.08	2.00	.04	252.70	**.00**	.99
			θ CORR	.05	2.00	.02	669.50	**.00**	1.00

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR = correlation between true and estimated parameters.

In addition, Bonferroni post-hoc multiple comparison tests were performed to examine pairwise differences as a function of Test Type. Table 19 shows the mean differences and associated p-values. It can be seen that the 90-Pair test provided better results than the 30-Pair test in every case ($p < .05$), and the 30-Triplet test provided better results than the 30-Pair test in every case (*p*

< .05). Also, as expected, there were no significant differences between the 30-Triplet and 90-Pair tests (p > .05). Therefore, Hypothesis 2 was supported.

Table 19. Multiple Comparisons with Test Type for Study 2 Hypothesis 2.

Dependent Variable	Comparison	Mean Difference	Sig.
θ ABS	**30-Triplet - 30-Pair**	-.19	**.00**
	30-Triplet - 90-Pair	.02	.20
	30-Pair - 90-Pair	.21	**.00**
θ RMSE	**30-Triplet - 30-Pair**	-.25	**.00**
	30-Triplet - 90-Pair	.04	.06
	30-Pair - 90-Pair	.29	**.00**
θ PSD	**30-Triplet - 30-Pair**	-.24	**.00**
	30-Triplet - 90-Pair	.03	.31
	30-Pair - 90-Pair	.27	**.00**
θ CORR	**30-Triplet - 30-Pair**	.18	**.00**
	30-Triplet - 90-Pair	-.02	.24
	30-Pair - 90-Pair	-.19	**.00**

Note. Bonferroni Correction was used.

Hypothesis 3 proposed that test information would be higher for the 90-Pair test than the 30-Pair test, but 30-Triplet and 90-Pair tests would be comparable (no significant difference). Hypothesis 3 was tested using one-way ANOVA with three test types as the between subject factors and the test information as dependent variable. A univariate test revealed a significant effect of test type on test information, F (2, 3) = 579.50, p <. 001, η^2 = .99. Also, Bonferroni multiple comparison test was conducted for test information by three test types. Table 20 shows that the 90-Pair test yielded significantly higher test information than the 30-Pair test (p < .05), and the 30-Triplet test yielded significantly higher test information than the 30-Pair test (p < .05).

Also, there was no significant difference between the 30-Triplet and 90-Pair tests, which indicates they have comparable test information. Therefore, hypothesis 3 was supported.

Table 20. Multiple Comparisons with Test Type for Study 2 Hypothesis 3.

Dependent Variable	Comparisons	Mean Difference	Sig.
OTI	**30-Triplet - 30-Pair**	49.62	**.00**
	30-Triplet - 90-Pair	-8.44	.06
	30-Pair - 90-Pair	-58.05	**.00**

Note. OTI = Overall Item Information index; Bonferroni Correction was used.

Although it was not a proposed hypothesis, one-way ANOVA results also revealed a significant effect of test type on average item information, F (2, 6) = 1177.01, p <. 001, η^2= .99. The multiple comparison test result shown in Table 21 also indicates that significantly higher average item information was found for the 30-Triplet test than the 30-Pair and 90-Pair tests.

Table 21. Multiple Comparisons of Three Test Types on Average OII.

Dependent Variable	Comparisons	Mean Difference	Sig.
Average OII	**30-Triplet - 30-Pair**	1.65	**.00**
	30-Triplet - 90-Pair	1.60	**.00**
	30-Pair - 90-Pair	-.05	.75

Note. OII = Overall Item Information index; Bonferroni Correction was used.

Study3

Table 22 presents descriptive statistics for the single-statement (SS) and MFC Big Five personality measure and the criterion measures. All of the ordinal response (Likert-type) measures were scored using the traditional summative approach: reverse score responses to negatively worded items and sum values corresponding to the endorsed item categories to obtain a scale score. The MFC personality measure was scored using the GGUM-RANK MCMC algorithm. In Table 22, note that the mean of the personality traits for the MFC measure are approximately zero because multivariate standard normal priors were used for estimation. The reliability of the MFC personality scales was computed using the marginal reliability equation described in Method. For all other measures, reliability was estimated using coefficient alpha.

As can be seen in Table 22, the coefficient alphas for SS Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism scales were .80, .87, .86, .79, and .87, respectively. The MFC marginal reliabilities for the same traits were .69, .68, .60, .70, and .68, respectively. The lower marginal reliabilities are consistent with other studies involving MFC personality measures (e.g., Brown, 2010; Chernyshenko et al., 2009). This might be explained by reduced variance in trait scores due to regression to the mean associated with Bayesian estimation and short (20-triplet) tests, or possibly somewhat low intrablock discrimination (see Table 25). Conversely, the marginal reliabilities could be better indicators of reliability than the SS alpha values, which may be inflated by single subject response consistency bias (Stark et al., 2014).

Table 22. Descriptive Statistics for Personality and Criterion Measures.

Measure	# of Items or statements	Scale Statistics					
		N	M	SD	Min.	Max.	Reliability
O-SS	12	495	45.22	7.29	19.00	60.00	0.80
O-MFC	12	495	0.00	0.83	-2.16	2.13	**0.69**
C-SS	12	495	44.50	7.81	19.00	60.00	0.87
C-MFC	12	495	0.00	0.83	-2.53	2.12	**0.68**
E-SS	12	495	39.01	8.38	14.00	60.00	0.86
E-MFC	12	495	0.00	0.80	-2.12	1.61	**0.60**
A-SS	12	495	44.76	6.63	18.00	60.00	0.79
A-MFC	12	495	0.00	0.83	-2.27	2.67	**0.70**
N-SS	12	495	30.84	8.98	12.00	59.00	0.87
N-MFC	12	495	0.00	0.82	-1.75	2.23	**0.68**
SWLS	5	495	22.86	7.04	30.00	5.00	0.90
AGG	12	495	29.94	10.61	48.00	12.00	0.91
PA	6	495	25.69	5.82	28.00	7.00	0.93
NA	6	495	18.11	6.89	32.00	8.00	0.92
HR	10	495	27.86	9.78	40.00	10.00	0.90
HI	10	495	33.08	9.50	40.00	10.00	0.90
HA	10	495	34.46	9.58	40.00	10.00	0.90
HS	10	495	34.16	9.35	40.00	10.00	0.90
HE	10	495	30.75	9.30	40.00	10.00	0.89
HC	10	495	31.16	9.73	40.00	10.00	0.92

Note. O = Openness; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism; SWLS = Life Satisfaction; PA = Positive Affect; NA = Negative Affect; AGG = Aggression; HR = Holland Realistic; HI = Holland Investigative; HA = Holland Artistic; HS = Holland Social; HE = Holland Enterprising; HC = Holland Conventional. SS = Single Statement Likert-type Measure, MFC = Multidimensional Forced Choice Measure Reliability estimates for triplet MFC personality measures are marginal reliabilities, but reliability estimates for the other measures are Cronbach's coefficient alpha.

Construct validity

Table 23 presents the multi-trait multi-method (MTMM) correlations between the SS and MFC Big Five scores. Convergent validity correlations between the corresponding SS and MFC constructs are shown in bold. These monotrait heteromethod correlations ranged from .50 to .68; they are lower than the .75 to .87 correlations found by Heggestad et al. (2006), with measures composed of the same statements. However, those authors used an unpacking and repackaging strategy to score the MFC measure, rather than an MFC IRT model. In contrast, the correlations are closer to those reported in Chernyshenko et al. (2009), who used the MUPP IRT model (Stark et al., 2005) to score MFC measures having a moderate degree of overlap.

Of greater concern in this study is the high heterotrait monomethod correlations for MFC triplets which relate to discriminant validity. For example, the correlations between Openness and Agreeableness and between Conscientiousness and Neuroticism were .78 and -.75. This is intriguing because the Likert-type correlations were much lower and respondents had no incentives to try to distort their answers by answering in a socially desirable way. More research is needed to investigate whether the manner with which statements were combined in the MFC measure induced a response set that adversely affected discriminant validity (no attempt was made to control socially desirability responding); whether the placement of the MFC questionnaire in the online survey invoked a response set; and whether aspects of the MCMC algorithm led to inflated trait correlations because of the brevity of the measure.

Table 23. Correlations Between Single Statement (SS) and Multidimensional Forced Choice (MFC) Big Five Scores.

		SS					MFC				
Format	Construct	O	C	E	A	N	O	C	E	A	N
SS	Openness	(.80)									
	Conscuentiousness	.34**	(.87)								
	Extraversion	.26**	.37**	(.86)							
	Agreeableness	.46**	.46**	.31**	(.79)						
	Neuroticism	-.18**	-.61**	-.53**	-.47**	(.87)					
MFC	Openness	**.53****	.15**	.03	.33**	-.10*	(.69)				
	Conscuentiousness	.13**	**.61****	.24**	.29**	-.43**	.42**	(.68)			
	Extraversion	.12**	.24**	**.68****	.22**	-.40**	.17**	.36**	(.60)		
	Agreeableness	.36**	.32**	.25**	**.50****	-.37**	.78**	.63**	.46**	(.70)	
	Neuroticism	-.06	-.47**	-.46**	-.28**	**.66****	-.20**	-.75**	-.64**	-.55**	(.68)

Note. N= 495, * = *p* < .05; ** = *p* < .01; Bold coefficients indicate monotrait heteromethod correlations; Values enclosed in parentheses are reliabilities; SS= singe-statement responses via Likert type sum scores; MFC = Multidimensional RANK responses via GGUM-RANK scores; O = Openness; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism.

Criterion-Related Validity

Table 24 shows the correlations of the SS and MFC personality scores with the criterion variables (e.g., life satisfaction, aggression, positive and negative affect, and vocational interests). Overall, the correlations are consistent with previous meta-analytic findings. For example, Schmidt and Shultz (2008) found significant relationships between Conscientiousness, Extraversion, Agreeableness, and Neuroticism with life satisfaction (*r* = .25, .28, .14, and -.38), positive affect (*r* = .27, .44, .12, and .30), and negative affect (*r* = -.20, -.18, -.20, and .54). Barrick, Mount and Gupta (2003) found significant relationships between Openness to Experience and Investigative (*r* = .25), Artistic (*r* = .39), and Social (*r* = .12) interests; between Extraversion and Social (*r* = .29) and Enterprising (*r* = .41) interests; and between Agreeableness and Social (*r* = .15) interests. Several studies have also found that Agreeableness and Neuroticism correlate with aggression (Gleason et al., 2004; Graziano et al., 1996; Miller et al., 2003; Suls et al., 1998). Overall, the results in Table 24 indicated that the SS and MFC measures exhibited a similar pattern of correlations with outcomes, but the MFC correlations were generally lower. Ordinarily, this findings might be used to suggest that the correlations between SS (Likert-type) Big Five measures and Likert-type criterion variables were inflated by common method bias, but the better discriminant validity of the SS measure casts doubt on this explanation. Clearly, more research is needed to understand the intricacies of MFC triplet test construction and how discriminant validity may affect criterion-related validities in applied settings.

