October 2021

Comparison of Parameter Estimation Approaches for Multi-Unidimensional Pairwise Preference Tests

Naidan Tu
University of South Florida

Comparison of Parameter Estimation Approaches for Multi-Unidimensional Pairwise Preference Tests

by

Naidan Tu

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Psychology
College of Arts and Sciences
University of South Florida

Major Professor: Stephen Stark, Ph.D.
Seang-Hwane Joo, Ph.D.
Brenton M. Wiernik, Ph.D.
Marina A. Bornovalova, Ph.D.

Date of Approval:
September 20, 2021

Keywords: Markov chain Monte Carlo (MCMC), Item Response Theory (IRT), Generalized Graded Unfolding Model (GGUM), Forced Choice

DEDICATION

I would like to express my genuine gratitude to my advisor, Dr. Stephen Stark, for trusting in my capabilities and guiding me as I am learning. With his mentorship, I have been able to continuingly accumulate knowledge and develop expertise. I would also like to thank Dr. Seang-Hwane Joo and Dr. Philseok Lee for their tremendous contribution to the project. I am also grateful to my other committee members, Dr. Brenton M. Wiernik and Dr. Marina A. Bornovalova for their helpful feedback. Last but not least, a unique thanks to my family and friends for their endless care and support.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

Multidimensional forced choice (MFC) testing has been proposed as a way of reducing response biases in noncognitive measurement. Although early item response theory (IRT) research focused on illustrating that trait scores with normative properties could be obtained using various MFC models and formats, more recent attention has been devoted to exploring the processes involved in test construction and how that influences MFC scores. This research compared two approaches for estimating Multi-Unidimensional Pairwise Preference model (MUPP; Stark et al., 2005) parameters based on the Generalized Graded Unfolding Model (GGUM; Roberts et al., 2000). More specifically, we compared the efficacy of statement and person parameter estimation based on a "two-step" process, developed by Stark et al. (2005) with a more recently developed "direct" estimation approach (Lee et al., 2019) in a Monte Carlo study that also manipulated test length, test dimensionality, sample size, and the correlations between generating thetas for each dimension. Results indicated that the two approaches had similar scoring accuracy, although the two-step approach had better statement parameter recovery than the direct approach. Implications, limitations, and recommendations for future MFC research and practice are discussed.

CHAPTER ONE:

INTRODUCTION

Past research on multidimensional forced choice (MFC) testing has focused on developing psychometric methods that yield scores with normative properties, resistance to response biases such as faking, and validity for predicting organizational outcomes. Models such as the Multi-Unidimensional Pairwise Preference model (MUPP; Stark et al., 2005) and Thurstonian IRT model (TIRT; Brown & Maydeu-Olivares, 2011) have been shown to produce scores with normative properties (Brown et al., 2011; Joo et al., 2019; Lee et al., 2019, Stark et al., 2012). Meta-analysis (Cao & Drasgow, 2019) and a recent primary study (Wetzel et al., 2020) have shown that MFC measures are less susceptible to faking than rating scale measures, and MFC measures have criterion validity similar to rating scale measures in research contexts (Wetzel & Frick, 2019; Zhang et al., 2020). And MUPP personality tests, in particular, have been shown to predict citizenship behaviors, counterproductive work behaviors, and attrition in personnel screening environments (Drasgow et al., 2012; Stark et al., 2014) and utility for job classification (Nye et al., 2020).

As MFC item response theory (IRT) research is now well into its second decade and the number of applications has increased, more attention is being devoted to improving testing efficiency (Joo et al., 2019), differential item functioning detection (Lee et al., 2020), and, importantly, parameter estimation approaches that are fundamental to all applications (Lee et al., 2019). This research focuses specifically on the latter. Accurately estimating statement and person parameters is the first, and probably most important, step in applying MFC models.

Different estimation approaches have been proposed to analyze noncognitive responses. Models such as the TIRT MFC model are based on the dominance model assumption that the relationship between latent trait level and the probability of endorsing or agreeing with a noncognitive statement is monotonic. Other models, such as the MUPP, allow one to use either a dominance or ideal point model to compute the probability of agreeing with statements composing MFC items and, accordingly, the preferential choice probabilities. Ideal point models assume that the relationship between latent trait level and the probability of agreeing with a noncognitive statement may be nonmonotonic, and Stark et al. (2005) recommended the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) for MUPP applications because it was found to fit rating scale personality data as good or better than popular dominance models (Stark et al., 2006). Other researchers have since found good fit for the GGUM to vocational interests (Tay et al., 2009), emotional intelligence (Cho et al., 2015), job satisfaction (Carter & Dalal, 2010), and attachment style (Sun, 2017) data. However, there is a continuing need for research exploring the efficacy of parameter estimation with ideal point models. This study specifically aims to advance MFC applications by focusing on the relative efficacy of two approaches to MUPP model parameter estimation based on the GGUM.

Two approaches have been proposed to estimate MUPP model parameters. Stark and colleagues (2005) described a two-step approach to MUPP test construction and scoring that begins with estimating GGUM (Roberts et al., 2000) parameters for a pool of noncognitive statements reflecting low, medium, and high levels of each trait of interest. The statements are administered using a four-point response rating scale format (Strongly Disagree, Disagree, Agree, Strongly Agree) to a large sample of examinees in a pretest study with instructions to answer as honestly and accurately as possible. The “honest” responses are dichotomized and calibrated, one

trait at a time. For high-stakes uses, social desirability ratings are also obtained for the pool of statements in a "fake-good" study, or by gathering subject matter expert (SME) judgments; the mean self- or SME rating for each statement serves as a social desirability parameter estimate. MUPP tests are then constructed for assessment purposes by pairing statements, similar in extremity and social desirability, using a table of test design specifications that delineates the combinations of dimensions that will compose the forced choice items. The preponderance of these pairings should be multidimensional to enhance resistance to faking, and care should be taken to ensure each item provides adequate IRT information for scoring. Finally, the MUPP tests are administered for assessment purposes, and the forced choice response data are scored using, for example, a multidimensional Bayes modal method (Stark et al., 2005) or some recent alternatives (e.g., Guan, Sun, & Carter, 2021; Lee et al., 2019).

As an alternative, Lee et al. (2019) developed a Markov chain Monte Carlo (MCMC) method for estimating MUPP statement and person parameters directly from forced choice responses. Unlike the earlier two-step approach that is advantageous for CAT, because it allows any number of pairwise preference tests to be constructed dynamically after a statement pool has been calibrated (Stark et al., 2012), this "direct" approach takes into account the potential interplay of statements composing the administered pairwise preference items, and it is arguably better suited for building parallel test forms and differential item functioning analysis. The purpose of this research was therefore to investigate the overall efficacy and comparability of these two approaches to MUPP statement and person parameter estimation through a Monte Carlo simulation. Sample size, test dimensionality, number of items (pairs) per dimension, and the correlations among dimensions (i.e., latent trait correlations) were manipulated to explore estimation accuracy and precision in a range of realistic experimental conditions. The next

sections of this paper present the MUPP and GGUM models, the Monte Carlo study details, the hypotheses and research questions, and the statistical analysis of the simulation results. Results are then summarized, and the implications are discussed.

The MUPP model

The MUPP model (Stark, 2002; Stark et al., 2005) assumes that when a respondent is presented with a pair of statements (*s* and *t*) and is asked to select the statement in each pair that is more "like me", the respondent evaluates each statement separately until a preference is reached. The preferential choice decision is operationalized as agreeing with one statement and disagreeing with the other. These agree/disagree joint probabilities depend on the respondent's trait levels and the statement parameters estimated using an appropriate unidimensional IRT model. Mathematically, the probability of preferring statement *s* to statement *t* in an item is defined as:

$$P_{(s>t)_i}\left(\theta_{d_s}, \theta_{d_t}\right) = \frac{P_{st}\{1,0\}}{P_{st}\{1,0\}+P_{st}\{0,1\}} = \frac{P_s\{1\}P_t\{0\}}{P_s\{1\}P_t\{0\}+P_s\{0\}P_t\{1\}} \quad (1)$$

where *i* represents the i^{th} item, θ_{d_s} and θ_{d_t} are the latent trait values for a respondent on dimension d_s and d_t, respectively, $P_{st}\{1,0\}$ is the joint probability of endorsing statement s and not endorsing statement t, $P_{st}\{0,1\}$ is the joint probability of endorsing statement t and not endorsing statement s, $P_s\{1\}$ and $P_t\{1\}$ are the conditional probabilities of endorsing statement *s* and *t*, respectively, $P_s\{0\}$ and $P_t\{0\}$ are the conditional probabilities of not endorsing statement *s* and *t*, respectively. To compute $P_s\{1\}$, $P_t\{1\}$, $P_s\{0\}$, $P_t\{0\}$, the dichotomous version of the GGUM was used for this study, in accordance with Stark et al.'s (2005, 2006) recommendations.

The GGUM

The GGUM (Roberts et al., 2000) is an ideal point model that assumes the probability of agreeing with a statement is a function of the distance between the statement and the respondent on the latent trait continuum. The closer the statement is to the respondent, the more likely the

respondent will agree. GGUM is often used in noncognitive assessment and can be applied to both dichotomous and polytomous responses. In the dichotomous version, the probabilities of agreeing (Z=1) and disagreeing (Z=0) with a statement are:

$$P(1) = (Z = 1|\theta) = \frac{\exp(\alpha[(\theta-\delta)-\tau])+\exp(\alpha[2(\theta-\delta)-\tau])}{\gamma} \quad (2.1)$$

$$P(0) = P(Z = 0|\theta) = \frac{1+\exp(\alpha[3(\theta-\delta)])}{\gamma} \quad (2.2)$$

$$\gamma = 1 + \exp(\alpha[3(\theta - \delta)]) + \exp(\alpha[(\theta - \delta) - \tau]) + \exp(\alpha[2(\theta - \delta) - \tau])$$

where θ is the latent trait value of a respondent, α, δ, and τ are the statement discrimination, location, and threshold parameters, and γ is the sum of the numerators in Equations (2.1) and (2.2).

CHAPTER TWO:

METHOD

This study investigated the recovery and comparability of MUPP statement and person parameters using the two-step and direct approaches described above. Although the comparability of statement parameters is of interest, the main question for applied purposes concerns the similarity of person parameter estimates (i.e., the trait scores). The simulation conditions in this study were based on previous simulation research involving the MUPP model (Stark et al., 2005; Stark et al., 2012; Joo et al., 2018; Lee et al., 2019) as well as conditions that are being explored in field research and practice.

Simulation study design

Five independent variables were manipulated: (1) sample size (N=400, N=800), (2) number of dimensions in the MUPP tests (6D, 12D), (3) number of items (pairs) per dimension (4, 6, 8), (4) estimation approach (direct, two-step), and (5) the correlations between generating thetas for each dimension (.00, .30). Sample size has been explored in several studies involving GGUM and MUPP estimation, and research has shown that at least 400 respondents are needed for reasonably accurate statement parameter estimates (e.g., de la Torre et al., 2006; Joo et al., 2017; Roberts et al., 2000; Stark et al., 2005). 6D and 12D tests were used to explore the effect of dimensionality, reflective of some widely used MFC vocational interest (Wang et al., 2017) and personality tests (Aon, 2015; Stark et al., 2014; White & Young, 1998). The effect of test length was explored, as in previous studies, by manipulating the number of items (pairs) per dimension (Stark et al., 2005; Stark et al., 2012). Total test length was determined by multiplying the

number of dimensions and the number of items per dimension (IPD) to keep the ratio of test length to dimensionality constant. For example, 6D tests of 4, 6, and 8 IPD resulted in test lengths of 24, 36, and 48 items, and similarly configured 12D tests had 48, 72, and 96 items, respectively. This approach has been used to ensure that each dimension (i.e., measured construct) is represented the same number of times with tests of different dimensionality. Whereas early simulation research explored nonadaptive testing with 10, 20, and 40 IPD (e.g., Stark et al., 2005, 2012), more recent studies have found adequate trait estimation with "shorter" (fewer IPD) measures (e.g., Lee et al., 2019). This research therefore systematically explored trait estimation with measures considerably fewer IPD than originally recommended.