Table 24. Criterion-Related Validity Coefficients of Personality Traits using Single Statement Responses and MFC Responses.

	SS					MFC				
Crietrion Variables	O	C	E	A	N	O	C	E	A	N
SWLS	.06	**.47**	**.43**	**.28**	**-.59**	-.01	**.32**	**.32**	**.18**	**-.42**
AGG	**-.28**	**-.42**	**-.16**	**-.55**	**.46**	**-.31**	**-.42**	**-.18**	**-.46**	**.38**
PA	**.22**	**.53**	**.49**	**.42**	**-.67**	.04	**.33**	**.33**	**.24**	**-.46**
NA	**-.21**	**-.53**	**-.35**	**-.36**	**.68**	**-.11**	**-.43**	**-.28**	**-.30**	**.48**
HR	.03	.00	**.19**	-.04	**-.11**	-.07	-.08	.07	-.08	.01
HI	**.21**	**.11**	**.13**	.09	-.07	**.11**	.02	.02	.06	.01
HA	**.39**	.00	**.17**	**.19**	.01	**.27**	-.05	.07	**.17**	.04
HS	**.26**	**.18**	**.39**	**.35**	**-.19**	**.12**	.07	**.22**	**.20**	**-.11**
HE	.09	**.23**	**.40**	.09	**-.26**	-.05	.08	**.24**	.04	**-.20**
HC	.01	**.19**	**.21**	.09	**-.18**	-.08	.05	.04	-.03	-.09

Note. N = 495; values in bold are statistically significant (*p* < .05); SWLS = Life Satisfaction; PA = Positive Affect; NA = Negative Affect; AGG = Aggression; HR = Holland Realistic; HI = Holland Investigative; HA = Holland Artistic; HS = Holland Social; HE = Holland Enterprising; HC = Holland Conventional; O = Openness; C = Conscientiousness; E = Extraversion; A = Agreeableness; N = Neuroticism.

For readers interested in the estimated GGUM-RANK statement parameters and standard errors (PSD) values for the items of the MFC triplet personality measure, the detailed results are shown in Table 25. It can be seen that many of the τ PSD values were quite large, suggesting that a larger sample might have been helpful for parameter estimation to handle the potential effects of unmotivated, careless, or socially desirable responding among the online (Amazon Mechanical Turk) participants. Alternatively, the PSDs may have been large because the average intrablock discrimination was 1.07 (values ranged from .68 to 1.47), which would fall into the category of low intrablock discrimination in the Study 1 simulation.

Table 25. Item Parameters and Standard Errors for 20-Triplet MFC Personality Measure.

Item	Alpha	PSD	Delta	PSD	Tau	PSD	Item	Alpha	PSD	Delta	PSD	Tau	PSD
1	1.83	0.23	0.98	0.15	-1.33	0.31	11	2.00	0.27	1.23	0.28	-1.96	0.39
	0.61	0.12	2.21	0.41	-0.83	0.81		0.73	0.18	-1.57	0.51	-0.53	0.81
	1.00	0.15	2.22	0.32	-1.30	0.58		1.69	0.22	2.19	0.30	-1.44	0.46
2	1.15	0.17	1.97	0.36	-2.03	0.48	12	0.88	0.14	-1.84	0.42	-0.84	0.57
	0.91	0.15	1.90	0.49	-0.82	0.65		0.67	0.11	-2.00	0.42	-1.50	0.66
	0.99	0.15	-2.24	0.34	-0.25	0.52		0.60	0.16	1.23	0.43	-0.65	0.70
3	1.57	0.21	-1.64	0.45	-1.65	0.59	13	1.89	0.23	1.01	0.16	-1.60	0.29
	1.16	0.16	-2.28	0.33	-0.69	0.66		1.11	0.16	2.10	0.31	-2.24	0.42
	0.85	0.14	-2.03	0.39	-0.85	0.79		0.72	0.15	2.04	0.47	-0.03	0.58
4	1.41	0.22	1.38	0.31	-1.48	0.44	14	1.59	0.23	1.57	0.41	-1.57	0.59
	1.21	0.17	1.73	0.28	-1.96	0.46		0.96	0.16	-2.22	0.37	-1.01	0.77
	0.66	0.15	-2.00	0.42	-0.34	0.70		1.31	0.20	2.22	0.32	-0.60	0.59
5	1.55	0.33	1.02	0.35	-0.86	0.47	15	0.63	0.19	-1.23	0.57	-1.04	0.70
	0.62	0.15	-1.47	0.47	-0.77	0.80		0.84	0.16	-1.46	0.42	-0.31	0.52
	1.50	0.20	2.16	0.31	-1.66	0.46		0.81	0.14	1.76	0.44	-1.59	0.55
6	1.32	0.17	-2.09	0.35	-0.67	0.48	16	0.65	0.11	-2.07	0.43	-0.80	0.84
	0.67	0.13	1.78	0.51	-0.94	0.74		0.95	0.16	1.58	0.47	-1.38	0.64
	0.75	0.13	2.26	0.36	-1.24	0.70		1.67	0.21	2.14	0.32	-1.26	0.48
7	0.60	0.13	-1.71	0.44	-0.88	0.74	17	1.77	0.26	1.10	0.23	-1.91	0.32
	0.84	0.15	-1.35	0.43	-0.27	0.54		1.02	0.18	-1.74	0.43	-0.90	0.58
	0.82	0.13	2.10	0.38	-1.73	0.55		0.54	0.17	-0.91	0.46	-0.82	0.86
8	1.33	0.21	0.91	0.21	-0.55	0.55	18	0.84	0.14	2.33	0.33	-0.98	0.78
	1.07	0.17	1.58	0.36	-0.87	0.66		1.16	0.15	2.04	0.36	-1.13	0.59
	1.26	0.17	2.00	0.33	-1.43	0.56		1.47	0.18	1.37	0.19	-0.94	0.49
9	1.02	0.19	1.52	0.42	-1.35	0.68	19	0.49	0.16	0.13	0.45	-0.40	0.72
	1.60	0.26	-2.09	0.42	-0.35	0.52		0.69	0.13	-1.79	0.44	-1.62	0.56
	0.96	0.15	2.15	0.36	-1.04	0.71		0.87	0.14	1.85	0.37	-1.32	0.51
10	1.06	0.19	-1.73	0.39	-0.51	0.60	20	1.76	0.22	-2.21	0.36	-0.91	0.47
	0.73	0.15	-1.50	0.49	-0.94	0.79		1.14	0.16	1.92	0.40	-1.38	0.70
	1.13	0.18	1.43	0.39	-1.44	0.59		0.83	0.13	-2.02	0.47	-0.51	0.71

Table 25 shows estimated statement parameters and standard errors for the MFC triplet personality measure. Result suggest that statement parameters were well estimated based on standard error. Comparing to alpha and delta item parameters, tau parameter showed somewhat higher standard errors. This is consistent with the simulation findings. Average intrablock discrimination of the MFC triplet measure was 1.07 (ranging from .68 to 1.47), which belongs to the low intrablock discrimination condition in the simulation study.

Finally, Figure 11 shows the overall item information (OII) values of the 20-triplet MFC personality measure. It is readily apparent that four of the items provided very little information – Item 7 (OII = 0.59), Item 12 (OII = 0.54), Item 15 (OII = 0.56), and Item 19 (OII = 0.49) – which is problematic with such a short measure. (The corresponding average intrablock item discrimination values were: Item 7 (avg. α =0.75), Item 12 (avg. α =0.72), Item 15 (avg. α = 0.76), and Item 19 (avg. α =0.68). The important implication is that to ensure adequate measurement precision, item pretesting may remain an important step in MFC test construction, despite the ability to simultaneously estimate statement parameters and score responses with the MCMC algorithm developed in this research.

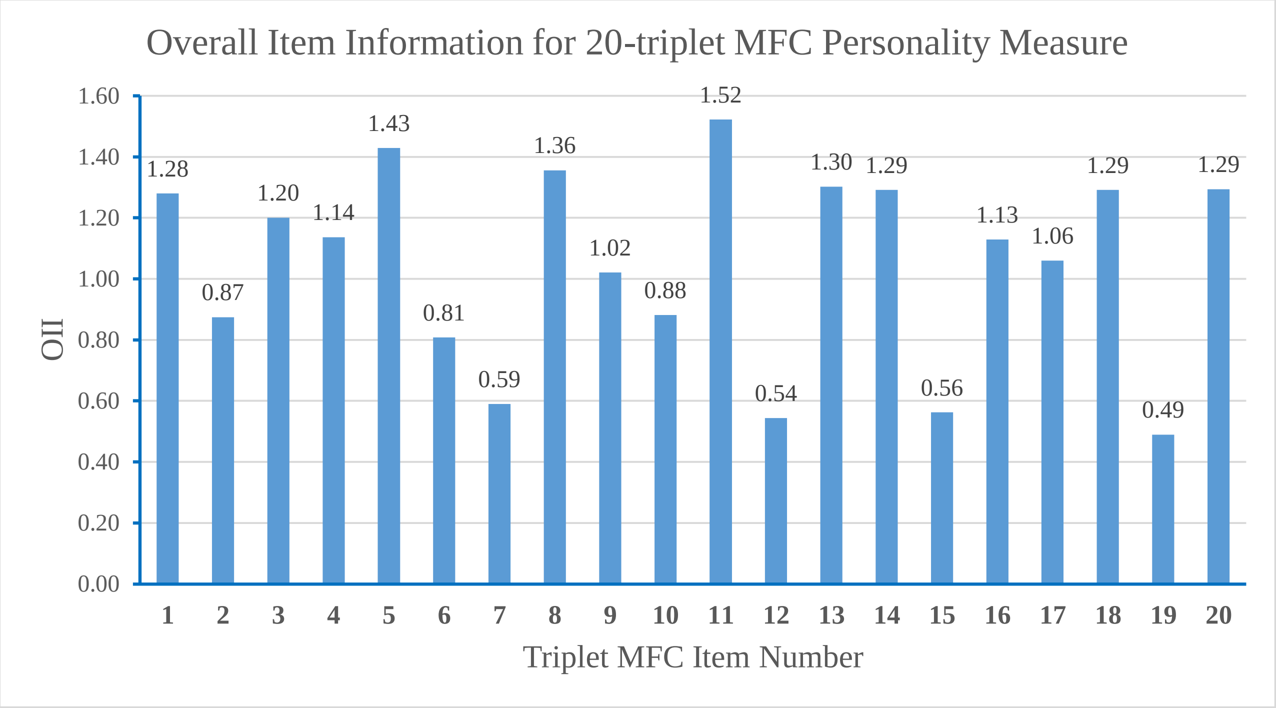


Figure 11. Individual item information for 20-triplet MFC personality measure

CHAPTER FOUR:

DISCUSSION AND CONCLUSION

This dissertation research had multiple goals. The first aim was to develop an MCMC algorithm for estimating GGUM-RANK statement and person parameters simultaneously from MFC rank responses. The second aim was to investigate the recovery of statement and person parameters with MFC triplet tests and compare with the results for MFC pair tests in select conditions. The third aim was to investigate how manipulating statement parameters influences overall item and test information and compare the information provided by MFC triplet and pair measures. The fourth aim was to examine the correspondence between GGUM-RANK MFC triplet and SS Big Five personality scores using data collected from online research participants. An overarching goal of this dissertation is to provide practitioners and researchers with practical guidelines for constructing effective MFC measures.

Findings and Implications from Proposed Studies

The main findings and practical implications of these studies are as follows. ***First***, with regard to sample size, larger sample size yielded more accurate statement parameter estimates, but sample size had little influence on person parameter estimates. The results suggest that at least 250 respondents are needed for GGUM-RANK estimation with MFC triplets test involving highly discriminating statements, and larger samples (e.g., N=500) are recommended for statement parameter estimation when measures are developed for high-stakes decision making.

Second, with regard to test length, 30-Triplets may be sufficient for scoring with 10-dimension MFC measures, provided that the triplets are pretested to ensure adequate intrablock discrimination and OII. For example, in the 30-Triplet, High Intrablock Discrimination conditions of Study 1, the average correlation between true and estimated person parameters was above .90. Importantly, using short MFC triplet measures should decrease the “cognitive load” on respondents, relative to long MFC triplet measures, and in turn reduce test fatigue, careless responding, and completion time.

Third, intrablock discrimination was found to be of primary importance for estimation accuracy. MFC items involving statements with high discrimination parameters produced more accurate parameter estimation and higher overall item and test information. Thus, researchers and practitioners are strongly encouraged to create MFC tests comprising highly discriminating statements to ensure sufficient measurement precision. Importantly, the GGUM-RANK MCMC direct estimation process will help practitioners to more accurately evaluate item discrimination by taking into account potential interactions among statements within a block. This should also lead to more effective MFC item analysis and facilitate construction of parallel MFC test forms.

Fourth, intrablock location variability had little to no effect on overall item and test information statistics and parameter recovery. This result has implications for creating fake-resistant MFC measures, because it shows that statements within a block can be matched more closely on location (extremity) and social desirability without adversely affecting the psychometric quality of the items.

Fifth, MFC triplet measures will outperform MFC pair measures of similar length and intrablock discrimination in terms of estimation accuracy. In this research, 30-Triplet tests consistently yielded better discrimination and location parameter recovery than 30-Pair tests, and the 30-Triplet tests were nearly as good as 90-Pair tests in terms of overall test information and person parameter recovery. In addition, triplet measures had approximately 2.8 times higher average overall item information than pair measures. Together, these results show the potential psychometric benefits of using the triplet format, provided that perceived item “difficulty” does not lead to aberrant responding with real examinees.

Lastly, this dissertation not only developed GGUM-RANK estimation methods but also illustrated their viability for applied use. Although questions were raised concerning discriminant validity of the MFC measure in the empirical investigation using online research participants, the study showed that an MFC measure, which was not pretested with real examinees to ensure item quality, could yield patterns of correlation with outcomes similar to Likert-type Big Five measures that have been widely in applied research. This empirical example will open the way for practical applications of the GGUM-RANK IRT model.

Limitations and Suggestions for Future Research

First, due to extremely long simulation run-times, Study 1 and Study 2 considered a limited number of simulation conditions out of all possibilities that may be seen in real MFC testing applications. These simulations explored parameter recovery exclusively with 10-dimension tests, but MFC tests of higher dimensionality are used in some applied settings. For example, TAPAS personality tests (Stark et al., 2014) used for military personnel testing have measured 13-15 dimensions with multidimensional pairwise preference items, and the Occupational Personality Questionnaire (OPQ32; Brown & Bartram, 2009b) measures up to 32 work-related behaviors using MFC triplets. Thus, simulation research is needed to explore the accuracy of GGUM-RANK scoring with measures involving more than 10 dimensions (e.g., 20 dimensions or 30 dimensions). In addition, this dissertation considered just two levels of intrablock location parameter variability (large and small SD). Future simulation research should examine a wider variety of location parameter variability conditions; e.g., the effect of using all positive or all negative location parameters within MFC blocks on overall item and test information and parameter recovery. Brown and Maydeu-Olivares (2011) suggested that MFC items should be created by mixing positively and negatively worded statements to ensure more accurate parameter estimation with their Thurstonian model. That is analogous to mixing statements with positive and negative GGUM-RANK location (δ) parameters within MFC blocks. However, the results of this investigation do not directly support or contradict that recommendation. If future GGUM-RANK research finds that all positive or all negative statements can be used in MFC blocks without adversely affecting parameter estimation, then there will be potentially greater resistance to faking and related forms of response distortion.

Second, empirical Study 3 showed that the MFC and SS personality measures had similar patterns of correlation with criterion variables, but the discriminant validity of the MFC measure was questionable. As previously mentioned, the MFC measure was constructed by recombining the statements of a SS Big Five measure with emphasis on balancing positive and negative wording across MFC blocks. Had the measure been constructed from a pool of discriminating statements following the guidelines of Stark et al. (2005), for example, better results may have been observed. To address similar problems in future MFC applications, practitioners should carefully pretest statement pools and select statements with good psychometric properties for creating MFC measures. It may also be beneficial to match statements within blocks on location and social desirability if there is any potential for deliberate or unintentional socially desirable responding. Thus, there remains a need for validity research using GGUM-RANK MFC triplet measures with real examinees in research and organizational settings.

Third, this research suggests that MFC triplet tests provide greater measurement precision MFC pairs. However, it has been suggested that block size (i.e., number of statements within an MFC block), is positively associated with a respondent's "cognitive load" (Brown & Maydeu-Olivares, 2011). That is, it is more cognitive demanding to respond to a triplet than a pair. The potential positive relationship between cognitive load and block size may affect test-taking anxiety, positive affect, applicant reaction, test fairness, and adverse impact (e.g., Converse et al., 2008). Future empirical research should, therefore, examine the effect of MFC block size (e.g., pairs vs. triplets vs. tetrads) on respondents' perceived cognitive load, reactions toward testing, test fairness, and adverse impact in high-stake contexts.

Lastly, this research focused on single-sample parameter estimation and scoring. To facilitate applications, research is needed to develop methods for assessing GGUM-RANK model-data fit, linking item and person parameters across different subpopulations, and equating MFC test forms comprising different subsets of items. The development of differential item and test functioning methods for GGUM-RANK measures would help to support multinational testing efforts and permit meaningful cross-cultural comparisons.

In closing, there is increasing interest in the use of MFC measures for noncognitive measurement in I-O and educational settings. This research has extended methods developed in previous investigations (e.g., Hontangas et al., 2015; Seybert, 2013; Stark et al., 2005) by introducing an MCMC algorithm for estimating GGUM-RANK person and statement parameters from MFC triplet responses and by examining how parameter recovery is influenced by overall item and test information. It is hoped that this research provides a solid foundation for applied research and a springboard for future psychometric development efforts.

REFERENCES

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

Anguiano-Carrasco, C. MacCann, C., Geiger, M., Seybert, J. M., & Roberts, R. D. (2014). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment, 33,* 83-97.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

Barrick, M. R., Mount, M. K., & Gupta, R. (2003). Meta-analysis of the relationship between the Five-Factor Model of personality and Holland's occupational types. *Personnel Psychology, 56,* 45-74.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: what do we know and where do we go next?. *International Journal of Selection and Assessment*, *9*, 9-30.

Bartram, D. (2005). The Great Eight competencies: a criterion-centric approach to validation. *Journal of Applied Psychology, 90*, 1185-1203.

Bartram. D., & Burke, E. (2013). Handbook of test security. Routledge, In Wollack, J. A., & Fremer, J. J. (2013), *Industrial/Organizational Testing Case Studies*.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: a review and meta-analysis. *Journal of Applied Psychology*, *92*, 410-424.

Bock, R. D., & Aitken, M.(1981), Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395-414.

Bono, J. E., & Judge, T. A. (2004). Personality and transformational and transactional leadership: a meta-analysis. *Journal of Applied Psychology, 89*, 901-910.

Borman, W. C., Buck, D. E., Hanson, M. A., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001a). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*, 965-973.

Borman, W. C., Penner, L. A., Allen, T. D., & Motowidlo, S. J. (2001b). Personality predictors of citizenship performance. *International Journal of Selection and Assessment*, *9*, 52-69.

Bowen, C. C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis*, *10*, 240-259.

Brown, A. (2010). *How IRT can solve problems of ipsative data* (Doctoral dissertation). University of Barcelona, Spain.

Brown, A., & Bartram, D. (2009a, April). *Doing less but getting more: improving forced-choice measures with IRT.* Poster presented at the 24th Annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Brown, A., & Bartram, D. (2009b). *Development and psychometric properties of OPQ32r* (Supplement to the OPQ32 technical manual). Thames Ditton, UK: SHL Group Limited.

Brown, A., & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York: Routledge/Taylor & Francis Group.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*, 460-502.

Brown, A., & Maydeu-Olivares, A. (2014). Modeling forced-choice response formats. In P. Irwing, Booth, T., & Hughes, D. (Eds.), *The Wiley Handbook of Psychometric Testing*. John Wiley & Sons.

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: a measurement model for the Buss–Perry Aggression Questionnaire. *Journal of Research in Personality*, *35*, 138-167.

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

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009) Normative Scoring of Multidimensional Pairwise Preference Personality Scales Using IRT: Empirical Comparisons With Other Formats, *Human Performance, 22, 105-127.*

Cheung, M.W.L, & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*, 55-77.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267-307. Clemans, W. V. (1966). *An anlyticial and empirical examination of some properties of ipsative measures* (Psychometrika Monograph No. 14). Richmond, VA : Psychometric Society.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic

Colley, S. K., Lincolne, J., & Neal A. (2013). An examination of the relationship amongst profiles of perceived organizational values, safety climate and safety outcomes, *Safety Science, 51*, 69-76.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality tests and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155-169.

de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2012). *Some extensions of the multiunidimensional pairwise preference model.* Paper presented at the 26th annual meeting of the Society for Industrial and Organizational Psychology, Chicago, IL.

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.

DeNeve, K. M. & Cooper, H (1998). The happy personality: A meta-analysis of 137 personality traits and subjective well-being. *Psychological Bulletin, 124,* 197-229.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of personality assessment*, *49*, 71-75.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*, 143-156.

Doornik, J. A. (2009). An Object-Oriented Matrix Programming Language Ox 6.

Ferrando, P. J., Demestre, J., Anguiano-Carrasco, C., & Chico, E. (2011). Evaluación TRI de la escala IE de Rotter: un nuevo enfoque y algunas consideraciones. *Psicothema*, *23*, 282-288.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science,* 457-472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6,* 721-741.