Test design

Statement parameter generation. For comparison with previous research focusing on MUPP direct estimation, statement parameters were selectively drawn from Lee et al. (2019). In that study, location parameters (δ) were sampled from a combination of uniform distributions, consistent with U[-2, 2]; threshold parameters (τ) were sampled from U[-1.4, - .40]; and discrimination parameters (α) were sampled from distributions reflecting low (U[.75, 1.25]) and high discrimination (U[1.75, 2.25]). Because comparing estimation in low and high discrimination conditions was not a goal of this study, some additional parameters reflecting moderate discrimination were sampled from U[1.25, 1.75]. These statement parameters were tabulated and used to develop MUPP tests with the intended design specifications for each experimental condition.

Test specifications. 6D tests having 4 IPD (6D/4) and 6D tests having 6 IPD (6D/6) were created by systematically selecting from the pool of generating statement parameters such that the mean and standard deviation of the discrimination (α), location (δ), and threshold (τ) parameters were

approximately equal for every dimension measured in every test. Then 6D tests having 8 IPD (6D/8) were created by duplicating the 6D/4 test parameters and modifying the associated dimension numbers to create new items with the same psychometric properties for the longer test. Finally, 12D tests of 4, 6, and 8 IPD (12D/4, 12D/6, 12D/8) were created by duplicating the respective 6D test parameters and modifying the dimension numbers to reflect the higher dimensionality. As noted above, care was taken to ensure each dimension would be measured similarly well within a test by requiring each dimension to be represented the same number of times.

Tables 1-3 present detailed test specifications for the 6D/4, 6D/6, and 6D/8 MUPP tests, and Table 4 shows the corresponding means and standard deviations of the statement parameters. In Tables 1-3, the columns labeled *s* and *t* indicate the dimension(s) represented in the respective pairwise preference items, and the columns labeled α_s, δ_s, τ_s, α_t, δ_t, and τ_t show the corresponding statement parameters. For illustration, note that in the 24-item 6D/4 test each dimension is represented in 8 different pairwise preference items, and in the 36-item 6D/6 test each dimension is represented in 12 different items. Note also that the means and standard deviations of the respective statement parameters are approximately equal for every dimension in every test, as shown in Table 4. The corresponding specifications for the 12D/4, 12D/6, and 12D/8 tests are shown in Appendix Tables A and B.

Data generation

Person parameter generation. On each replication in each experimental condition, an NxD matrix of latent trait parameters (θ) for the designated sample size and dimensionality was sampled from a multivariate normal distribution with zero means and the correlations between the generating thetas for each dimension set to .00 or .30.

Table 1. Test Specifications for the 6D/4 MUPP Test.

Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t	Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t
1	1	6	1.81	.60	-1.37	.75	-1.78	-1.17	13	1	5	.75	1.82	-1.37	.78	-1.61	-1.12
2	1	3	1.83	-1.99	-.55	1.88	-.69	-1.24	14	4	5	1.80	-.68	-1.31	2.13	-1.03	-.91
3	3	4	.81	-1.49	-.55	1.31	-1.12	-.41	15	2	4	.97	.56	-.98	1.88	1.23	-1.12
4	2	4	1.77	1.36	-.65	1.83	-1.37	-1.31	16	3	6	2.06	-.10	-.97	1.94	.32	-.51
5	3	4	1.98	.85	-.76	1.51	1.78	-1.28	17	2	5	1.98	-1.25	-1.12	1.92	-.29	-.68
6	3	6	1.33	1.60	-.51	1.25	1.45	-1.12	18	4	5	.82	-1.91	-1.22	1.48	.98	-.65
7	3	6	1.76	1.92	-.94	1.45	-1.08	-.95	19	1	2	2.00	1.25	-.73	1.34	.91	-1.21
8	1	2	1.35	1.08	-.69	1.49	-.59	-.53	20	1	3	2.01	-1.63	-.80	1.52	-.48	-1.13
9	1	5	1.56	-.22	-1.22	1.26	1.35	-.46	21	2	6	1.94	1.93	-1.13	1.97	-1.54	-.94
10	2	6	1.88	-1.71	-1.20	1.83	-.49	-.56	22	4	6	1.85	.37	-.58	2.08	.99	-.69
11	1	4	1.78	-1.16	-.91	2.11	1.45	-.46	23	2	5	1.76	-1.47	-.95	1.92	-1.72	-1.33
12	5	6	1.84	1.89	-.54	1.93	1.87	-.97	24	3	5	1.76	-1.87	-.85	1.75	.18	-1.22

Table 2. Test Specifications for the 6D/6 MUPP Test.

Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t	Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t
1	1	6	1.81	.60	-1.37	.75	-1.78	-1.17	19	1	5	.75	1.82	-1.37	.78	-1.61	-.67
2	1	3	1.83	-1.99	-.55	1.88	-.69	-1.24	20	4	5	1.80	-.68	-1.31	2.13	-1.03	-.91
3	3	4	.81	-1.49	-.55	1.31	-1.12	-.41	21	2	4	.97	.56	-.98	1.88	1.23	-1.12
4	2	4	1.77	1.36	-.99	1.83	-1.37	-1.31	22	3	6	2.06	-.10	-.97	1.94	.32	-.51
5	3	4	1.98	.85	-.76	1.51	1.78	-1.28	23	2	5	1.98	-1.25	-1.12	1.92	-.29	-.68
6	3	6	1.33	1.6	-.51	1.25	1.45	-1.12	24	4	5	.82	-1.91	-1.22	1.48	.98	-.65
7	3	6	1.76	1.92	-.94	1.45	-1.08	-.95	25	1	2	2.00	1.25	-.73	1.34	.91	-1.21
8	1	2	1.35	1.08	-.69	1.49	-.59	-.53	26	1	3	2.01	-1.63	-.80	1.52	-.48	-1.13
9	1	5	1.56	-.22	-1.22	1.26	1.35	-.46	27	2	6	1.94	1.93	-1.13	1.97	-1.54	-.94
10	2	6	1.88	-1.71	-1.20	1.83	-.49	-.56	28	4	6	1.85	.37	-.58	2.08	.99	-.69
11	1	4	1.78	-1.16	-.91	2.11	1.45	-.46	29	2	5	1.76	-1.47	-.95	1.92	-1.72	-1.33
12	5	6	1.84	1.89	-.54	1.93	1.87	-.57	30	3	5	1.76	-1.87	-.85	1.75	.18	-.66
13	5	2	1.84	-.69	-.37	1.88	-.85	-1.2	31	5	1	.87	.56	-1.22	1.81	.01	-.69
14	5	4	1.81	1.60	-1.37	1.88	.04	-1.24	32	3	1	1.76	.37	-.76	.93	.30	-.55
15	4	5	1.80	-1.51	-.55	2.15	-1.59	-1.37	33	6	3	2.17	-.10	-1.17	2.13	-.59	-1.12
16	6	3	1.76	1.62	-.94	1.83	-1.09	-1.31	34	4	2	.76	-.29	-1.01	.89	.40	-.53
17	2	6	1.76	.04	-.94	1.83	-.88	-1.21	35	6	1	.84	-.71	-1.20	1.75	.02	-.68
18	4	2	2.11	1.63	-.76	2.12	.35	-.57	36	3	1	.93	-1.09	-.24	2.15	-.49	-1.37

MUPP response data generation. True statement parameters and true trait scores were used to compute MUPP response probabilities via Equations (1), (2.1), and (2.2). These probabilities were then compared with random uniform numbers. If the computed probability was larger than the random number, the response was recorded as 1 (statement *s* preferred to statement *t*); otherwise it was recorded as 0 (statement *t* preferred to statement *s*).

Estimation approaches

Two-step approach. Dichotomous single-statement (i.e., rating scale) response data were generated based on the GGUM using the true statement parameters and trait scores for each dimension in the pairwise preference tests described above. Specifically, each GGUM response probability was computed using Equation (2.1) and compared with a random uniform number. If the response probability was larger than the random number, then the response was coded as 1 (agree); otherwise 0 (disagree). The generated response data for the statements representing each dimension were then calibrated, one dimension at a time, using an Ox (Doornik, 2009) MCMC program for GGUM estimation developed by Joo et al. (2017). Next, these estimated parameters were treated as "fixed" and used to score the MUPP responses data described above.

Direct approach. Statement and person parameters were estimated directly from the generated MUPP response data using an Ox (Doornik, 2009) MCMC program for GGUM-RANK estimation (Lee et al., 2019) which includes the MUPP model as a special case.

Replication. Due to long runtimes for MCMC estimation in exploratory work, ranging from 1 to 20 hours per replication in the experimental conditions, 20 replications per condition were performed. This number is consistent with recently published studies on MUPP estimation using the direct approach (e.g., Lee et al., 2019).

Table 3. Test Specifications for the 6D/8 MUPP Test.

Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t	Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t
1	1	6	1.81	.60	-1.37	.75	-1.78	-1.17	25	1	5	.75	1.82	-1.37	.78	-1.61	-1.12
2	1	3	1.83	-1.99	-.55	1.88	-.69	-1.24	26	4	5	1.80	-.68	-1.31	2.13	-1.03	-.91
3	3	4	.81	-1.49	-.55	1.31	-1.12	-.41	27	2	4	.97	.56	-.98	1.88	1.23	-1.12
4	2	4	1.77	1.36	-.65	1.83	-1.37	-1.31	28	3	6	2.06	-.1	-.97	1.94	.32	-.51
5	3	4	1.98	.85	-.76	1.51	1.78	-1.28	29	2	5	1.98	-1.25	-1.12	1.92	-.29	-.68
6	3	6	1.33	1.60	-.51	1.25	1.45	-1.12	30	4	5	.82	-1.91	-1.22	1.48	.98	-.65
7	3	6	1.76	1.92	-.94	1.45	-1.08	-.95	31	1	2	2.00	1.25	-.73	1.34	.91	-1.21
8	1	2	1.35	1.08	-.69	1.49	-.59	-.53	32	1	3	2.01	-1.63	-.80	1.52	-.48	-1.13
9	1	5	1.56	-.22	-1.22	1.26	1.35	-.46	33	2	6	1.94	1.93	-1.13	1.97	-1.54	-.94
10	2	6	1.88	-1.71	-1.20	1.83	-.49	-.56	34	4	6	1.85	.37	-.58	2.08	.99	-.69
11	1	4	1.78	-1.16	-.91	2.11	1.45	-.46	35	2	5	1.76	-1.47	-.95	1.92	-1.72	-1.33
12	5	6	1.84	1.89	-.54	1.93	1.87	-.97	36	3	5	1.76	-1.87	-.85	1.75	.18	-1.22
13	5	2	1.81	.60	-1.37	.75	-1.78	-1.17	37	5	1	.75	1.82	-1.37	.78	-1.61	-1.12
14	5	4	1.83	-1.99	-.55	1.88	-.69	-1.24	38	3	1	1.80	-.68	-1.31	2.13	-1.03	-.91
15	4	3	.81	-1.49	-.55	1.31	-1.12	-.41	39	6	3	.97	.56	-.98	1.88	1.23	-1.12
16	6	3	1.77	1.36	-.65	1.83	-1.37	-1.31	40	4	3	2.06	-.10	-.97	1.94	.32	-.51
17	4	3	1.98	.85	-.76	1.51	1.78	-1.28	41	6	1	1.98	-1.25	-1.12	1.92	-.29	-.68
18	4	2	1.33	1.60	-.51	1.25	1.45	-1.12	42	3	1	.82	-1.91	-1.22	1.48	.98	-.65
19	4	2	1.76	1.92	-.94	1.45	-1.08	-.95	43	5	6	2.00	1.25	-.73	1.34	.91	-1.21
20	5	6	1.35	1.08	-.69	1.49	-.59	-.53	44	5	4	2.01	-1.63	-.80	1.52	-.48	-1.13
21	5	1	1.56	-.22	-1.22	1.26	1.35	-.46	45	6	2	1.94	1.93	-1.13	1.97	-1.54	-.94
22	6	2	1.88	-1.71	-1.20	1.83	-.49	-.56	46	3	2	1.85	.37	-.58	2.08	.99	-.69
23	5	3	1.78	-1.16	-.91	2.11	1.45	-.46	47	6	1	1.76	-1.47	-.95	1.92	-1.72	-1.33
24	1	2	1.84	1.89	-.54	1.93	1.87	-.97	48	4	1	1.76	-1.87	-.85	1.75	.18	-1.22

Table 4. Parameter Means and Standard Deviations (in parentheses) in the 6D Tests.

		Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Total
6D/4	mean α	1.64	1.62	1.64	1.64	1.64	1.65	1.64
		(.42)	(.40)	(.41)	(.41)	(.44)	(.46)	(.42)
	mean δ	-.03	-.03	-.03	-.03	-.03	-.03	-.03
		(1.43)	(1.40)	(1.38)	(1.42)	(1.37)	(1.39)	(1.40)
	mean τ	-.96	-.97	-.87	-.96	-.86	-.86	-.91
		(.32)	(.25)	(.26)	(.40)	(.33)	(.25)	(.30)
6D/6	mean α	1.64	1.63	1.65	1.64	1.65	1.65	1.64
		(.44)	(.44)	(.44)	(.46)	(.46)	(.48)	(.45)
	mean δ	-.03	-.03	-.03	-.03	-.03	-.03	-.03
		(1.16)	(1.16)	(1.23)	(1.32)	(1.31)	(1.26)	(1.24)
	mean τ	-.96	-.95	-.87	-.94	-.85	-.92	-.91
		(.33)	(.26)	(.32)	(.36)	(.37)	(.27)	(.32)
6D/8	mean α	1.64	1.63	1.64	1.64	1.64	1.63	1.64
		(.42)	(.42)	(.40)	(.40)	(.42)	(.42)	(.41)
	mean δ	-.03	-.03	-.03	-.03	-.03	-.03	-.03
		(1.35)	(1.35)	(1.36)	(1.36)	(1.35)	(1.35)	(1.35)
	mean τ	-.91	-.92	-.92	-.92	-.91	-.92	-.91
		(.32)	(.25)	(.33)	(.33)	(.32)	(.25)	(.30)

MCMC prior distributions and initial values

In both the two-step and direct estimation conditions, four-parameter beta priors (1.5, 1.5, .25, 3), (2, 2, -3, 3), and (2, 2, -3, 1) were used for estimating the (α, δ, τ) statement parameters, respectively. For the four-parameter beta distribution, the first two hyperparameters influence the shape and the last two set the support (range). A four-parameter beta distribution is a common choice for item parameter priors in MCMC IRT estimation because the hyperparameters can be changed to produce a wide variety of distribution forms. For example, a (5, 5, -3, 3) four-parameter beta distribution closely resembles a standard normal distribution. $N(0, 1)$ was used as a prior for GGUM person parameter estimation, and a multivariate standard normal distribution with zero covariances among dimensions was used as a prior for MUPP person parameter estimation. All the initial values for $\alpha, \tau,$ and θ were set to 1, -1, and 0, respectively. δ values were initialized to 1 or -1 aligning with the signs of the true δ values, as in previous studies (e.g., Lee et al., 2019), as it is easy for subject matter experts to judge whether most noncognitive statements reflect a negative (lower) or positive (higher) level of a trait.

MCMC convergence checks, number of iterations, and number of chains

Convergence represents the status that the Markov chains have reached their stationary state. Convergence can be assessed using the Gelman-Rubin diagnostic index (Gelman & Rubin, 1992), which compares the variability of MCMC samples after "burn-in" within parallel chains with the variability of the samples between parallel chains. If the ratio of variability between parallel chains to within parallel chains is less than 1.2 (i.e., $\hat{R}$<1.2), practical convergence has been reached (Brooks & Gelman, 1998). In exploratory work, it was found that 50,000 iterations were generally needed for the direct approach to achieve convergence, and fewer than 30,000 iterations were needed for the two-step approach. Therefore, in the current study, 50,000

iterations were designated to foster convergence of both estimation approaches, and the first 25,000 iterations in each chain were discarded as burn-in to exclude pre-stationary samples. Following previous studies, three chains were used.

Indices of estimation accuracy and information

Four indices were calculated to evaluate parameter estimation accuracy. First, for each replication, Pearson correlations (CORRs) between true and estimated parameters were calculated, and absolute biases (ABSs) were computed as the average of the absolute differences between true and estimated parameters across statements or respondents. For example, ABS ($\widehat{\alpha}$) $= \frac{\sum_j |\widehat{\alpha}-\alpha|}{S}$, where S is the total number of statements, *j* represents the j^{th} statement, $\widehat{\alpha}$ is the parameter estimate, and α is the true parameter. Root mean square errors (RMSEs) were calculated for each replication by taking the square root of the average of the squared difference between true and estimated parameters across statements or respondents. For example, RMSE $(\widehat{\theta}_d) = \sqrt{\frac{\sum_i (\widehat{\theta}_d-\theta_d)^2}{N}}$ where N is the total number of respondents, *d* represents the dimension, *i* represents the i^{th} person (simulee), $\widehat{\theta}_d$ is the estimated person parameter, and θ_d is the true parameter for the d^{th} dimension. Posterior standard deviations (PSDs) were also calculated by taking the square root of the variance of the MCMC posterior samples after burn-in for each replication. To have a single value of each estimation effectiveness index for each condition, the obtained CORR, ABS, RMSE, and PSD values for statement and person parameter estimates were averaged across replications and dimensions. Larger CORR and smaller RMSE, PSD, and ABS values indicate better generating parameter recovery. To compare the quality of the MUPP tests, MUPP item information (Joo et al., 2018) for each dimension in every test was also computed. The person parameter estimates in the 800 sample size conditions from both the two-

step and the direct approaches were used for information calculation so that the population distribution was better represented than the 400 sample size conditions.

Hypotheses and research questions

It is well-known in the IRT literature that larger samples are beneficial to statement parameter estimation, and test length is positively related to person parameter estimation. In this study, better recovery of the generating (true) parameters is indicated by larger Pearson correlations and lower absolute bias, RMSE and PSD statistics. Thus, two hypotheses were proposed.

> *Hypothesis 1*: Large sample size (N=800) conditions will have better statement parameter recovery than small sample size (N=400) conditions across all types of MFC tests.

As in previous research by Stark et al. (2005, 2012), test length was defined in terms of IPD and allowed to increase as dimensionality increased, in order to ensure similar measurement precision for comparing 6D/4 and 12D/4, 6D/6 and 12D/6, and 6D/8 and 12D/8 tests. Consequently, the following hypothesis two hypotheses were proposed:

> *Hypothesis 2*: Tests with more IPD will result in better person parameter recovery than tests with fewer IPD.
>
> *Hypothesis 3*: Dimensionality will have no effect on person and statement parameter recovery.

The two-step approach estimates statement parameters from single-statement responses in a forced-choice test development stage, and then uses the estimated statement parameters to score MUPP tests developed and administered for assessment purposes. In contrast, the direct approach estimates statement and person parameters directly from MUPP responses. Because there has been no previous research comparing the efficacy of the two estimation approaches for

statement parameters, rather than offering hypotheses, the following research questions were proposed.

Research question 1: Will the two-step or direct estimation approach provide better statement parameter recovery?

Research question 2: Will the two-step or direct estimation approach provide better person parameter recovery?

Research question 3: How closely will the respective statement and person parameter estimates from the two-step and direct approaches accord, as indicated by Pearson correlations?

Research question 4: Will the correlations between measured dimensions (i.e., correlations of .00 or .30 between generating thetas) influence person parameter recovery?

Simulation research has shown that IRT trait estimates are robust (or perhaps insensitive) to estimation error in item parameters. For example, Stark et al. (2011) found that latent trait estimates based on SME and IRT estimates of statement location correlated above .90, when the SME and IRT location estimates correlated as low as .6. Similarly, Seybert (2013) found that trait scores based on the general and simple hyperbolic cosine models exhibited high correlations, although the former model was substantially more complex. Based on these findings, the following hypothesis was proposed.

Hypothesis 4: MUPP trait scores for the two-step and direct approaches will be highly correlated (e.g., .90).

The hypotheses and research questions above were tested using a combination of MANOVAs and ANOVAs with the parameter recovery and parameter difference indices as the dependent variables and the manipulating factors as the independent variables. Partial eta squared (η_p^2) and

eta squared (η^2) were used as effect size indices for the MANOVAs and ANOVAs, respectively. Values of .01, .06, and .14 are considered as small, medium, and large effects, respectively (Cohen, 1988).

CHAPTER THREE:

RESULTS

Table 5 presents the average convergence rates across replications for each simulation condition and each statement parameter. The convergence rates in the two-step conditions were 1.00, indicating all the estimated parameters reached convergence status. For the direct approach, convergence rates were mostly high, ranging from .80 to 1.00, except in the 12D/4 conditions, where convergence rates for alpha and delta were low (.42 to .65). Encountering difficulty in this condition is not surprising given that 4 IPD is well below previous "test length" recommendations (Stark et al., 2005, 2012), and these conditions were included intentionally to explore performance at the low end of the possible range. Nevertheless, to facilitate comparisons across conditions and accurately reflect true performance, all the parameter estimates, regardless of convergence status, were used to compute the parameter recovery indices.

Table 6 presents the statement parameter recovery results for the uncorrelated theta ($\rho_{gen} = .00$) conditions, averaged across dimensions and replications. Overall, the two-step statement parameter estimates had smaller error and higher correlations with the true (generating) parameters than the direct estimates, and the difference was most obvious in the recovery of τ. Specifically, across the two-step conditions, the correlations between true and estimated τ parameters ranged from .79 to .90, absolute biases ranged from .12 to .19, RMSEs ranged from .17 to .25, and PSDs ranged from .16 to .28; however, in the direct conditions, the correlations between true and estimated τ parameters ranged from .60 to .66, absolute biases ranged from .19 to .22, RMSEs ranged from .24 to .27, and PSDs ranged from .70 to .72. For

Table 5. Average Convergence Rates across Conditions.

Sample Size	Dimensions	IPD	ρ_{gen}	Two-Step Approach				Direct Approach			
				Overall	α	δ	τ	Overall	α	δ	τ
400	6	4	.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
			.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		6	.00	1.00	1.00	1.00	1.00	.99	1.00	.98	1.00
			.30	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
		8	.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
			.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	12	4	.00	1.00	1.00	1.00	1.00	.75	.65	.62	.98
			.30	1.00	1.00	1.00	1.00	.75	.65	.62	.98
		6	.00	1.00	1.00	1.00	1.00	.95	.97	.89	1.00
			.30	1.00	1.00	1.00	1.00	.95	.94	.90	1.00
		8	.00	1.00	1.00	1.00	1.00	.99	1.00	.98	1.00
			.30	1.00	1.00	1.00	1.00	.99	.99	.98	1.00
800	6	4	.00	1.00	1.00	1.00	1.00	.99	1.00	.99	1.00
			.30	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
		6	.00	1.00	1.00	1.00	1.00	.99	1.00	.98	1.00
			.30	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
		8	.00	1.00	1.00	1.00	1.00	1.00	1.00	.99	1.00
			.30	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	12	4	.00	1.00	1.00	1.00	1.00	.61	.43	.42	.96
			.30	1.00	1.00	1.00	1.00	.63	.47	.47	.96
		6	.00	1.00	1.00	1.00	1.00	.90	.89	.80	1.00
			.30	1.00	1.00	1.00	1.00	.88	.85	.80	1.00
		8	.00	1.00	1.00	1.00	1.00	.98	.98	.95	1.00
			.30	1.00	1.00	1.00	1.00	.96	.96	.94	1.00

Note. ρ_{gen} = Correlations between generating thetas for each dimension.

discrimination parameters, in the two-step conditions, the correlations between true and estimated α parameters ranged from .74 to .90, absolute biases ranged from .15 to .24, RMSEs ranged from .19 to .30, and PSDs ranged from .19 to .32, whereas, in the direct conditions, the correlations ranged from .64 to .86, absolute biases ranged from .18 to .28, RMSEs ranged from .23 to .35, and PSDs ranged from .23 to .41. For location parameters, in the two-step conditions, the correlations between true and estimated δ parameters ranged from .98 to .99, absolute biases ranged from .13 to .20, RMSEs ranged from .18 to .26, and PSDs ranged from .17 to .31, and in the direct conditions, the correlations ranged from .97 to .99, absolute biases ranged from .15 to .25, RMSEs ranged from .20 to .32, and PSDs ranged from .21 to .42. As expected, larger sample size was associated with better statement parameter recovery, and there were only negligible differences across corresponding 6D and 12D conditions. Interestingly, statement parameter recovery was also slightly better with more IPD, perhaps because the MCMC algorithms estimated person parameters along with statement parameters, and person parameter estimation generally improves with test length leading to better overall performance. (In the two-step conditions, however, note that the GGUM person parameter estimates from step one were discarded; the only interest was the statement parameters, which were treated as fixed in the subsequent MUPP test scoring step.) The correlations between the respective two-step and direct statement parameter estimates were also computed to examine concordance. The average overall correlation was .74. The average correlations for α, δ, and τ were .69, .98, and .57, respectively (Research Question 3). A similar pattern of statement parameter estimation results was observed for the ρ_{gen} = .30 conditions, indicating the correlations between generating thetas for each dimension had no effect on statement parameter recovery. The full results for the ρ_{gen} = .30 conditions can be found in Appendix C.

Table 7 presents the recovery results for person parameters averaged across dimensions and replications. The two-step approach and the direct approach yielded similar person parameter recovery. Specifically, in the two-step conditions, the correlation between true and estimated thetas ranged from .78 to .90, absolute biases ranged from .33 to .47, RMSEs ranged from .43 to .63, and PSDs ranged from .41 to .58. In the direct conditions, the correlation between true and estimated thetas ranged from .77 to .90, absolute biases ranged from .33 to .49, RMSEs ranged from .43 to .64, and PSDs ranged from .41 to .62. The person parameter estimates for the two-step and direct approaches were highly correlated, with correlations of .98 in both the $\rho_{gen} = .00$ and $\rho_{gen} = .30$ conditions, and negligible differences in terms of estimation efficacy (Research Question 3). As expected, person parameter recovery was better in conditions with more IPD, and the results were very similar across conditions with corresponding dimensionality and sample size, with one exception: for the direct approach, estimation was slightly better in the 6D/4 conditions than in the corresponding 12D/4 conditions where some convergence concerns were noted. Full person parameter recovery results for each dimension within the tests can be found in Appendix C.

Test information was also calculated for each dimension in the 6D and 12D tests using the true statement parameters and the person parameter estimates from the large sample size (N=800) conditions. These results are shown in Tables 8 and 9. As intended, the information values were highly similar across the dimensions within each test because the respective statement parameters were selected to produce nearly equal means and standard deviations across dimensions during test design. Also, as expected, tests with more IPD provided proportionally higher information than tests with fewer IPD. For additional insight into the comparability of the various tests, reliabilities for each test were also computed by squaring the

Table 6. Statement Parameter Recovery Results for the Uncorrelated Dimensions ($\rho_{gen} = .00$) Conditions.

Sample Size	Dimensions	IPD	Recovery Statistics	Two-step Approach			Direct Approach		
				α	δ	τ	α	δ	τ
400	6	4	ABS	.22	.20	.19	.26	.24	.22
			RMSE	.27	.26	.25	.32	.31	.27
			PSD	.32	.31	.28	.39	.37	.72
			CORR	.80	.99	.79	.71	.98	.61
		6	ABS	.21	.17	.15	.24	.19	.22
			RMSE	.26	.23	.21	.30	.26	.27
			PSD	.28	.24	.21	.33	.28	.71
			CORR	.84	.99	.86	.79	.98	.61
		8	ABS	.20	.17	.16	.23	.20	.20
			RMSE	.25	.23	.21	.28	.26	.25
			PSD	.26	.24	.22	.31	.28	.71
			CORR	.83	.99	.83	.78	.97	.63
	12	4	ABS	.24	.20	.18	.28	.25	.21
			RMSE	.30	.26	.24	.35	.32	.26
			PSD	.32	.31	.28	.41	.42	.72
			CORR	.74	.99	.79	.64	.97	.60
		6	ABS	.21	.18	.16	.24	.19	.21
			RMSE	.27	.24	.22	.30	.25	.26
			PSD	.28	.24	.21	.34	.29	.71
			CORR	.83	.98	.84	.78	.98	.63
		8	ABS	.20	.17	.16	.23	.19	.20
			RMSE	.25	.23	.22	.28	.25	.25
			PSD	.26	.24	.22	.31	.28	.71
			CORR	.83	.99	.81	.78	.98	.63

Table 6 (Continued)

800	6	4	ABS	.18	.17	.15	.22	.20	.20
			RMSE	.22	.23	.20	.27	.26	.25
			PSD	.25	.24	.21	.31	.29	.71
			CORR	.86	.99	.82	.80	.98	.63
		6	ABS	.16	.13	.12	.20	.15	.21
			RMSE	.21	.18	.17	.25	.20	.25
			PSD	.21	.17	.16	.26	.21	.70
			CORR	.90	.99	.90	.86	.99	.66
		8	ABS	.15	.13	.12	.18	.15	.19
			RMSE	.19	.19	.17	.23	.21	.24
			PSD	.19	.17	.16	.23	.21	.71
			CORR	.90	.99	.88	.86	.99	.66
	12	4	ABS	.19	.17	.15	.25	.20	.20
			RMSE	.24	.22	.20	.31	.26	.25
			PSD	.24	.24	.21	.34	.33	.72
			CORR	.85	.98	.85	.72	.98	.62
		6	ABS	.16	.14	.12	.20	.16	.21
			RMSE	.21	.19	.17	.25	.21	.26
			PSD	.21	.18	.16	.26	.22	.70
			CORR	.90	.98	.90	.86	.99	.64
		8	ABS	.15	.13	.12	.18	.15	.19
			RMSE	.19	.19	.17	.23	.21	.24
			PSD	.19	.17	.16	.23	.21	.71
			CORR	.90	.99	.87	.86	.98	.66

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

correlation between the respective true and estimated thetas. As shown in Table 10, reliabilities were similar for the two-step and direct estimation approaches, and tests having more IPD had higher reliabilities than tests with fewer IPD.

Statistical significance tests of hypotheses and research questions

To formally test the proposed hypotheses and answer the research questions, MANOVAs were conducted. The results are presented in Table 11. Note that for MANOVA tests shown in bold in the Parameter column, the Box's tests of equality of covariance matrices were significant, indicating the homogeneity of variance-covariance matrices assumption were violated. Therefore, the Pillai's trace criterion, which is more robust to assumption violations (Tabachnick, Fidell, & Ullman, 2007), was used. The MANOVA results showed that sample size, IPD, dimensionality, and estimation approaches had statistically significant effects ($p < .05$) on the statement parameter recovery indices, and IPD, estimation approaches, and magnitude of correlations between generating thetas for each dimension had statistically significant effects on person parameter recovery indices. However, follow-up univariate tests, shown in Table 12, indicated that the effect of dimensionality on each statement parameter recovery index was not significant; in addition, neither estimation approach nor correlation between generating thetas had a significant effect on person parameter recovery indices. This indicates that there were differences on linear combinations of the parameter recovery indices, but not on the indices considered separately. Because the purpose was to test the significance of differences for each parameter recovery index, emphasis was thus placed on the univariate test results. There was a significant effect of sample size on α ABS, α RMSE, α PSD, α CORR, δ ABS, δ RMSE, δ PSD, δ CORR, with large effect sizes: $\eta 2$ = .39, .40, .43, .31, .36, .33, .31, and .16, respectively. τ parameters are sometimes difficult to estimate (e.g., Lee et al., 2019); therefore, the effect of sample size was

Table 7. Person Parameter Recovery Results across Conditions.

Sample Size	IPD	Recovery Statistics	Two-step Approach				Direct Approach			
			6D/.00	12D/.00	6D/.30	12D/.30	6D/.00	12D/.00	6D/.30	12D/.30
400	4	ABS	.47	.47	.45	.45	.47	.48	.46	.48
		RMSE	.62	.62	.59	.59	.62	.64	.60	.62
		PSD	.58	.57	.55	.57	.60	.61	.57	.62
		CORR	.79	.78	.81	.81	.78	.77	.80	.78
	6	ABS	.40	.40	.38	.38	.40	.40	.38	.39
		RMSE	.53	.53	.51	.50	.53	.53	.51	.51
		PSD	.49	.49	.47	.47	.50	.50	.49	.50
		CORR	.85	.85	.86	.87	.85	.85	.86	.86
	8	ABS	.35	.35	.34	.34	.35	.35	.34	.34
		RMSE	.45	.46	.44	.44	.45	.45	.44	.44
		PSD	.42	.42	.41	.41	.44	.43	.42	.43
		CORR	.89	.89	.90	.90	.89	.89	.90	.90
800	4	ABS	.46	.47	.45	.45	.46	.49	.45	.48
		RMSE	.61	.63	.59	.59	.61	.64	.59	.62
		PSD	.57	.57	.55	.57	.58	.61	.56	.62
		CORR	.79	.78	.81	.81	.79	.77	.81	.78
	6	ABS	.39	.39	.38	.38	.39	.39	.38	.38
		RMSE	.52	.52	.50	.50	.53	.53	.50	.50
		PSD	.49	.49	.47	.48	.49	.49	.48	.50
		CORR	.85	.85	.87	.87	.85	.85	.87	.86
	8	ABS	.34	.34	.33	.33	.34	.34	.33	.33
		RMSE	.44	.45	.43	.43	.45	.45	.43	.43
		PSD	.42	.42	.41	.41	.43	.42	.41	.43
		CORR	.89	.90	.90	.90	.89	.90	.90	.90

Note. ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR = correlation between true and estimated parameters; 6D/.00 = 6 uncorrelated dimension; 6D/.30 = 6 dimension with .30 correlations between dimensions; 12D/.00 = 12 uncorrelated dimension; 12D/.30 = 12 dimension with .30 correlations between dimensions.

Table 8. Test Information by Dimension in the 6D Test Conditions.

ρ_{gen}		Estimation Approach	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6
.00	6D/4	Direct	2.81	3.26	2.83	3.37	3.10	3.33
		Two-Step	2.81	3.27	2.82	3.39	3.10	3.31
	6D/6	Direct	4.64	4.96	4.83	4.91	4.76	5.16
		Two-Step	4.63	4.96	4.83	4.92	4.76	5.13
	6D/8	Direct	5.99	6.55	6.26	6.25	5.98	6.54
		Two-Step	5.98	6.54	6.27	6.27	5.99	6.53
.30	6D/4	Direct	2.78	3.29	2.88	3.43	3.28	3.41
		Two-Step	2.78	3.30	2.88	3.46	3.28	3.41
	6D/6	Direct	4.62	5.03	4.85	5.02	5.02	5.23
		Two-Step	4.61	5.02	4.85	5.03	5.03	5.22
	6D/8	Direct	6.09	6.66	6.34	6.34	6.09	6.64
		Two-Step	6.09	6.64	6.35	6.34	6.09	6.63

Note. ρ_{gen} = Correlations between generating thetas for each dimension.

Table 9. Test Information by Dimension in the 12D Test Conditions.