Gleason, K. A., Jensen-Campbell, L. A., & Richardson, D. S. (2004). Agreeableness as a predictor of aggression in adolescence. *Aggressive Behavior, 30,* 43-61.

Graziano, W. G., Jensen-Campbell, L. A., & Hair, E. C. (1996). Perceiving interpersonal conflict and reacting to it: The case for Agreeableness. *Journal of Personality and Social Psychology, 70*, 820-835.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36,* 341-355.

Hastings, W. K. (1970). Monte Carlo simulation methods using Markov chains and their applications. *Biometrika, 57,* 97-109.

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. (2014). Response Styles and Personality Traits A Multilevel Analysis. *Journal of Cross-Cultural Psychology, 45*, 1028-1045.

He, J., & van de Vijver, F. J. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55*, 794-800.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167-184.

Holland, J. L. (1985). *Making vocational choices: A theory of vocational personalities and work environments* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Hontangas, P. M, de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing Traditional and IRT Scoring of Forced-Choice Tests, *Applied Psychological Measurement,* 1-15.

Hough, M. H., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effects of response distortion on those validities. *Journal of Applied Psychology, 75*, 581-595.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and I-O psychology: Reflections, progress and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 272-290.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869-879.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution?. *Human Performance, 13,* 371-388.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153-162.

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.

Joo, S., Lee, P, & Stark, S. (2016). Individual Paper Accepted: “Information Functions of Multidimensional Forced-Choice IRT models”. Individual Paper Presentation will be conducted at 2016 NCME (National Council on Measurement in Education), Washington D.C.

Judge, T. A., & Hurst, C. (2007). Capitalizing on one's advantages: role of core self-evaluations. *Journal of Applied Psychology, 92*, 1212-1227.

Koenig, J. A., & Roberts, J. S. (2007). Linking parameters estimated with the generalized graded unfolding model: A comparison of the accuracy of characteristic curve methods. *Applied Psychological Measurement*, *31,* 504-524.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90,* 442–452.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66,* 81-95.

McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, *8,* 222-248.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*, 531-552.

Metropolis, N., Rosenbulth, A. W., Rosenbluth, M. N., Teller A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087-1091.

Miller, J. D., Lynam, D., & Leukefeld, C. (2003). Examining antisocial behavior through the five-factor model of personality. *Aggressive Behavior, 29*, 497-514.

Morris, M. A. (2003). *A meta-analytic investigation of vocational interest-based job fit, and its relationship to job satisfaction, performance, and turnover*(Doctoral dissertation, ProQuest Information & Learning).

Muthén, L. K., & Muthén, B. O. (1998–2015). Mplus *user's guide (7th ed.)*. Los Angeles, CA:

Muthén & Muthén. Retrieved from www.statmodel.com

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

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*, 153-172.

Reise, P. S., & Yu, J. (1990). Parameter Recovery in the Graded Response Model Using MULTILOG, *Journal of Educational Measurement*, *27*, 133-144.

Roberts, J. S. (2000). GGUM2000 [Computer software]. *College Park, MD: Author*.

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.

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). *O*NET Interest Profiler Short Form Psychometric Characteristics: Summary.* O*NET Resource Center.

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*, 3-30.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, *39*, 111-121.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, *124*, 262-274.

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology*, *91*, 613-621.

Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*, 1-20.

Semadar, A., Robins, G., & Ferris, G. R. (2006). Comparing the validity of multiple social effectiveness constructs in the prediction of managerial job performance. *Journal of Organizational Behavior*, *27,* 443-461.

Seybert, J. (2013). A New Item Response Theory Model for Estimating Person Ability and Item Parameters for Multidimensional Rank Order Responses (Doctoral dissertation). Retrieved from http://scholarcommons.usf.edu/etd/4942.

Sinha, R., Oswald, F., Imus, A., & Schmitt, N. (2011). Criterion-focused approach to reducing adverse impact in college admissions. *Applied Measurement in Education*, *24*, 137-161.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30,* 298-321.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment: The generalized graded unfolding model for multi-dimensional paired comparison responses* (Doctoral dissertation). University of Illinois at Urbana-Champaign. Urbana-Champaign, IL.

Stark, S., Chernyshenko, O.S., Chan, K.Y., Lee, W.C., & Drasgow, F. (2001). Effects of the Testing Situation on Item Responding: Cause for Concern. *Journal of Applied Psychology, 86,* 943-953.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, *29,* 184-203.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2012). Constructing fake-resistant personality tests using item response theory: High stakes personality testing with multidimensional pairwise preferences. *New Perspectives on Faking in Personality Assessments; NY: Oxford University Press*, 214-239.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., & Farmer, W. L. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology, 26,* 153-164.

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods, 15,* 463-487.

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

Stark, S., & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26, 208-227.*

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134,* 138-161.

Suls, J., Martin, R., & David, J. P. (1998). Person-environment fit and its limits: Agreeableness, neuroticism, and emotional reactivity to interpersonal conflict. *Psychology Bulletin, 24,* 88-98.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701-1728.

Van Rooy, D. L., & Viswesvaran, C. (2004). Emotional intelligence: A meta-analytic investigation of predictive validity and nomological net. *Journal of Vocational Behavior*, *65*, 71-95.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, *19,* 175-199.

White, L. A., & Young, M. C. (1998). *Development and validation of the Assessment of Individual Motivation (AIM).* Paper presented at the Annual Meeting of the American Psychological Association, San Francisco, CA.

Appendix A: Derivatives for GGUM-RANK Item Information Functions (IIFs)

For IIFs, the first and second partial derivatives of Equation 13 are needed for cases involving statements on different dimensions. Begin by defining the rank response probabilities and their derivatives with respect to dimension *d1*, *d2*, and *d3* as:

$$A = P_{d1}(1), \qquad A' = P'_{d1}(1) \;\& \qquad A'' = P_{d1}{}''(1)$$

$$B = P_{d2}(1), \qquad B' = P'_{d2}(1) \;\& \qquad B'' = P_{d2}{}''(1)$$

$$C = P_{d3}(1), \qquad C' = P'_{d3}(1) \;\& \qquad C'' = P_{d3}{}''(1)$$

$$D = P_{d1}(0) = 1 - P_{d1}(1) \; and \; D' = P_{d1}{}'(0) = -P_{d1}{}'(1) = -A'$$

$$E = P_{d2}(0) = 1 - P_{d2}(1) \; and \; E' = P_{d2}{}'(0) = -P_{d2}{}'(1) = -B'$$

$$F = P_{d3}(0) = 1 - P_{d3}(1) \; and \; F' = P_{d3}{}'(0) = -P_{d3}{}'(1) = -C'$$

where, *d1* = first dimension 1, *d2*= second dimension, *d3* = third dimension.

First partial derivatives of GGUM-RANK triplet probability functions follow:

$$\frac{\boldsymbol{\partial P_{m=1}(\theta)}}{\boldsymbol{\partial \theta_{d1}}} = \frac{\partial}{\partial \theta_{d1}}\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{BF}{BF+EC}\right)$$

$$= \left(\frac{AEF}{AEF+DBF+DEC}\right)'\left(\frac{BF}{BF+EC}\right)$$

$$= \left(\frac{(A'EF)(AEF+DBF+DEC)-(AEF)(A'EF+D'BF+D'EC)}{(AEF+DBF+DEC)^2}\right)\left(\frac{BF}{BF+EC}\right).$$

$$\frac{\boldsymbol{\partial P_{m=1}(\theta)}}{\boldsymbol{\partial \theta_{d2}}} = \frac{\partial}{\partial \theta_{d2}}\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{BF}{BF+EC}\right)$$

$$= \left(\frac{AEF}{AEF+DBF+DEC}\right)'\left(\frac{BF}{BF+EC}\right) + \left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{BF}{BF+EC}\right)'$$

$$= \left(\frac{(AE'F)(AEF+DBF+DEC)-(AEF)(AE'F+DB'F+DE'C)}{(AEF+DBF+DEC)^2}\right)\left(\frac{BF}{BF+EC}\right)$$

$$+ \left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{(B'F)(BF+EC)-(BF)(B'F+E'C)}{(BF+EC)^2}\right).$$

$$\frac{\boldsymbol{\partial P_{m=1}(\theta)}}{\boldsymbol{\partial \theta_{d3}}} = \frac{\partial}{\partial \theta_{d3}}\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{BF}{BF+EC}\right)$$

$$= \left(\frac{AEF}{AEF+DBF+DEC}\right)'\left(\frac{BF}{BF+EC}\right) + \left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{BF}{BF+EC}\right)'$$

$$= \left(\frac{(AEF')(AEF+DBF+DEC)-(AEF)(AEF'+DBF'+DEC')}{(AEF+DBF+DEC)^2}\right)\left(\frac{BF}{BF+EC}\right)$$

$$+ \left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{(BF')(BF+EC)-(BF)(BF'+EC')}{(BF+EC)^2}\right).$$

For the ***second partial derivatives*** of GGUM-RANK triplet probability functions:

$$\frac{\boldsymbol{\partial^2 P_{m=1}(\theta)}}{\boldsymbol{\partial \theta_{d1}^2}} = \frac{\partial}{\partial \theta_{d1}}\left[\left(\frac{(A'EF)(AEF+DBF+DEC)-(AEF)(A'EF+D'BF+D'EC)}{(AEF+DBF+DEC)^2}\right)\left(\frac{BF}{BF+EC}\right)\right].$$

Let us define

$\alpha_1^2 = [(A''EF)(AEF + DBF + DEC) + (A'EF)(A'EF + D'BF + D'EC) - (A'EF)(A'EF + D'BF + D'EC) - (AEF)(A''EF + D''BF + D''EC)](AEF + DBF + DEC)^2 - 2(AEF + DBF + DEC)(A'EF + D'BF + D'EC)[(A'EF)(AEF + DBF + DEC) - (AEF)(A'EF + D'BF + D'EC)]$.

Therefore,

$$\frac{\partial^2 P_{m=1}(\boldsymbol{\theta})}{\partial \theta_{d1}^2} = \frac{\alpha_1^2}{(AEF+DBF+DEC)^4}\left(\frac{BF}{BF+EC}\right),$$

$$\frac{\boldsymbol{\partial^2 P_{m=1}(\theta)}}{\boldsymbol{\partial \theta_{d2}^2}} = \frac{\partial}{\partial \theta_{d2}}\left[\left(\frac{(AE'F)(AEF+DBF+DEC)-(AEF)(AE'F+DB'F+DE'C)}{(AEF+DBF+DEC)^2}\right)\left(\frac{BF}{BF+EC}\right) + \right.$$

$$\left.\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{(B'F)(BF+EC)-(BF)(B'F+E'C)}{(BF+EC)^2}\right)\right]$$

Let us define

$\beta_1^2 = [(AE''F)(AEF + DBF + DEC) + (AE'F)(AE'F + DB'F + DE'C) - (AE'F)(AE'F + DB'F + DE'C) - (AEF)(AE''F + DB''F + DE''C)](AEF + DBF + DEC)^2 - 2(AEF + DBF + DEC)(AE'F + DB'F + DE'C)[(AE'F)(AEF + DBF + DEC) - (AEF)(AE'F + DB'F + DE'C)]$,

$\delta_1 = [(B''F)(BF + EC) + (B'F)(B'F + E'C) - (B'F)(B'F + E'C) - (BF)(B''F + E''C)](BF + EC)^2 - 2(BF + EC)(B'F + E'C)[(B'F)(BF + EC) - (BF)(B'F + E'C)]$.

Therefore,

$$\frac{\partial^2 P_{m=1}(\boldsymbol{\theta})}{\partial \theta_{d2}^2} = \frac{\beta_1^2}{(AEF+DBF+DEC)^4}\left(\frac{BF}{BF+EC}\right) +$$

$$\left(\frac{(AE'F)(AEF+DBF+DEC)-(AEF)(AE'F+DB'F+DE'C)}{(AEF+DBF+DEC)^2}\right)\left(\frac{(B'F)(BF+EC)-(BF)(B'F+E'C)}{(BF+EC)^2}\right) +$$

$$\left(\frac{(AE'F)(AEF+DBF+DEC)-(AEF)(AE'F+DB'F+DE'C)}{(AEF+DBF+DEC)^2}\right)\left(\frac{(B'F)(BF+EC)-(BF)(B'F+E'C)}{(BF+EC)^2}\right) +$$

$$\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{\delta_1}{(BF+EC)^4}\right)$$

$$= \frac{\beta_1^2}{(AEF+DBF+DEC)^4}\left(\frac{BF}{BF+EC}\right) +$$

$$2\left(\frac{(AE'F)(AEF+DBF+DEC)-(AEF)(AE'F+DB'F+DE'C)}{(AEF+DBF+DEC)^2}\right)\left(\frac{(B'F)(BF+EC)-(BF)(B'F+E'C)}{(BF+EC)^2}\right) +$$

$$\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{\delta_1}{(BF+EC)^4}\right) .$$

$$\boldsymbol{\frac{\partial^2 P_{m=1}(\theta)}{\partial \theta_{d3}^2}} = \frac{\partial}{\partial \theta_{d3}}\left[\left(\frac{(AEF')(AEF+DBF+DEC)-(AEF)(AEF'+DBF'+DEC')}{(AEF+DBF+DEC)^2}\right)\left(\frac{BF}{BF+EC}\right) +\right.$$

$$\left.\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{(BF')(BF+EC)-(BF)(BF'+EC')}{(BF+EC)^2}\right)\right] .$$

Let us define

$\gamma_1^2 = [(AEF'')(AEF + DBF + DEC) + (AEF')(AEF' + DBF' + DEC') - (AEF')(AEF' + DBF' + DEC') - (AEF)(AEF'' + DBF'' + DEC'')](AEF + DBF + DEC)^2 - 2(AEF + DBF + DEC)(AEF' + DBF' + DEC')[(AEF')(AEF + DBF + DEC) - (AEF)(AEF' + DBF' + DEC')]$,

$$\varepsilon_1 = [(BF'')(BF+EC) + (BF')(BF'+EC') - (BF')(BF'+EC') - (BF)(BF''+EC'')](BF+EC)^2 - 2(BF+EC)(BF'+EC')[(BF')(BF+EC) - (BF)(BF'+EC')] .$$

Therefore,

$$\frac{\partial^2 P_{m=1}(\boldsymbol{\theta})}{\partial \theta_{d3}^2} = \frac{\gamma_1^2}{(AEF+DBF+DEC)^4}\left(\frac{BF}{BF+EC}\right) +$$

$$\left(\frac{(AEF')(AEF+DBF+DEC)-(AEF)(AEF'+DBF'+DEC')}{(AEF+DBF+DEC)^2}\right)\left(\frac{(BF')(BF+EC)-(BF)(BF'+EC')}{(BF+EC)^2}\right) +$$

$$\left(\frac{(AEF')(AEF+DBF+DEC)-(AEF)(AEF'+DBF'+DEC')}{(AEF+DBF+DEC)^2}\right)\left(\frac{(BF')(BF+EC)-(BF)(BF'+EC')}{(BF+EC)^2}\right) +$$

$$\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{\varepsilon_1}{(BF+EC)^4}\right)$$

$$= \frac{\gamma_1^2}{(AEF+DBF+DEC)^4}\left(\frac{BF}{BF+EC}\right) +$$

$$2\left(\frac{(AEF')(AEF+DBF+DEC)-(AEF)(AEF'+DBF'+DEC')}{(AEF+DBF+DEC)^2}\right)\left(\frac{(BF')(BF+EC)-(BF)(BF'+EC')}{(BF+EC)^2}\right) +$$

$$\left(\frac{AEF}{AEF+DBF+DEC}\right)\left(\frac{\varepsilon_1}{(BF+EC)^4}\right).$$

The same process can be applied to obtain the first and second partial derivatives for *m=2 to m=6* in Equation 13.

Appendix B: Item Parameters for 10-D MFC Triplet Tests in Study 1

Table B1.

Test Specification for the 10-Dimension Triplet Test with Low Alpha and Low Delta SD condition

			Statement Parameters						Statement Parameters		
Item Block #	Statement	Dimension	Alpha	Delta	Tau	Item Block #	Statement	Dimension	Alpha	Delta	Tau
1	1	1	0.97	1.10	-1.37	16	46	4	0.76	-1.18	-1.14
	2	2	1.03	0.56	-0.98		47	6	0.96	-1.63	-0.56
	3	3	0.88	0.79	-0.55		48	10	1.16	-1.73	-0.69
2	4	1	0.95	-1.23	-0.55	17	49	5	0.82	1.40	-1.05
	5	6	1.04	-1.56	-0.80		50	8	1.24	1.91	-0.68
	6	9	0.82	-1.87	-1.28		51	10	0.78	1.87	-1.12
3	7	4	0.85	-1.91	-1.22	18	52	3	1.22	-1.89	-0.97
	8	5	0.90	-1.61	-0.67		53	4	0.95	-1.45	-1.27
	9	8	0.77	-1.31	-0.54		54	5	0.75	-1.97	-0.43
4	10	6	1.09	0.59	-1.17	19	55	1	1.10	1.25	-0.73
	11	7	1.00	0.11	-0.99		56	7	1.06	0.91	-1.21
	12	8	1.16	0.64	-0.67		57	10	0.83	1.50	-0.48
5	13	2	0.82	0.88	-0.99	20	58	1	1.07	-1.04	-0.80
	14	6	0.97	1.01	-1.24		59	3	0.77	-1.25	-1.13
	15	7	0.79	1.36	-0.53		60	5	0.83	-1.63	-0.41
6	16	8	0.77	-0.75	-1.04	21	61	5	0.94	0.82	-0.84
	17	9	1.12	-0.25	-0.94		62	6	0.96	1.40	-0.57
	18	10	0.80	-0.95	-0.79		63	7	0.77	1.11	-1.05
7	19	2	0.85	1.30	-1.29	22	64	3	1.15	1.54	-0.84
	20	3	1.21	1.56	-0.94		65	4	0.98	1.91	-1.12
	21	5	1.09	1.92	-0.48		66	6	1.21	1.35	-0.51
8	22	1	0.80	1.05	-0.69	23	67	7	1.06	-1.80	-0.95
	23	4	1.02	0.72	-1.31		68	8	1.12	-1.50	-1.33
	24	9	0.77	1.34	-0.72		69	9	1.04	-1.10	-0.80
9	25	1	1.04	-0.76	-1.31	24	70	3	0.78	0.18	-0.85
	26	2	0.82	-0.41	-0.86		71	8	0.96	0.77	-0.66
	27	10	0.75	-0.97	-0.51		72	10	1.09	0.47	-1.20
10	28	3	0.80	-1.22	-0.54	25	73	2	1.08	-1.38	-0.59
	29	6	1.14	-0.90	-1.12		74	3	0.78	-1.09	-0.76
	30	8	1.13	-0.69	-0.65		75	9	1.05	-0.81	-1.24
11	31	4	1.19	-0.77	-0.46	26	76	1	0.79	0.04	-1.22
	32	8	1.01	-0.35	-1.20		77	5	1.09	0.39	-1.00
	33	9	1.07	-0.18	-0.91		78	10	0.77	-0.28	-0.46
12	34	6	0.91	-1.16	-0.95	27	79	4	0.92	1.81	-0.48
	35	7	1.18	-0.78	-0.90		80	5	1.15	1.22	-1.27
	36	9	0.98	-0.60	-1.13		81	6	0.78	1.45	-0.87
13	37	2	1.18	-1.10	-0.56	28	82	4	1.03	0.26	-0.58
	38	7	1.11	-0.83	-1.13		83	5	0.91	0.13	-1.25
	39	10	0.85	-0.53	-0.45		84	9	0.76	0.71	-0.81
14	40	1	1.20	1.62	-0.62	29	85	2	1.08	1.94	-1.22
	41	9	1.03	1.91	-0.91		86	3	1.09	1.61	-0.86
	42	10	0.88	1.40	-1.31		87	4	1.22	1.33	-0.41
15	43	1	0.93	-1.85	-0.73	30	88	2	0.87	0.34	-0.84
	44	2	0.97	-1.39	-1.30		89	7	1.08	0.66	-1.20
	45	7	0.78	-1.31	-1.39		90	8	1.06	0.89	-0.47

Table B2

Test Specification for the 10-Dimension Triplet Test with Low Alpha and High Delta SD Condition