ρ_{gen}		Estimation Approach	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10	Dim 11	Dim 12
.00	12D/4	Direct	2.83	3.41	2.88	3.48	3.19	3.48	2.84	3.42	2.87	3.50	3.18	3.46
		Two-Step	2.83	3.34	2.86	3.46	3.16	3.41	2.83	3.35	2.86	3.44	3.15	3.39
	12D/6	Direct	4.62	5.02	4.87	4.97	4.82	5.23	4.61	5.03	4.86	4.98	4.82	5.25
		Two-Step	4.63	5.00	4.85	4.95	4.80	5.20	4.63	5.00	4.84	4.96	4.82	5.22
	12D/8	Direct	6.01	6.60	6.28	6.28	6.01	6.60	6.00	6.59	6.28	6.28	6.01	6.61
		Two-Step	6.01	6.58	6.26	6.26	5.99	6.57	5.99	6.56	6.27	6.27	5.99	6.57
.30	12D/4	Direct	2.78	3.42	2.86	3.47	3.31	3.46	2.81	3.38	2.86	3.48	3.27	3.44
		Two-Step	2.77	3.31	2.87	3.44	3.26	3.39	2.79	3.29	2.88	3.42	3.23	3.38
	12D/6	Direct	4.60	5.00	4.87	5.00	4.96	5.23	4.59	5.01	4.87	5.00	4.98	5.25
		Two-Step	4.62	5.00	4.85	5.00	4.98	5.22	4.61	5.01	4.83	4.99	4.99	5.22
	12D/8	Direct	6.05	6.58	6.30	6.30	6.06	6.58	6.05	6.58	6.30	6.29	6.06	6.59
		Two-Step	6.07	6.60	6.32	6.32	6.07	6.60	6.07	6.60	6.32	6.32	6.07	6.61

Note. ρ_{gen} = Correlations between generating thetas for each dimension.

Table 10. Test Reliability.

Estimation Approach	ρ_{gen}	Sample Size	6D/4	6D/6	6D/8	12D/4	12D/6	12D/8
Direct	.00	400	.61	.72	.79	.59	.72	.79
		800	.62	.72	.79	.59	.72	.81
	.30	400	.64	.74	.81	.61	.74	.81
		800	.66	.76	.81	.61	.74	.81
	Average		.63	.74	.80	.60	.73	.81
Two-Step	.00	400	.62	.72	.79	.61	.72	.79
		800	.62	.72	.79	.61	.72	.81
	.30	400	.66	.74	.81	.66	.76	.81
		800	.66	.76	.81	.66	.76	.81
	Average		.64	.74	.80	.63	.74	.81

Note. ρ_{gen} = Correlations between generating thetas for each dimension.

only significant on τ RMSE with $\eta 2 = .17$. This partially supported Hypothesis 1 that large sample size conditions would show better statement parameter recovery than small sample size conditions. IPD had statistically significant effects on person parameter (θ) recovery with large effect sizes ranging from .95 to .97. This supported Hypothesis 2 that tests with more IPD would result in better person parameter recovery than tests with fewer IPD. Interestingly, the number of IPD also had statistically significant effects on indices of α and δ parameter recovery, even after Bonferroni correction, with large effect sizes ranging from .25 to .42. This was likely an indirect benefit of improved person parameter estimation stemming from MCMC joint estimation. Hypothesis 3, which stated that test dimensionality would have no effect on statement and person parameter recovery, was also supported by univariate tests. With regard to research questions, estimation approach had statistically significant effects on statement parameter recovery, with effect sizes ranging from .20 to .99 (Research Question 1). However, consistent with a previous study by Stark et al. (2011), there was no effect on person parameter recovery (Research Question 2). Indeed, person parameters based on the two estimation approaches correlated .98, supporting Hypothesis 4. This is an important finding for practice because it suggests that both

test development methods are viable for constructing MUPP tests for assessment purposes. Finally, using correlated (.30) vs. uncorrelated (.00) generating person parameters for the various dimensions had no significant effect on estimation (Research Question 4). Bonferroni post-hoc multiple comparisons were also conducted for IPD to perform pairwise comparisons between the means of 4, 6, and 8 IPD conditions. The results in Table 13 suggest that 8 IPD led to significantly better person parameter recovery than 6 IPD, which led to better person parameter recovery than 4 IPD. This further supported Hypothesis 2. For statement parameter recovery, 6 IPD and 8 IPD were significantly better than 4 IPD for α and δ recovery, but there was no difference between the longer test conditions.

Table 11. Multivariate Tests of Between Subjects Effects for Hypotheses and Research Questions.

Hypothesis	Factor	Parameter	Pillai's Trace	F	Hypothesis df	Error df	Sig	η_p^2
Hypothesis 1	Sample size	**Statement parameter**	.84	14.72	12	35	**.00**	.84
		Person parameter	.15	1.88	4	43	.13	.15
Hypothesis 2	IPD	**Statement parameter**	1.57	10.54	24	70	**.00**	.78
		Person parameter	1.55	37.39	8	86	**.00**	.78
Hypothesis 3	Dimensionality	**Statement parameter**	.46	2.45	12	35	**.02**	.46
		Person parameter	.03	.34	4	43	.85	.03
Research question 1&2	Estimation Approach	**Statement parameter**	1.00	3884.08	12	35	**.00**	1.00
		Person parameter	.40	7.09	4	43	**.00**	.40
Research question 4	Correlation	**Statement parameter**	.37	1.70	12	35	.11	.37
		Person parameter	.46	8.96	4	43	**.00**	.46

Table 12. Univariate Tests of Between Subjects Effects for Hypotheses and Research Questions.

Hypothesis	Factor	Parameter	DV	SS	df	Mean Square	F	Sig	η^2
Hypothesis 1	Sample size	Statement parameter	ABS_a	.02	1.00	.02	29.77	**.00**	.39
			ABS_b	.02	1.00	.02	25.71	**.00**	.36
			ABS_tau	.01	1.00	.01	6.27	.02	.12
			CORR_a	.07	1.00	.07	20.47	**.00**	.31
			CORR_b	.00	1.00	.00	8.84	.01	.16
			CORR_tau	.02	1.00	.02	1.43	.24	.03
			RMSE_a	.04	1.00	.04	30.63	**.00**	.40
			RMSE_b	.03	1.00	.03	22.90	**.00**	.33
			RMSE_tau	.01	1.00	.01	9.42	**.00**	.17
			PSD_a	.07	1.00	.07	35.11	**.00**	.43
			PSD_b	.06	1.00	.06	20.15	**.00**	.31
			PSD_tau	.01	1.00	.01	0.19	.67	.00
Hypothesis 2	IPD	Statement parameter	ABS_a	.02	2.00	.01	7.61	**.00**	.25
			ABS_b	.02	2.00	.01	11.99	**.00**	.35
			ABS_tau	.00	2.00	.00	1.12	.33	.05
			CORR_a	.07	2.00	.04	10.43	**.00**	.32
			CORR_b	.00	2.00	.00	3.57	.04	.14
			CORR_tau	.02	2.00	.01	0.61	.55	.03
			RMSE_a	.02	2.00	.01	7.73	**.00**	.26
			RMSE_b	.03	2.00	.01	10.74	**.00**	.32
			RMSE_tau	.00	2.00	.00	1.44	.25	.06
			PSD_a	.05	2.00	.02	9.60	**.00**	.30
			PSD_b	.08	2.00	.04	16.29	**.00**	.42
			PSD_tau	.01	2.00	.01	0.09	.92	.00
		Person parameter	ABS_theta	.13	2.00	.06	651.73	**.00**	.97
			RMSE_theta	.23	2.00	.12	544.99	**.00**	.96
			PSD_theta	.21	2.00	.10	452.21	**.00**	.95
			CORR_theta	.09	2.00	.05	449.49	**.00**	.95
Hypothesis 3	Dimensionality	Statement parameter	ABS_a	.00	1.00	.00	.70	.41	.02
			ABS_b	.00	1.00	.00	.52	.48	.01
			ABS_tau	.00	1.00	.00	.00	.98	.00
			CORR_a	.01	1.00	.01	1.32	.26	.03
			CORR_b	.00	1.00	.00	1.85	.18	.04
			CORR_tau	.00	1.00	.00	.01	.93	.00
			RMSE_a	.00	1.00	.00	.84	.37	.02
			RMSE_b	.00	1.00	.00	.92	.34	.02
			RMSE_tau	.00	1.00	.00	.00	.96	.00
			PSD_a	.00	1.00	.00	.06	.81	.00
			PSD_b	.00	1.00	.00	.50	.48	.01
			PSD_tau	.00	1.00	.00	.00	.99	.00
Research question 1&2	Approach	Statement parameter	ABS_a	.02	1.00	.02	17.66	**.00**	.28
			ABS_b	.01	1.00	.01	11.59	**.00**	.20
			ABS_tau	.04	1.00	.04	133.81	**.00**	.74
			CORR_a	.05	1.00	.05	14.33	**.00**	.24
			CORR_b	.00	1.00	.00	17.35	**.00**	.27
			CORR_tau	.55	1.00	.55	615.39	**.00**	.93
			RMSE_a	.02	1.00	.02	17.16	**.00**	.27
			RMSE_b	.02	1.00	.02	12.01	**.00**	.21
			RMSE_tau	.04	1.00	.04	87.00	**.00**	.65
			PSD_a	.04	1.00	.04	15.09	**.00**	.25
			PSD_b	.04	1.00	.04	11.26	**.00**	.20
			PSD_tau	3.08	1.00	3.08	3497.78	**.00**	.99
		Person parameter	ABS_theta	.00	1.00	.00	.08	.78	.00
			RMSE_theta	.00	1.00	.00	.08	.79	.00
			PSD_theta	.00	1.00	.00	.80	.38	.02
			CORR_theta	.00	1.00	.00	.09	.76	.00
Research question 4	Correlation	Person parameter	ABS_theta	.00	1.00	.00	.74	.40	.02
			RMSE_theta	.01	1.00	.01	1.19	.28	.03
			PSD_theta	.00	1.00	.00	.20	.66	.00
			CORR_theta	.00	1.00	.00	1.32	.26	.03

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

Table 13. Multiple Comparisons with IPD.

Dependent Variable	Comparison	Mean Difference	Sig
ABS_a	4 - 6	.03	.04
	4 - 8	.04	.00
	6 - 8	.01	.67
ABS_b	4 - 6	.04	.00
	4 - 8	.04	.00
	6 - 8	-.00	1.00
ABS_tau	4 - 6	.01	.96
	4 - 8	.02	.45
	6 - 8	.01	1.00
CORR_a	4 - 6	-.08	.00
	4 - 8	-.08	.00
	6 - 8	.00	1.00
CORR_b	4 - 6	-.00	.16
	4 - 8	-.01	.04
	6 - 8	-.00	1.00
CORR_tau	4 - 6	-.04	.92
	4 - 8	-.03	1.00
	6 - 8	.01	1.00
RMSE_a	4 - 6	.04	.04
	4 - 8	.05	.00
	6 - 8	-.04	.63
RMSE_b	4 - 6	.05	.00
	4 - 8	.05	.00
	6 - 8	.00	1.00
RMSE_tau	4 - 6	.01	.87
	4 - 8	.02	.30
	6 - 8	.01	1.00
PSD_a	4 - 6	.05	.02
	4 - 8	.07	.00
	6 - 8	.02	.58
PSD_b	4 - 6	.09	.00
	4 - 8	.09	.00
	6 - 8	.00	1.00
PSD_tau	4 - 6	.04	1.00
	4 - 8	.03	1.00
	6 - 8	-.00	1.00

Table 13 (Continued)

ABS_theta	4 - 6	.08	.00
	4 - 8	.13	.00
	6 - 8	.05	.00
RMSE _theta	4 - 6	.10	.00
	4 - 8	.17	.00
	6 - 8	.07	.00
PSD_theta	4 - 6	.09	.00
	4 - 8	.16	.00
	6 - 8	.07	.00
CORR _theta	4 - 6	-.07	.00
	4 - 8	-.11	.00
	6 - 8	-.04	.00

Note. Bonferroni Correction was used.