			Statement Parameters						Statement Parameters		
Item Block #	Statement	Dimension	Alpha	Delta	Tau	Item Block #	Statement	Dimension	Alpha	Delta	Tau
1	1	1	0.97	0.6	-1.37	16	46	4	0.76	0.35	-1.14
	2	2	1.03	-1.89	-0.98		47	6	0.96	1.56	-0.56
	3	3	0.88	-1.4	-0.55		48	10	1.16	-1.08	-0.69
2	4	1	0.95	-1.23	-0.55	17	49	5	0.82	1.35	-1.05
	5	6	1.04	-0.26	-0.80		50	8	1.24	-0.29	-0.68
	6	9	0.82	1.37	-1.28		51	10	0.78	-1.25	-1.12
3	7	4	0.85	-0.69	-1.22	18	52	3	1.22	-0.1	-0.97
	8	5	0.90	0.74	-0.67		53	4	0.95	-1.91	-1.27
	9	8	0.77	1.89	-0.54		54	5	0.75	0.63	-0.43
4	10	6	1.09	-0.6	-1.17	19	55	1	1.10	1.25	-0.73
	11	7	1.00	1.36	-0.99		56	7	1.06	0.91	-1.21
	12	8	1.16	1.87	-0.67		57	10	0.83	-1.18	-0.48
5	13	2	0.82	1.16	-0.99	20	58	1	1.07	-1.63	-0.80
	14	6	0.97	1.96	-1.24		59	3	0.77	-0.68	-1.13
	15	7	0.79	-0.59	-0.53		60	5	0.83	0.95	-0.41
6	16	8	0.77	-1.63	-1.04	21	61	5	0.94	-1.15	-0.84
	17	9	1.12	0.99	-0.94		62	6	0.96	1.45	-0.57
	18	10	0.80	-0.41	-0.79		63	7	0.77	0.26	-1.05
7	19	2	0.85	0.57	-1.29	22	64	3	1.15	1.91	-0.84
	20	3	1.21	1.92	-0.94		65	4	0.98	1.23	-1.12
	21	5	1.09	-0.7	-0.48		66	6	1.21	-0.61	-0.51
8	22	1	0.80	1.08	-0.69	23	67	7	1.06	-1.47	-0.95
	23	4	1.02	-1.37	-1.31		68	8	1.12	-1.72	-1.33
	24	9	0.77	-1.25	-0.72		69	9	1.04	0.65	-0.80
9	25	1	1.04	-0.97	-1.31	24	70	3	0.78	-1.87	-0.85
	26	2	0.82	0.69	-0.86		71	8	0.96	0.18	-0.66
	27	10	0.75	1.6	-0.51		72	10	1.09	0.54	-1.20
10	28	3	0.80	-1.09	-0.54	25	73	2	1.08	-1.73	-0.59
	29	6	1.14	-1.54	-1.12		74	3	0.78	0.85	-0.76
	30	8	1.13	0.98	-0.65		75	9	1.05	-0.69	-1.24
11	31	4	1.19	1.45	-0.46	26	76	1	0.79	0.04	-1.22
	32	8	1.01	0.11	-1.20		77	5	1.09	-0.97	-1.00
	33	9	1.07	-1.16	-0.91		78	10	0.77	1.71	-0.46
12	34	6	0.91	-0.49	-0.95	27	79	4	0.92	1.18	-0.48
	35	7	1.18	1.89	-0.90		80	5	1.15	-1.33	-1.27
	36	9	0.98	1.62	-1.13		81	6	0.78	-0.64	-0.87
13	37	2	1.18	-1	-0.56	28	82	4	1.03	0.37	-0.58
	38	7	1.11	-1.67	-1.13		83	5	0.91	0.97	-1.25
	39	10	0.85	0.85	-0.45		84	9	0.76	-1.55	-0.81
14	40	1	1.20	1.91	-0.62	29	85	2	1.08	1.94	-1.22
	41	9	1.03	0.81	-0.91		86	3	1.09	0.82	-0.86
	42	10	0.88	-0.68	-1.31		87	4	1.22	-0.68	-0.41
15	43	1	0.93	-1.85	-0.73	30	88	2	0.87	0.48	-0.84
	44	2	0.97	0.47	-1.30		89	7	1.08	-1.71	-1.20
	45	7	0.78	0.34	-1.39		90	8	1.06	-1.86	-0.47

Table B3

Test Specification for the 10-Dimension Triplet Test with High Alpha and Low Delta SD Condition

			Statement Parameters						Statement Parameters		
Item Block #	Statement	Dimension	Alpha	Delta	Tau	Item Block #	Statement	Dimension	Alpha	Delta	Tau
1	1	1	1.81	1.10	-1.37	16	46	4	1.95	-1.18	-1.14
	2	2	1.85	0.56	-0.98		47	6	1.83	-1.63	-0.56
	3	3	2.00	0.79	-0.55		48	10	2.17	-1.73	-0.69
2	4	1	1.83	-1.23	-0.55	17	49	5	2.05	1.40	-1.05
	5	6	2.12	-1.56	-0.80		50	8	1.92	1.91	-0.68
	6	9	2.11	-1.87	-1.28		51	10	1.98	1.87	-1.12
3	7	4	2.19	-1.91	-1.22	18	52	3	2.06	-1.89	-0.97
	8	5	2.22	-1.61	-0.67		53	4	1.92	-1.45	-1.27
	9	8	1.84	-1.31	-0.54		54	5	1.82	-1.97	-0.43
4	10	6	1.94	0.59	-1.17	19	55	1	2.00	1.25	-0.73
	11	7	1.77	0.11	-0.99		56	7	2.19	0.91	-1.21
	12	8	1.88	0.64	-0.67		57	10	1.89	1.50	-0.48
5	13	2	1.84	0.88	-0.99	20	58	1	2.01	-1.04	-0.80
	14	6	2.20	1.01	-1.24		59	3	2.13	-1.25	-1.13
	15	7	2.12	1.36	-0.53		60	5	2.15	-1.63	-0.41
6	16	8	2.09	-0.75	-1.04	21	61	5	1.96	0.82	-0.84
	17	9	1.82	-0.25	-0.94		62	6	1.97	1.40	-0.57
	18	10	1.94	-0.95	-0.79		63	7	1.81	1.11	-1.05
7	19	2	1.93	1.30	-1.29	22	64	3	1.85	1.54	-0.84
	20	3	1.76	1.56	-0.94		65	4	1.88	1.91	-1.12
	21	5	1.86	1.92	-0.48		66	6	2.08	1.35	-0.51
8	22	1	2.23	1.05	-0.69	23	67	7	1.76	-1.80	-0.95
	23	4	1.83	0.72	-1.31		68	8	1.92	-1.50	-1.33
	24	9	1.82	1.34	-0.72		69	9	1.92	-1.10	-0.80
9	25	1	1.93	-0.76	-1.31	24	70	3	1.76	0.18	-0.85
	26	2	1.94	-0.41	-0.86		71	8	1.75	0.77	-0.66
	27	10	2.23	-0.97	-0.51		72	10	2.00	0.47	-1.20
10	28	3	1.91	-1.22	-0.54	25	73	2	1.96	-1.38	-0.59
	29	6	1.93	-0.90	-1.12		74	3	1.98	-1.09	-0.76
	30	8	2.15	-0.69	-0.65		75	9	1.88	-0.81	-1.24
11	31	4	2.11	-0.77	-0.46	26	76	1	2.15	0.04	-1.22
	32	8	2.20	-0.35	-1.20		77	5	2.05	0.39	-1.00
	33	9	1.78	-0.18	-0.91		78	10	2.17	-0.28	-0.46
12	34	6	2.24	-1.16	-0.95	27	79	4	1.93	1.81	-0.48
	35	7	2.05	-0.78	-0.90		80	5	2.16	1.22	-1.27
	36	9	1.94	-0.60	-1.13		81	6	2.13	1.45	-0.87
13	37	2	2.13	-1.10	-0.56	28	82	4	1.85	0.26	-0.58
	38	7	1.99	-0.83	-1.13		83	5	1.83	0.13	-1.25
	39	10	2.11	-0.53	-0.45		84	9	1.84	0.71	-0.81
14	40	1	2.09	1.62	-0.62	29	85	2	2.04	1.94	-1.22
	41	9	2.13	1.91	-0.91		86	3	2.09	1.61	-0.86
	42	10	1.80	1.40	-1.31		87	4	2.16	1.33	-0.41
15	43	1	1.90	-1.85	-0.73	30	88	2	2.00	0.34	-0.84
	44	2	2.18	-1.39	-1.30		89	7	1.88	0.66	-1.20
	45	7	2.17	-1.31	-1.39		90	8	2.03	0.89	-0.47

Table B4

Test Specification for the 10-Dimension Triplet Test with High Alpha and High Delta SD Condition

			Statement Parameters						Statement Parameters		
Item Block #	Statement	Dimension	Alpha	Delta	Tau	Item Block #	Statement	Dimension	Alpha	Delta	Tau
1	1	1	1.81	0.60	-1.37	16	46	4	1.95	0.35	-1.14
	2	2	1.85	-1.89	-0.98		47	6	1.83	1.56	-0.56
	3	3	2.00	-1.40	-0.55		48	10	2.17	-1.08	-0.69
2	4	1	1.83	-1.23	-0.55	17	49	5	2.05	1.35	-1.05
	5	6	2.12	-0.26	-0.80		50	8	1.92	-0.29	-0.68
	6	9	2.11	1.37	-1.28		51	10	1.98	-1.25	-1.12
3	7	4	2.19	-0.69	-1.22	18	52	3	2.06	-0.10	-0.97
	8	5	2.22	0.74	-0.67		53	4	1.92	-1.91	-1.27
	9	8	1.84	1.89	-0.54		54	5	1.82	0.63	-0.43
4	10	6	1.94	-0.60	-1.17	19	55	1	2.00	1.25	-0.73
	11	7	1.77	1.36	-0.99		56	7	2.19	0.91	-1.21
	12	8	1.88	1.87	-0.67		57	10	1.89	-1.18	-0.48
5	13	2	1.84	1.16	-0.99	20	58	1	2.01	-1.63	-0.80
	14	6	2.20	1.96	-1.24		59	3	2.13	-0.68	-1.13
	15	7	2.12	-0.59	-0.53		60	5	2.15	0.95	-0.41
6	16	8	2.09	-1.63	-1.04	21	61	5	1.96	-1.15	-0.84
	17	9	1.82	0.99	-0.94		62	6	1.97	1.45	-0.57
	18	10	1.94	-0.41	-0.79		63	7	1.81	0.26	-1.05
7	19	2	1.93	0.57	-1.29	22	64	3	1.85	1.91	-0.84
	20	3	1.76	1.92	-0.94		65	4	1.88	1.23	-1.12
	21	5	1.86	-0.70	-0.48		66	6	2.08	-0.61	-0.51
8	22	1	2.23	1.08	-0.69	23	67	7	1.76	-1.47	-0.95
	23	4	1.83	-1.37	-1.31		68	8	1.92	-1.72	-1.33
	24	9	1.82	-1.25	-0.72		69	9	1.92	0.65	-0.80
9	25	1	1.93	-0.97	-1.31	24	70	3	1.76	-1.87	-0.85
	26	2	1.94	0.69	-0.86		71	8	1.75	0.18	-0.66
	27	10	2.23	1.60	-0.51		72	10	2.00	0.54	-1.20
10	28	3	1.91	-1.09	-0.54	25	73	2	1.96	-1.73	-0.59
	29	6	1.93	-1.54	-1.12		74	3	1.98	0.85	-0.76
	30	8	2.15	0.98	-0.65		75	9	1.88	-0.69	-1.24
11	31	4	2.11	1.45	-0.46	26	76	1	2.15	0.04	-1.22
	32	8	2.20	0.11	-1.20		77	5	2.05	-0.97	-1.00
	33	9	1.78	-1.16	-0.91		78	10	2.17	1.71	-0.46
12	34	6	2.24	-0.49	-0.95	27	79	4	1.93	1.18	-0.48
	35	7	2.05	1.89	-0.90		80	5	2.16	-1.33	-1.27
	36	9	1.94	1.62	-1.13		81	6	2.13	-0.64	-0.87
13	37	2	2.13	-1.00	-0.56	28	82	4	1.85	0.37	-0.58
	38	7	1.99	-1.67	-1.13		83	5	1.83	0.97	-1.25
	39	10	2.11	0.85	-0.45		84	9	1.84	-1.55	-0.81
14	40	1	2.09	1.91	-0.62	29	85	2	2.04	1.94	-1.22
	41	9	2.13	0.81	-0.91		86	3	2.09	0.82	-0.86
	42	10	1.80	-0.68	-1.31		87	4	2.16	-0.68	-0.41
15	43	1	1.90	-1.85	-0.73	30	88	2	2.00	0.48	-0.84
	44	2	2.18	0.47	-1.30		89	7	1.88	-1.71	-1.20
	45	7	2.17	0.34	-1.39		90	8	2.03	-1.86	-0.47

Appendix C: Item Parameters for 10-D MFC Pair Tests in Study 2

Table C1

Test Specification for 30-Pairs Test with 10-Dimension

			30-Pair						30-Pair		
Item Block #	Statement	Dimension	Alpha	Delta	Tau	Item Block #	Statement	Dimension	Alpha	Delta	Tau
1	1	1	1.81	0.6	-1.37	16	31	6	1.83	1.56	-0.56
	2	2	1.85	-1.89	-0.98		32	10	2.17	-1.08	-0.69
2	3	1	1.83	-1.23	-0.55	17	33	8	1.92	-0.29	-0.68
	4	9	2.11	1.37	-1.28		34	10	1.98	-1.25	-1.12
3	5	5	2.22	0.74	-0.67	18	35	3	2.06	-0.1	-0.97
	6	8	1.84	1.89	-0.54		36	5	1.82	0.63	-0.43
4	7	6	1.94	-0.6	-1.17	19	37	1	2.00	1.25	-0.73
	8	7	1.77	1.36	-0.99		38	7	2.19	0.91	-1.21
5	9	2	1.84	1.16	-0.99	20	39	1	2.01	-1.63	-0.80
	10	7	2.12	-0.59	-0.53		40	3	2.13	-0.68	-1.13
6	11	8	2.09	-1.63	-1.04	21	41	5	1.96	-1.15	-0.84
	12	9	1.82	0.99	-0.94		42	6	1.97	1.45	-0.57
7	13	3	1.76	1.92	-0.94	22	43	4	1.88	1.23	-1.12
	14	5	1.86	-0.7	-0.48		44	6	2.08	-0.61	-0.51
8	15	1	2.23	1.08	-0.69	23	45	7	1.76	-1.47	-0.95
	16	4	1.83	-1.37	-1.31		46	8	1.92	-1.72	-1.33
9	17	2	1.94	0.69	-0.86	24	47	3	1.76	-1.87	-0.85
	18	10	2.23	1.6	-0.51		48	8	1.75	0.18	-0.66
10	19	6	1.93	-1.54	-1.12	25	49	3	1.98	0.85	-0.76
	20	8	2.15	0.98	-0.65		50	9	1.88	-0.69	-1.24
11	21	4	2.11	1.45	-0.46	26	51	1	2.15	0.04	-1.22
	22	9	1.78	-1.16	-0.91		52	10	2.17	1.71	-0.46
12	23	6	2.24	-0.49	-0.95	27	53	4	1.93	1.18	-0.48
	24	9	1.94	1.62	-1.13		54	5	2.16	-1.33	-1.27
13	25	2	2.13	-1	-0.56	28	55	4	1.85	0.37	-0.58
	26	10	2.11	0.85	-0.45		56	5	1.83	0.97	-1.25
14	27	9	2.13	0.81	-0.91	29	57	3	2.09	0.82	-0.86
	28	10	1.80	-0.68	-1.31		58	4	2.16	-0.68	-0.41
15	29	2	2.18	0.47	-1.30	30	59	2	2.00	0.48	-0.84
	30	7	2.17	0.34	-1.39		60	7	1.88	-1.71	-1.20

Table C2

Test Specification for 90-Pairs Test with 10-Dimension

Item Block #	Statement	Dimension	90-Pair Alpha	Delta	Tau
1	1	1	1.81	0.6	-1.37
	2	2	1.85	-1.89	-0.98
2	3	1	1.81	0.6	-1.37
	4	3	2.00	-1.4	-0.55
3	5	2	1.85	-1.89	-0.98
	6	3	2.00	-1.4	-0.55
4	7	1	1.83	-1.23	-0.55
	8	6	2.12	-0.26	-0.80
5	9	1	1.83	-1.23	-0.55
	10	9	2.11	1.37	-1.28
6	11	6	2.12	-0.26	-0.80
	12	9	2.11	1.37	-1.28
7	13	4	2.19	-0.69	-1.22
	14	5	2.22	0.74	-0.67
8	15	4	2.19	-0.69	-1.22
	16	8	1.84	1.89	-0.54
9	17	5	2.22	0.74	-0.67
	18	8	1.84	1.89	-0.54
10	19	6	1.94	-0.6	-1.17
	20	7	1.77	1.36	-0.99
11	21	6	1.94	-0.6	-1.17
	22	8	1.88	1.87	-0.67
12	23	7	1.77	1.36	-0.99
	24	8	1.88	1.87	-0.67
13	25	2	1.84	1.16	-0.99
	26	6	2.20	1.96	-1.24
14	27	2	1.84	1.16	-0.99
	28	7	2.12	-0.59	-0.53
15	29	6	2.20	1.96	-1.24
	30	7	2.12	-0.59	-0.53

Table C2 (Continued)

Item Block #	Statement	Dimension	90-Pair		
			Alpha	Delta	Tau
16	31	8	2.09	-1.63	-1.04
	32	9	1.82	0.99	-0.94
17	33	8	2.09	-1.63	-1.04
	34	10	1.94	-0.41	-0.79
18	35	9	1.82	0.99	-0.94
	36	10	1.94	-0.41	-0.79
19	37	2	1.93	0.57	-1.29
	38	3	1.76	1.92	-0.94
20	39	2	1.93	0.57	-1.29
	40	5	1.86	-0.7	-0.48
21	41	3	1.76	1.92	-0.94
	42	5	1.86	-0.7	-0.48
22	43	1	2.23	1.08	-0.69
	44	4	1.83	-1.37	-1.31
23	45	1	2.23	1.08	-0.69
	46	9	1.82	-1.25	-0.72
24	47	4	1.83	-1.37	-1.31
	48	9	1.82	-1.25	-0.72
25	49	1	1.93	-0.97	-1.31
	50	2	1.94	0.69	-0.86
26	51	1	1.93	-0.97	-1.31
	52	10	2.23	1.6	-0.51
27	53	2	1.94	0.69	-0.86
	54	10	2.23	1.6	-0.51
28	55	3	1.91	-1.09	-0.54
	56	6	1.93	-1.54	-1.12
29	57	3	1.91	-1.09	-0.54
	58	8	2.15	0.98	-0.65
30	59	6	1.93	-1.54	-1.12
	60	8	2.15	0.98	-0.65

Table C2 (Continued)

			90-Pair		
Item Block #	Statement	Dimension	Alpha	Delta	Tau
31	61	4	2.11	1.45	-0.46
	62	8	2.20	0.11	-1.20
32	63	4	2.11	1.45	-0.46
	64	9	1.78	-1.16	-0.91
33	65	8	2.20	0.11	-1.20
	66	9	1.78	-1.16	-0.91
34	67	6	2.24	-0.49	-0.95
	68	7	2.05	1.89	-0.90
35	69	6	2.24	-0.49	-0.95
	70	9	1.94	1.62	-1.13
36	71	7	2.05	1.89	-0.90
	72	9	1.94	1.62	-1.13
37	73	2	2.13	-1	-0.56
	74	7	1.99	-1.67	-1.13
38	75	2	2.13	-1	-0.56
	76	10	2.11	0.85	-0.45
39	77	7	1.99	-1.67	-1.13
	78	10	2.11	0.85	-0.45
40	79	1	2.09	1.91	-0.62
	80	9	2.13	0.81	-0.91
41	81	1	2.09	1.91	-0.62
	82	10	1.80	-0.68	-1.31
42	83	9	2.13	0.81	-0.91
	84	10	1.80	-0.68	-1.31
43	85	1	1.90	-1.85	-0.73
	86	2	2.18	0.47	-1.30
44	87	1	1.90	-1.85	-0.73
	88	7	2.17	0.34	-1.39
45	89	2	2.18	0.47	-1.30
	90	7	2.17	0.34	-1.39

Table C2 (Continued)

			90-Pair		
Item Block #	Statement	Dimension	Alpha	Delta	Tau
46	91	4	1.95	0.35	-1.14
	92	6	1.83	1.56	-0.56
47	93	4	1.95	0.35	-1.14
	94	10	2.17	-1.08	-0.69
48	95	6	1.83	1.56	-0.56
	96	10	2.17	-1.08	-0.69
49	97	5	2.05	1.35	-1.05
	98	8	1.92	-0.29	-0.68
50	99	5	2.05	1.35	-1.05
	100	10	1.98	-1.25	-1.12
51	101	8	1.92	-0.29	-0.68
	102	10	1.98	-1.25	-1.12
52	103	3	2.06	-0.1	-0.97
	104	4	1.92	-1.91	-1.27
53	105	3	2.06	-0.1	-0.97
	106	5	1.82	0.63	-0.43
54	107	4	1.92	-1.91	-1.27
	108	5	1.82	0.63	-0.43
55	109	1	2.00	1.25	-0.73
	110	7	2.19	0.91	-1.21
56	111	1	2.00	1.25	-0.73
	112	10	1.89	-1.18	-0.48
57	113	7	2.19	0.91	-1.21
	114	10	1.89	-1.18	-0.48
58	115	1	2.01	-1.63	-0.80
	116	3	2.13	-0.68	-1.13
59	117	1	2.01	-1.63	-0.80
	118	5	2.15	0.95	-0.41
60	119	3	2.13	-0.68	-1.13
	120	5	2.15	0.95	-0.41

Table C2 (Continued)

			90-Pair		
Item Block #	Statement	Dimension	Alpha	Delta	Tau
61	121	5	1.96	-1.15	-0.84
	122	6	1.97	1.45	-0.57
62	123	5	1.96	-1.15	-0.84
	124	6	1.97	1.45	-0.57
63	125	6	1.97	1.45	-0.57
	126	7	1.81	0.26	-1.05
64	127	3	1.85	1.91	-0.84
	128	4	1.88	1.23	-1.12
65	129	3	1.85	1.91	-0.84
	130	6	2.08	-0.61	-0.51
66	131	4	1.88	1.23	-1.12
	132	6	2.08	-0.61	-0.51
67	133	7	1.76	-1.47	-0.95
	134	8	1.92	-1.72	-1.33
68	135	7	1.76	-1.47	-0.95
	136	9	1.92	0.65	-0.80
69	137	8	1.92	-1.72	-1.33
	138	9	1.92	0.65	-0.80
70	139	3	1.76	-1.87	-0.85
	140	8	1.75	0.18	-0.66
71	141	3	1.76	-1.87	-0.85
	142	10	2.00	0.54	-1.20
72	143	8	1.75	0.18	-0.66
	144	10	2.00	0.54	-1.20
73	145	2	1.96	-1.73	-0.59
	146	3	1.98	0.85	-0.76
74	147	2	1.96	-1.73	-0.59
	148	9	1.88	-0.69	-1.24
75	149	3	1.98	0.85	-0.76
	150	9	1.88	-0.69	-1.24

Table C2 (Continued)

			90-Pair		
Item Block #	Statement	Dimension	Alpha	Delta	Tau
76	151	1	2.15	0.04	-1.22
	152	5	2.05	-0.97	-1.00
77	153	1	2.15	0.04	-1.22
	154	10	2.17	1.71	-0.46
78	155	5	2.05	-0.97	-1.00
	156	10	2.17	1.71	-0.46
79	157	4	1.93	1.18	-0.48
	158	5	2.16	-1.33	-1.27
80	159	4	1.93	1.18	-0.48
	160	6	2.13	-0.64	-0.87
81	161	5	2.16	-1.33	-1.27
	162	6	2.13	-0.64	-0.87
82	163	4	1.85	0.37	-0.58
	164	5	1.83	0.97	-1.25
83	165	4	1.85	0.37	-0.58
	166	9	1.84	-1.55	-0.81
84	167	5	1.83	0.97	-1.25
	168	9	1.84	-1.55	-0.81
85	169	2	2.04	1.94	-1.22
	170	3	2.09	0.82	-0.86
86	171	2	2.04	1.94	-1.22
	172	4	2.16	-0.68	-0.41
87	173	3	2.09	0.82	-0.86
	174	4	2.16	-0.68	-0.41
88	175	2	2.00	0.48	-0.84
	176	7	1.88	-1.71	-1.20
89	177	2	2.00	0.48	-0.84
	178	8	2.03	-1.86	-0.47
90	179	7	1.88	-1.71	-1.20
	180	8	2.03	-1.86	-0.47

Appendix D1: Single-Statement Personality Items

Below are sixty statements representing Big 5 personality constructs. Using a 1-5 scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree), indicate your level of agreement with each statement by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

1. Have a good word for everyone.
2. Believe that others have good intentions.
3. Respect others.
4. Accept people as they are.
5. Am concerned about others.
6. Trust what people say.
7. Sympathize with others' feelings.
8. Treat all people equally.
9. Cut others to pieces.
10. Get back at others.
11. Contradict others.
12. Am out for my own personal gain.
13. Am always prepared.
14. Get chores done right away.
15. Carry out my plans.
16. Complete tasks successfully.
17. Do things according to a plan.
18. Am exacting in my work.
19. Finish what l start.
20. Follow through with my plans.
21. Waste my time.
22. Find it difficult to get down to work.
23. Don't put my mind on the task at hand.
24. Need a push to get started.
25. Feel comfortable around people.
26. Make friends easily.
27. Am the life of the party.
28. Know how to captivate people.
29. Start conversations.
30. Warm up quickly to others.
31. Talk to a lot of different people at parties.
32. Cheer people up.
33. Keep in the background.
34. Would describe my experiences as somewhat dull.
35. Don't like to draw attention to myself.

36. Find it difficult to approach others.
37. Often feel blue.
38. Dislike myself.
39. Am often down in the dumps
40. Have frequent mood swings.
41. Feel threatened easily.
42. Seldom feel blue.
43. Feel comfortable with myself.
44. Rarely get irritated.
45. Am not easily bothered by things.
46. Am very pleased with myself.
47. Am relaxed most of the time.
48. Am not easily frustrated.
49. Believe in the importance of art.
50. Have a vivid imagination.
51. Carry the conversation to a higher level.
52. Enjoy thinking about things.
53. Enjoy wild flights of fantasy.
54. Get excited by new ideas.
55. Have a rich vocabulary.
56. Am not interested in abstract ideas.
57. Do not like art.
58. Do not enjoy going to art museums.
59. Rarely look for a deeper meaning in things.
60. Am not interested in theoretical discussions.

Appendix D2: Triplet Multidimensional Forced Choice Personality Measure

Below are 20 triplets measuring Big 5 personality constructs. Please rank the following statements on a scale of 1, 2, 3 from “most like me” to “least like me”. Please be open and honest in your responding.

	MFC Items	RANK
1	Respect others. Have a rich vocabulary. Follow through with my plans.	
2	Get excited by new ideas. Warm up quickly to others. Do not put my mind on the task at hand.	
3	Feel comfortable with myself. Do not enjoy going to art museums. Keep in the background.	
4	Am always prepared. Accept people as they are. Seldom feel blue.	
5	Am the life of the party. Am out for my own personal gain. Often feel blue.	
6	Rarely look for a deeper meaning in things. Cheer people up. Carry out my plans.	
7	Rarely get irritated. Am not interested in abstract ideas. Know how to captivate people.	
8	Have a good word for everyone. Am exacting in my work.	

	Have a vivid imagination.	
9	Do things according to a plan. Cut others to pieces. Feel comfortable around people.	
10	Find it difficult to get down to work. Am relaxed most of the time. Enjoy thinking about things.	
11	Treat all people equally. Would describe my experiences as somewhat dull. Am often down in the dumps	
12	Waste my time. Find it difficult to approach others. Trust what people say.	
13	Am concerned about others. Believe in the importance of art. Feel threatened easily.	
14	Complete tasks successfully. Do not like to draw attention to myself. Have frequent mood swings.	
15	Am not easily bothered by things. Contradict others. Carry the conversation to a higher level.	
16	Am not interested in theoretical discussions. Talk to a lot of different people at parties. Dislike myself.	
17	Finish what l start. Get back at others. Am not easily frustrated.	
18	Enjoy wild flights of fantasy. Get chores done right away. Sympathize with others' feelings.	
19	Believe that others have good intentions.	

	Am very pleased with myself. Make friends easily.	
20	Do not like art. Start conversations. Need a push to get started.	

Appendix D3: O*NET Interest Profiler Short Form

Rounds, J., Su, R., Lewis, P., & Rivkin, D. (2010). O*NET Interest Profiler Short Form Psychometric Characteristics: Summary, The National Center for O*NET Development.

Below are sixty statements representing vocational interest constructs. Using a 1-5 scale below, indicate your level of agreement with each statement by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

1 = Dislike, 2 = Slightly dislike, 3 = Neither like not dislike, 4 = Slightly enjoy, 5 = Enjoy

1. Build kitchen cabinets
2. Lay brick or tile
3. Repair household appliances
4. Raise fish in a fish hatchery
5. Assemble electronic parts
6. Drive a truck to deliver packages to offices and homes
7. Test the quality of parts before shipment
8. Repair and install locks
9. Set up and operate machines to make products
10. Put out forest fires
11. Develop a new medicine
12. Study ways to reduce water pollution
13. Conduct chemical experiments
14. Study the movement of planets
15. Examine blood samples using a microscope
16. Investigate the cause of a fire
17. Develop a way to better predict the weather
18. Work in a biology lab
19. Invent a replacement for sugar
20. Do laboratory tests to identify diseases
21. Write books or plays
22. Play a musical instrument
23. Compose or arrange music
24. Draw pictures

25. Create special effects for movies
26. Paint sets for plays
27. Write scripts for movies or television shows
28. Perform jazz or tap dance
29. Sing in a band
30. Edit movies
31. Teach an individual an exercise routine
32. Help people with personal or emotional problems
33. Give career guidance to people
34. Perform rehabilitation therapy
35. Do volunteer work at a non-profit organization
36. Teach children how to play sports
37. Teach sign language to people with hearing disabilities
38. Help conduct a group therapy session
39. Take care of children at a day-care center
40. Teach a high-school class
41. Buy and sell stocks and bonds
42. Manage a retail store
43. Operate a beauty salon or barber shop
44. Manage a department within a large company
45. Start your own business
46. Negotiate business contracts
47. Represent a client in a lawsuit
48. Market a new line of clothing
49. Sell merchandise at a department store
50. Manage a clothing store
51. Develop a spreadsheet using computer software
52. Proofread records or forms
53. Load computer software into a large computer network
54. Operate a calculator
55. Keep shipping and receiving records
56. Calculate the wages of employees
57. Inventory supplies using a hand-held computer
58. Record rent payments
59. Keep inventory records
60. Stamp, sort, and distribute mail for an organization

Appendix D4: Satisfaction With Life Scale

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. Journal of personality assessment, 49(1), 71-75.

Below are five statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

1 = Strongly disagree, 2 = Disagree, 3 = Slightly disagree, 4 = Neither agree nor disagree,
5 = Slightly agree, 6 = Agree, 7 = Strongly agree

1. In most ways my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with my life.
4. So far I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing.

Appendix D5 : Buss-Perry Aggression Questionnaire

Bryant, F. B., & Smith, B. D. (2001). Refining the architecture of aggression: a measurement model for the Buss-Perry Aggression Questionnaire. Journal of Research in Personality, 35, 138–167.

Below are twelve statements that you may agree or disagree with. Using the 1 - 7 scale below, indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

1 = Strongly disagree, 2 = Disagree, 3 = Slightly disagree, 4 = Neither agree nor disagree, 5 = Slightly agree

1. Given enough provocation, I may hit another person.
2. There are people who pushed me so far that we came to blows.
3. I have threatened people I know.
4. I often find myself disagreeing with people.
5. I can't help getting into arguments when people disagree with me.
6. My friends say that I'm somewhat argumentative.
7. I flare up quickly but get over it quickly.
8. Sometimes I fly off the handle for no good reason.
9. I have trouble controlling my temper.
10. At times I feel I have gotten a raw deal out of life.
11. Other people always seem to get the breaks.
12. I wonder why sometimes I feel so bitter about things.

Appendix D6: SPANE Questionnaire

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. Social Indicators Research, 97(2), 143-156.

Please think about what you have been doing and experiencing during the past 4 weeks. Then report how much you experienced each of the following feelings, using the scale below. For each item, select a number from 1 to 5, and indicate that number on your response sheet.

1 = Very rarely or never, 2 = Rarely, 3 = Sometimes,
4 = Often, 5 = Very often or always

1. Joyful
2. Happy
3. Comfortable
4. Contented
5. Pleasant
6. Positive
7. Good
8. Negative
9. Irritated
10. Helpless
11. Sad
12. Unpleasant
13. Unpleasant
14. Bad
15. Afraid
16. Angry