CHAPTER FOUR:

DISCUSSION

This research compared two approaches to statement and person parameter estimation for the MUPP IRT model (Stark et al., 2005). A simulation study was conducted to compare the efficacy of the two-step approach to test construction and scoring proposed by Stark et al. (2005) with a more recently developed MCMC method (Lee et al., 2019) for estimating statement and person parameters directly from forced-choice responses. Results indicated the two-step approach was more effective in recovering generating statement parameters. However, the two-step and direct approaches were both effective in recovering generating person parameters, and the respective latent trait estimates correlated highly. This is a fundamentally important issue for practice. It confirms the findings in previous studies (e.g., Stark et al., 2011; Seybert, 2013) that IRT trait scores are robust to statement/item parameter estimation error and suggests that both the two-step and the direct approaches are viable for MUPP scoring. However, if researchers are interested in building MUPP CATs for efficiency, the two-step approach, which allows a large calibrated statement pool being created, is recommended with possible DIF screening of the statement pool prior to operational testing. In contrast, if interest lies primarily in constructing alternate static (nonadaptive) forms and examining measurement invariance across comparison groups, then the direct approach, which takes into account the potential interactions of statements within items, may be preferable. Toward that end, researchers should include some extra items in each test form to allow for deleting any that are problematic.

Importantly, the observed differences in statement parameter recovery for the two-step and direct approaches may be attributable to differences in the complexity of the estimation models: GGUM vs. MUPP based on the GGUM. An alternative explanation is that statement parameter estimates are influenced by context. More specifically, when statements are presented in pairs for comparative judgments, their parameters differ from when they are evaluated one at a time in a long inventory using an ordered-categorical response format. Future research involving "think-aloud" protocols and statistical tests of MUPP model assumptions may be needed to address this possibility. For MUPP test construction purposes, if the former is the case, the two-step approach should be preferred because it requires smaller sample size than the direct approach for statement calibration due to the model parsimony. If the latter is the case, the direct approach is recommended, but it appears substantially larger samples are needed for similar estimation efficacy. The large differences in the recovery of τ parameters for the two approaches and the difficulty in accurately estimating τ parameters also raise questions about the added value of allowing τ parameters to differ across statements. Future research might, therefore, explore parameter recovery with simpler models that either constrain τ parameters or focus exclusively on location and discrimination parameters.

In addition to these important findings, this study advanced understanding of estimation efficacy with short nonadaptive tests (having as few as 4 IPD), in connection with dimensionality, sample size, and the magnitude of correlations between dimensions. The simulation indicated that correlations between generating and estimated person parameters can exceed .8 with nonadaptive tests having as few as 6 IPD, although longer tests yielding correlations above .9 (8IPD) are desirable for high-stakes decision making. These results buttress the findings of Stark et al. (2012), which examined latent trait estimation for nonadaptive and adaptive MUPP tests.

The study is not without limitations. First, due to the long run time of the algorithms, only 20 replications were performed for each condition. Future research could explore other estimation algorithms (e.g., Hamiltonian Monte Carlo) that are considered more efficient and could potentially reduce the run time. Future research might also include more levels of the manipulated factors, such as tests of even higher dimensionality that may be of interest in practice. Besides MUPP estimation, research on MUPP linking and differential item functioning (DIF) detection is also needed to facilitate MFC applications. Research exploring different methods of linking with MUPP tests is now underway, as well as research to examine the efficacy of DIF detection using adaptations of methods developed for unidimensional IRT models.

REFERENCES

Aon, H. (2015). Trends in global employee engagement report. *Lincolnshire, IL: Aon Corp*.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, *7*(4), 434-455.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460-502.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of applied psychology*.

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

Cho, S., Drasgow, F., & Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychological assessment*, *27*(4), 1241.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences–second edition. 12 Lawrence Erlbaum Associates Inc. *Hillsdale, New Jersey*, *13*.

De La Torre, J., Stark, S., & Chernyshenko, O. S. (2006). Markov chain Monte Carlo estimation of item parameters for the generalized graded unfolding model. *Applied Psychological Measurement*, *30*(3), 216-232.

Doornik, J. A. (2009). *An object-oriented matrix language: Ox 6*. London, UK: Timberlake Consultants Press.

Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the tailored adaptive personality assessment system (TAPAS) to*

support army personnel selection and classification decisions. Drasgow Consulting Group Urbana IL.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, *7*(4), 457-472.

Guan, A., Sun, T., & Carter, N.T. (2021). *MUPPscore: An R script for expected a posteriori scoring of multi-unidimensional pairwise preference items*. PsyArXiv. http://doi.org/10.31234/osf.io/b6sq5

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences*, *123*, 229-235.

Lee, P., Joo, S. H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied psychological measurement*, *43*(3), 226-240.

Lee, P., Joo, S. H., & Stark, S. (2020). Detecting DIF in Multidimensional Forced Choice Measures Using the Thurstonian Item Response Theory Model. *Organizational Research Methods*, 1094428120959822.

Joo, S. H., Lee, P., & Stark, S. (2017). Evaluating anchor-item designs for concurrent calibration with the GGUM. *Applied psychological measurement*, *41*(2), 83-96.

Joo, S. H., Lee, P., & Stark, S. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model. *Journal of Educational Measurement*, *55*(3), 357-372.

Joo, S. H., Lee, P., & Stark, S. (2019). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods*, 1-12.

Nye, C.D., White, L.A., Drasgow, F., Prasad, J., Chernyshenko, O.S., & Stark, S. (2020) Examining personality for the selection and classification of soldiers: Validity and differential validity across jobs. *Military Psychology, 32*, 60-70.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*(1), 3-32.

Seybert, J. (2013). A new item response theory model for estimating person ability and item parameters for multidimensional rank order responses.

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., & 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*(3), 184-203.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*(1), 25.

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can subject matter experts' ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scales?. *Organizational Research Methods*, *14*(2), 256-278.

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*(3), 463-487.

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*(3), 153-164.

Sun, T. (2017). *When matches are ideal: Fitting measurement models to adult attachment data* (Doctoral dissertation).

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5, pp. 481-498). Boston, MA: Pearson.

Tay, L., Drasgow, F., Rounds, J., & Williams, B. A. (2009). Fitting measurement models to vocational interest data: Are dominance models ideal?. *Journal of Applied Psychology*, *94*(5), 1287.

Wang, W. C., Qiu, X. L., Chen, C. W., Ro, S., & Jin, K. Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*(8), 600-613.

Wetzel, E., & Frick, S. (2019). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment*.

Wetzel, E., Frick, S., & Brown, A. (2020). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*.

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

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2020). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, *23*(3), 569-590.

APPENDIX A:

TEST SPECIFICATIONS FOR THE MUPP TESTS

Table A1. Test Specifications for the 12D/4 MUPP Test.

Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t	Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t
1	1	6	1.81	.60	-1.37	.75	-1.78	-1.17	25	1	5	.75	1.82	-1.37	.78	-1.61	-1.12
2	1	3	1.83	-1.99	-.55	1.88	-.69	-1.24	26	4	5	1.80	-.68	-1.31	2.13	-1.03	-.91
3	3	4	.81	-1.49	-.55	1.31	-1.12	-.41	27	2	4	.97	.56	-.98	1.88	1.23	-1.12
4	2	4	1.77	1.36	-.65	1.83	-1.37	-1.31	28	3	6	2.06	-.10	-.97	1.94	.32	-.51
5	3	4	1.98	.85	-.76	1.51	1.78	-1.28	29	2	5	1.98	-1.25	-1.12	1.92	-.29	-.68
6	3	6	1.33	1.60	-.51	1.25	1.45	-1.12	30	4	5	.82	-1.91	-1.22	1.48	.98	-.65
7	3	6	1.76	1.92	-.94	1.45	-1.08	-.95	31	1	2	2.00	1.25	-.73	1.34	.91	-1.21
8	1	2	1.35	1.08	-.69	1.49	-.59	-.53	32	1	3	2.01	-1.63	-.80	1.52	-.48	-1.13
9	1	5	1.56	-.22	-1.22	1.26	1.35	-.46	33	2	6	1.94	1.93	-1.13	1.97	-1.54	-.94
10	2	6	1.88	-1.71	-1.20	1.83	-.49	-.56	34	4	6	1.85	.37	-.58	2.08	.99	-.69
11	1	4	1.78	-1.16	-.91	2.11	1.45	-.46	35	2	5	1.76	-1.47	-.95	1.92	-1.72	-1.33
12	5	6	1.84	1.89	-.54	1.93	1.87	-.97	36	3	5	1.76	-1.87	-.85	1.75	.18	-1.22
13	7	12	1.81	.60	-1.37	.75	-1.78	-1.17	37	7	11	.75	1.82	-1.37	.78	-1.61	-1.12
14	7	9	1.83	-1.99	-.55	1.88	-.69	-1.24	38	10	11	1.80	-.68	-1.31	2.13	-1.03	-.91
15	9	10	.81	-1.49	-.55	1.31	-1.12	-.41	39	8	10	.97	.56	-.98	1.88	1.23	-1.12
16	8	10	1.77	1.36	-.65	1.83	-1.37	-1.31	40	9	12	2.06	-.10	-.97	1.94	.32	-.51
17	9	10	1.98	.85	-.76	1.51	1.78	-1.28	41	8	11	1.98	-1.25	-1.12	1.92	-.29	-.68
18	9	12	1.33	1.60	-.51	1.25	1.45	-1.12	42	10	11	.82	-1.91	-1.22	1.48	.98	-.65
19	9	12	1.76	1.92	-.94	1.45	-1.08	-.95	43	7	8	2.00	1.25	-.73	1.34	.91	-1.21
20	7	8	1.35	1.08	-.69	1.49	-.59	-.53	44	7	9	2.01	-1.63	-.80	1.52	-.48	-1.13
21	7	11	1.56	-.22	-1.22	1.26	1.35	-.46	45	8	12	1.94	1.93	-1.13	1.97	-1.54	-.94
22	8	12	1.88	-1.71	-1.20	1.83	-.49	-.56	46	10	12	1.85	.37	-.58	2.08	.99	-.69
23	7	10	1.78	-1.16	-.91	2.11	1.45	-.46	47	8	11	1.76	-1.47	-.95	1.92	-1.72	-1.33
24	11	12	1.84	1.89	-.54	1.93	1.87	-.97	48	9	11	1.76	-1.87	-.85	1.75	.18	-1.22

Table A2. Test Specifications for the 12D/6 MUPP Test.

Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t	Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t
1	1	6	1.81	.60	-1.37	.75	-1.78	-1.17	37	1	5	.75	1.82	-1.37	.78	-1.61	-.67
2	1	3	1.83	-1.99	-.55	1.88	-.69	-1.24	38	4	5	1.80	-.68	-1.31	2.13	-1.03	-.91
3	3	4	.81	-1.49	-.55	1.31	-1.12	-.41	39	2	4	.97	.56	-.98	1.88	1.23	-1.12
4	2	4	1.77	1.36	-.99	1.83	-1.37	-1.31	40	3	6	2.06	-.10	-.97	1.94	.32	-.51
5	3	4	1.98	.85	-.76	1.51	1.78	-1.28	41	2	5	1.98	-1.25	-1.12	1.92	-.29	-.68
6	3	6	1.33	1.60	-.51	1.25	1.45	-1.12	42	4	5	.82	-1.91	-1.22	1.48	.98	-.65
7	3	6	1.76	1.92	-.94	1.45	-1.08	-.95	43	1	2	2.00	1.25	-.73	1.34	.91	-1.21
8	1	2	1.35	1.08	-.69	1.49	-.59	-.53	44	1	3	2.01	-1.63	-.80	1.52	-.48	-1.13
9	1	5	1.56	-.22	-1.22	1.26	1.35	-.46	45	2	6	1.94	1.93	-1.13	1.97	-1.54	-.94
10	2	6	1.88	-1.71	-1.20	1.83	-.49	-.56	46	4	6	1.85	.37	-.58	2.08	.99	-.69
11	1	4	1.78	-1.16	-.91	2.11	1.45	-.46	47	2	5	1.76	-1.47	-.95	1.92	-1.72	-1.33
12	5	6	1.84	1.89	-.54	1.93	1.87	-.57	48	3	5	1.76	-1.87	-.85	1.75	.18	-.66
13	5	2	1.84	-.69	-.37	1.88	-.85	-1.20	49	5	1	.87	.56	-1.22	1.81	.01	-.69
14	5	4	1.81	1.60	-1.37	1.88	.04	-1.24	50	3	1	1.76	.37	-.76	.93	.30	-.55
15	4	5	1.80	-1.51	-.55	2.15	-1.59	-1.37	51	6	3	2.17	-.10	-1.17	2.13	-.59	-1.12
16	6	3	1.76	1.62	-.94	1.83	-1.09	-1.31	52	4	2	.76	-.29	-1.01	.89	.40	-.53
17	2	6	1.76	.04	-.94	1.83	-.88	-1.21	53	6	1	.84	-.71	-1.20	1.75	.02	-.68
18	4	2	2.11	1.63	-.76	2.12	.35	-.57	54	3	1	.93	-1.09	-.24	2.15	-.49	-1.37
19	7	12	1.81	.60	-1.37	.75	-1.78	-1.17	55	7	11	.75	1.82	-1.37	.78	-1.61	-.67
20	7	9	1.83	-1.99	-.55	1.88	-.69	-1.24	56	10	11	1.80	-.68	-1.31	2.13	-1.03	-.91
21	9	10	.81	-1.49	-.55	1.31	-1.12	-.41	57	8	10	.97	.56	-.98	1.88	1.23	-1.12
22	8	10	1.77	1.36	-.99	1.83	-1.37	-1.31	58	9	12	2.06	-.10	-.97	1.94	.32	-.51
23	9	10	1.98	.85	-.76	1.51	1.78	-1.28	59	8	11	1.98	-1.25	-1.12	1.92	-.29	-.68
24	9	12	1.33	1.60	-.51	1.25	1.45	-1.12	60	10	11	.82	-1.91	-1.22	1.48	.98	-.65
25	9	12	1.76	1.92	-.94	1.45	-1.08	-.95	61	7	8	2.00	1.25	-.73	1.34	.91	-1.21
26	7	8	1.35	1.08	-.69	1.49	-.59	-.53	62	7	9	2.01	-1.63	-.80	1.52	-.48	-1.13
27	7	11	1.56	-.22	-1.22	1.26	1.35	-.46	63	8	12	1.94	1.93	-1.13	1.97	-1.54	-.94
28	8	12	1.88	-1.71	-1.20	1.83	-.49	-.56	64	10	12	1.85	.37	-.58	2.08	.99	-.69
29	7	10	1.78	-1.16	-.91	2.11	1.45	-.46	65	8	11	1.76	-1.47	-.95	1.92	-1.72	-1.33
30	11	12	1.84	1.89	-.54	1.93	1.87	-.57	66	9	11	1.76	-1.87	-.85	1.75	.18	-.66
31	11	8	1.84	-.69	-.37	1.88	-.85	-1.20	67	11	7	.87	.56	-1.22	1.81	.01	-.69
32	11	10	1.81	1.60	-1.37	1.88	.04	-1.24	68	9	7	1.76	.37	-.76	.93	.30	-.55
33	10	11	1.80	-1.51	-.55	2.15	-1.59	-1.37	69	12	9	2.17	-.10	-1.17	2.13	-.59	-1.12
34	12	9	1.76	1.62	-.94	1.83	-1.09	-1.31	70	10	8	.76	-.29	-1.01	.89	.40	-.53
35	8	12	1.76	.04	-.94	1.83	-.88	-1.21	71	12	7	.84	-.71	-1.20	1.75	.02	-.68
36	10	8	2.11	1.63	-.76	2.12	.35	-.57	72	9	7	.93	-1.09	-.24	2.15	-.49	-1.37

Table A3. Test Specifications for the 12D/8 MUPP Test.

Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t	Item	s	t	α_s	δ_s	τ_s	α_t	δ_t	τ_t
1	1	6	1.81	.60	-1.37	.75	-1.78	-1.17	49	1	5	.75	1.82	-1.37	.78	-1.61	-1.12
2	1	3	1.83	-1.99	-.55	1.88	-.69	-1.24	50	4	5	1.80	-.68	-1.31	2.13	-1.03	-.91
3	3	4	.81	-1.49	-.55	1.31	-1.12	-.41	51	2	4	.97	.56	-.98	1.88	1.23	-1.12
4	2	4	1.77	1.36	-.65	1.83	-1.37	-1.31	52	3	6	2.06	-.10	-.97	1.94	.32	-.51
5	3	4	1.98	.85	-.76	1.51	1.78	-1.28	53	2	5	1.98	-1.25	-1.12	1.92	-.29	-.68
6	3	6	1.33	1.60	-.51	1.25	1.45	-1.12	54	4	5	.82	-1.91	-1.22	1.48	.98	-.65
7	3	6	1.76	1.92	-.94	1.45	-1.08	-.95	55	1	2	2.00	1.25	-.73	1.34	.91	-1.21
8	1	2	1.35	1.08	-.69	1.49	-.59	-.53	56	1	3	2.01	-1.63	-.80	1.52	-.48	-1.13
9	1	5	1.56	-.22	-1.22	1.26	1.35	-.46	57	2	6	1.94	1.93	-1.13	1.97	-1.54	-.94
10	2	6	1.88	-1.71	-1.20	1.83	-.49	-.56	58	4	6	1.85	.37	-.58	2.08	.99	-.69
11	1	4	1.78	-1.16	-.91	2.11	1.45	-.46	59	2	5	1.76	-1.47	-.95	1.92	-1.72	-1.33
12	5	6	1.84	1.89	-.54	1.93	1.87	-.97	60	3	5	1.76	-1.87	-.85	1.75	.18	-1.22
13	5	2	1.81	.60	-1.37	.75	-1.78	-1.17	61	5	1	.75	1.82	-1.37	.78	-1.61	-1.12
14	5	4	1.83	-1.99	-.55	1.88	-.69	-1.24	62	3	1	1.80	-.68	-1.31	2.13	-1.03	-.91
15	4	3	.81	-1.49	-.55	1.31	-1.12	-.41	63	6	3	.97	.56	-.98	1.88	1.23	-1.12
16	6	3	1.77	1.36	-.65	1.83	-1.37	-1.31	64	4	3	2.06	-.10	-.97	1.94	.32	-.51
17	4	3	1.98	.85	-.76	1.51	1.78	-1.28	65	6	1	1.98	-1.25	-1.12	1.92	-.29	-.68
18	4	2	1.33	1.60	-.51	1.25	1.45	-1.12	66	3	1	.82	-1.91	-1.22	1.48	.98	-.65
19	4	2	1.76	1.92	-.94	1.45	-1.08	-.95	67	5	6	2.00	1.25	-.73	1.34	.91	-1.21
20	5	6	1.35	1.08	-.69	1.49	-.59	-.53	68	5	4	2.01	-1.63	-.80	1.52	-.48	-1.13
21	5	1	1.56	-.22	-1.22	1.26	1.35	-.46	69	6	2	1.94	1.93	-1.13	1.97	-1.54	-.94
22	6	2	1.88	-1.71	-1.20	1.83	-.49	-.56	70	3	2	1.85	.37	-.58	2.08	.99	-.69
23	5	3	1.78	-1.16	-.91	2.11	1.45	-.46	71	6	1	1.76	-1.47	-.95	1.92	-1.72	-1.33
24	1	2	1.84	1.89	-.54	1.93	1.87	-.97	72	4	1	1.76	-1.87	-.85	1.75	.18	-1.22
25	7	12	1.81	.60	-1.37	.75	-1.78	-1.17	73	7	11	.75	1.82	-1.37	.78	-1.61	-1.12
26	7	9	1.83	-1.99	-.55	1.88	-.69	-1.24	74	10	11	1.80	-.68	-1.31	2.13	-1.03	-.91
27	9	10	.81	-1.49	-.55	1.31	-1.12	-.41	75	8	10	.97	.56	-.98	1.88	1.23	-1.12
28	8	10	1.77	1.36	-.65	1.83	-1.37	-1.31	76	9	12	2.06	-.10	-.97	1.94	.32	-.51
29	9	10	1.98	.85	-.76	1.51	1.78	-1.28	77	8	11	1.98	-1.25	-1.12	1.92	-.29	-.68
30	9	12	1.33	1.60	-.51	1.25	1.45	-1.12	78	10	11	.82	-1.91	-1.22	1.48	.98	-.65
31	9	12	1.76	1.92	-.94	1.45	-1.08	-.95	79	7	8	2.00	1.25	-.73	1.34	.91	-1.21
32	7	8	1.35	1.08	-.69	1.49	-.59	-.53	80	7	9	2.01	-1.63	-.80	1.52	-.48	-1.13
33	7	11	1.56	-.22	-1.22	1.26	1.35	-.46	81	8	12	1.94	1.93	-1.13	1.97	-1.54	-.94
34	8	12	1.88	-1.71	-1.20	1.83	-.49	-.56	82	10	12	1.85	.37	-.58	2.08	.99	-.69
35	7	10	1.78	-1.16	-.91	2.11	1.45	-.46	83	8	11	1.76	-1.47	-.95	1.92	-1.72	-1.33
36	11	12	1.84	1.89	-.54	1.93	1.87	-.97	84	9	11	1.76	-1.87	-.85	1.75	.18	-1.22
37	11	8	1.81	.60	-1.37	.75	-1.78	-1.17	85	11	7	.75	1.82	-1.37	.78	-1.61	-1.12
38	11	10	1.83	-1.99	-.55	1.88	-.69	-1.24	86	9	7	1.80	-.68	-1.31	2.13	-1.03	-.91
39	10	9	.81	-1.49	-.55	1.31	-1.12	-.41	87	12	9	.97	.56	-.98	1.88	1.23	-1.12
40	12	9	1.77	1.36	-.65	1.83	-1.37	-1.31	88	10	9	2.06	-.10	-.97	1.94	.32	-.51
41	10	9	1.98	.85	-.76	1.51	1.78	-1.28	89	12	7	1.98	-1.25	-1.12	1.92	-.29	-.68
42	10	8	1.33	1.60	-.51	1.25	1.45	-1.12	90	9	7	.82	-1.91	-1.22	1.48	.98	-.65
43	10	8	1.76	1.92	-.94	1.45	-1.08	-.95	91	11	12	2.00	1.25	-.73	1.34	.91	-1.21
44	11	12	1.35	1.08	-.69	1.49	-.59	-.53	92	11	10	2.01	-1.63	-.80	1.52	-.48	-1.13
45	11	7	1.56	-.22	-1.22	1.26	1.35	-.46	93	12	8	1.94	1.93	-1.13	1.97	-1.54	-.94
46	12	8	1.88	-1.71	-1.20	1.83	-.49	-.56	94	9	8	1.85	.37	-.58	2.08	.99	-.69
47	11	9	1.78	-1.16	-.91	2.11	1.45	-.46	95	12	7	1.76	-1.47	-.95	1.92	-1.72	-1.33
48	7	8	1.84	1.89	-.54	1.93	1.87	-.97	96	10	7	1.76	-1.87	-.85	1.75	.18	-1.22

APPENDIX B:

PARAMETER MEANS AND STANDARD DEVIATIONS IN THE 12D TESTS

Table B1. Parameter Means and Standard Deviations (in parentheses) in the 12D Tests.

		Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Dim11	Dim12	Total
12D/4	mean	1.64	1.62	1.64	1.64	1.64	1.65	1.64	1.62	1.64	1.64	1.64	1.65	1.64
	α	(.42)	(.40)	(.41)	(.41)	(.44)	(.46)	(.42)	(.40)	(.41)	(.41)	(.44)	(.46)	(.42)
	mean	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03
	δ	(1.43)	(1.40)	(1.38)	(1.42)	(1.37)	(1.39)	(1.43)	(1.40)	(1.38)	(1.42)	(1.37)	(1.39)	(1.40)
	mean	-.96	-.97	-.87	-.96	-.86	-.86	-.96	-.97	-.87	-.96	-.86	-.86	-.91
	τ	(.32)	(.25)	(.26)	(.40)	(.33)	(.25)	(.32)	(.25)	(.26)	(.40)	(.33)	(.25)	(.30)
12D/6	mean	1.64	1.63	1.65	1.64	1.65	1.65	1.64	1.63	1.65	1.64	1.65	1.65	1.64
	α	(.44)	(.44)	(.44)	(.46)	(.46)	(.48)	(.44)	(.44)	(.44)	(.46)	(.46)	(.48)	(.45)
	mean	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03
	δ	(1.16)	(1.16)	(1.23)	(1.32)	(1.31)	(1.26)	(1.16)	(1.16)	(1.23)	(1.32)	(1.31)	(1.26)	(1.24)
	mean	-.96	-.95	-.87	-.94	-.85	-.92	-.96	-.95	-.87	-.94	-.85	-.92	-.91
	τ	(.33)	(.26)	(.32)	(.36)	(.37)	(.27)	(.33)	(.26)	(.32)	(.36)	(.37)	(.27)	(.32)
12D/8	mean	1.64	1.63	1.64	1.64	1.64	1.63	1.64	1.63	1.64	1.64	1.64	1.63	1.64
	α	(.42)	(.42)	(.40)	(.40)	(.42)	(.42)	(.42)	(.42)	(.40)	(.40)	(.42)	(.42)	(.41)
	mean	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03	-.03
	δ	(1.35)	(1.35)	(1.36)	(1.36)	(1.35)	(1.35)	(1.35)	(1.35)	(1.36)	(1.36)	(1.35)	(1.35)	(1.35)
	mean	-.91	-.92	-.92	-.92	-.91	-.92	-.91	-.92	-.92	-.92	-.91	-.92	-.91
	τ	(.32)	(.25)	(.33)	(.33)	(.32)	(.25)	(.32)	(.25)	(.33)	(.33)	(.32)	(.25)	(.30)

APPENDIX C:

PARAMETER RECOVERY RESULTS

Table C1. Person Parameter Recovery Results for the 6D and Uncorrelated Dimensions (ρ_{gen} = .00) Conditions.

Estimation Approach	Sample Size	IPD	Recovery Statistics	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Overall
Direct	400	4	ABS	.48	.46	.49	.46	.48	.46	.47
			RMSE	.64	.60	.65	.61	.63	.61	.62
			PSD	.61	.58	.61	.57	.61	.59	.60
			CORR	.77	.80	.76	.80	.77	.78	.78
		6	ABS	.41	.40	.41	.39	.39	.40	.40
			RMSE	.57	.53	.55	.51	.51	.51	.53
			PSD	.52	.49	.50	.49	.50	.49	.50
			CORR	.82	.85	.84	.86	.86	.86	.85
		8	ABS	.34	.34	.35	.35	.35	.34	.35
			RMSE	.45	.44	.45	.46	.46	.44	.45
			PSD	.44	.43	.44	.43	.44	.43	.44
			CORR	.89	.90	.89	.89	.89	.90	.89
	800	4	ABS	.47	.45	.48	.46	.47	.45	.46
			RMSE	.63	.59	.64	.60	.62	.59	.61
			PSD	.60	.57	.60	.57	.59	.57	.58
			CORR	.77	.80	.77	.80	.78	.81	.79
		6	ABS	.40	.39	.40	.39	.39	.39	.39
			RMSE	.56	.52	.53	.52	.52	.51	.53
			PSD	.51	.48	.50	.48	.49	.48	.49
			CORR	.83	.86	.85	.86	.85	.86	.85
		8	ABS	.35	.33	.34	.35	.34	.33	.34
			RMSE	.46	.43	.45	.46	.45	.43	.45
			PSD	.43	.42	.43	.43	.43	.42	.43
			CORR	.89	.90	.89	.89	.89	.90	.89
Two-Step	400	4	ABS	.48	.46	.49	.45	.47	.45	.47
			RMSE	.64	.60	.65	.60	.62	.60	.62
			PSD	.60	.56	.59	.56	.58	.57	.58
			CORR	.77	.80	.76	.80	.78	.80	.79
		6	ABS	.40	.39	.41	.38	.39	.40	.40
			RMSE	.55	.53	.56	.52	.52	.52	.53
			PSD	.51	.47	.49	.48	.49	.48	.49
			CORR	.83	.85	.84	.86	.86	.86	.85
		8	ABS	.34	.34	.35	.35	.35	.34	.35
			RMSE	.45	.44	.45	.46	.46	.44	.45
			PSD	.43	.42	.42	.42	.43	.42	.42
			CORR	.89	.90	.89	.89	.89	.90	.89

Table C1 (Continued)

800	4	ABS	.47	.45	.48	.45	.47	.45	.46
		RMSE	.63	.59	.63	.60	.62	.59	.61
		PSD	.59	.56	.59	.55	.58	.56	.57
		CORR	.77	.81	.78	.80	.79	.81	.79
	6	ABS	.40	.39	.40	.39	.39	.39	.39
		RMSE	.55	.52	.53	.51	.52	.51	.52
		PSD	.50	.48	.49	.47	.49	.48	.49
		CORR	.83	.86	.85	.86	.86	.86	.85
	8	ABS	.34	.33	.34	.34	.34	.33	.34
		RMSE	.45	.43	.45	.45	.45	.43	.44
		PSD	.43	.42	.43	.42	.43	.41	.42
		CORR	.89	.90	.89	.89	.89	.90	.89

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

Table C2. Person Parameter Recovery Results for the 12D and Uncorrelated Dimensions (ρ_{gen} = .00) Conditions.

Estimation Approach	Sample Size	IPD	Recovery Statistics	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	Overall
Direct	400	4	ABS	.50	.47	.51	.47	.48	.47	.49	.47	.52	.48	.49	.46	.48
			RMSE	.66	.62	.67	.63	.64	.61	.66	.62	.69	.63	.64	.61	.64
			PSD	.63	.58	.64	.59	.62	.60	.63	.59	.66	.58	.62	.61	.61
			CORR	.75	.79	.74	.77	.77	.79	.75	.79	.74	.79	.77	.78	.77
		6	ABS	.41	.39	.40	.39	.40	.39	.41	.39	.40	.39	.40	.39	.40
			RMSE	.57	.51	.53	.52	.52	.53	.57	.53	.54	.52	.52	.51	.53
			PSD	.51	.49	.50	.49	.50	.49	.52	.48	.50	.48	.50	.49	.50
			CORR	.83	.86	.85	.85	.85	.85	.81	.85	.84	.86	.85	.86	.85
		8	ABS	.35	.34	.35	.34	.35	.34	.35	.34	.35	.35	.35	.35	.35
			RMSE	.46	.45	.45	.45	.46	.44	.46	.45	.46	.46	.46	.45	.45
			PSD	.43	.43	.43	.42	.44	.42	.44	.43	.42	.43	.43	.42	.43
			CORR	.89	.89	.89	.89	.88	.89	.89	.89	.89	.89	.88	.89	.89
	800	4	ABS	.50	.47	.50	.47	.49	.48	.50	.47	.51	.47	.49	.47	.49
			RMSE	.67	.61	.66	.62	.64	.63	.67	.62	.67	.62	.64	.62	.64
			PSD	.62	.58	.63	.58	.62	.60	.64	.59	.63	.58	.61	.59	.61
			CORR	.75	.79	.75	.79	.77	.79	.74	.79	.74	.79	.77	.79	.77
		6	ABS	.41	.38	.40	.39	.40	.39	.40	.38	.40	.39	.40	.39	.39
			RMSE	.57	.51	.54	.51	.52	.51	.56	.51	.54	.51	.52	.52	.53
			PSD	.50	.48	.49	.48	.49	.48	.51	.48	.49	.48	.49	.49	.49
			CORR	.82	.86	.84	.86	.85	.86	.82	.86	.84	.86	.85	.86	.85
		8	ABS	.35	.34	.34	.34	.35	.33	.35	.34	.34	.35	.35	.34	.34
			RMSE	.46	.44	.45	.44	.46	.43	.46	.44	.45	.45	.46	.44	.45
			PSD	.43	.42	.42	.42	.43	.41	.43	.41	.42	.43	.43	.41	.42
			CORR	.89	.90	.90	.90	.89	.90	.89	.90	.89	.89	.89	.90	.90
Two-Step	400	4	ABS	.48	.46	.49	.46	.47	.46	.48	.47	.49	.47	.47	.46	.47
			RMSE	.64	.61	.64	.62	.62	.59	.65	.61	.66	.62	.63	.60	.62
			PSD	.60	.55	.59	.56	.58	.56	.60	.56	.59	.56	.57	.57	.57
			CORR	.76	.80	.77	.78	.78	.80	.76	.79	.76	.79	.79	.79	.78
		6	ABS	.41	.39	.40	.39	.40	.39	.41	.39	.40	.40	.40	.39	.40
			RMSE	.56	.51	.54	.52	.53	.53	.56	.52	.53	.53	.53	.51	.53
			PSD	.50	.47	.49	.47	.49	.48	.51	.48	.49	.47	.49	.48	.49
			CORR	.83	.86	.85	.86	.85	.85	.82	.86	.85	.85	.85	.86	.85
		8	ABS	.35	.34	.35	.34	.35	.34	.35	.34	.35	.35	.35	.35	.35
			RMSE	.45	.45	.46	.45	.46	.45	.45	.45	.46	.46	.47	.46	.46
			PSD	.43	.41	.42	.42	.43	.42	.43	.42	.42	.41	.43	.42	.42
			CORR	.89	.89	.89	.90	.89	.89	.89	.90	.89	.89	.88	.89	.89
	800	4	ABS	.48	.46	.49	.46	.47	.46	.48	.45	.49	.46	.47	.46	.47
			RMSE	.65	.60	.65	.61	.63	.61	.65	.61	.65	.61	.63	.61	.63
			PSD	.59	.56	.60	.55	.58	.56	.59	.56	.59	.56	.58	.56	.57
			CORR	.76	.80	.76	.80	.78	.80	.76	.80	.76	.79	.78	.80	.78
		6	ABS	.40	.38	.40	.38	.40	.39	.40	.38	.40	.39	.39	.39	.39
			RMSE	.55	.50	.53	.51	.52	.51	.55	.51	.54	.51	.52	.51	.52
			PSD	.50	.48	.50	.47	.49	.47	.51	.47	.50	.47	.48	.48	.49
			CORR	.83	.86	.84	.86	.85	.86	.83	.86	.84	.86	.86	.86	.85
		8	ABS	.35	.34	.34	.34	.35	.33	.35	.33	.34	.34	.35	.34	.34
			RMSE	.46	.44	.44	.44	.46	.43	.46	.44	.45	.45	.46	.44	.45
			PSD	.42	.41	.42	.42	.43	.41	.43	.41	.42	.42	.43	.41	.42
			CORR	.89	.90	.90	.90	.89	.90	.89	.90	.89	.89	.89	.90	.90

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

Table C3. Statement Parameter Recovery Results for the .30 Correlated Dimensions Conditions.

Sample Size	Dimensions	IPD	Recovery Statistics	Two-step Approach			Direct Approach		
				α	δ	τ	α	δ	τ
400	6	4	ABS	.22	.20	.18	.25	.24	.21
			RMSE	.29	.26	.24	.32	.31	.26
			PSD	.32	.31	.28	.38	.36	.72
			CORR	.78	.99	.79	.71	.98	.61
		6	ABS	.21	.17	.15	.24	.20	.22
			RMSE	.27	.23	.21	.30	.26	.27
			PSD	.28	.24	.21	.33	.29	.71
			CORR	.83	.99	.86	.79	.98	.62
		8	ABS	.20	.17	.15	.22	.20	.20
			RMSE	.25	.23	.21	.28	.27	.25
			PSD	.26	.24	.22	.31	.28	.71
			CORR	.83	.99	.83	.79	.98	.63
	12	4	ABS	.23	.21	.18	.30	.28	.22
			RMSE	.29	.27	.24	.37	.38	.27
			PSD	.32	.31	.28	.40	.45	.72
			CORR	.78	.98	.79	.59	.96	.59
		6	ABS	.21	.17	.16	.25	.21	.22
			RMSE	.27	.23	.21	.31	.28	.27
			PSD	.28	.24	.21	.33	.30	.71
			CORR	.83	.99	.84	.77	.98	.63
		8	ABS	.20	.17	.16	.23	.20	.20
			RMSE	.25	.23	.22	.29	.27	.25
			PSD	.26	.24	.22	.30	.28	.71
			CORR	.83	.99	.82	.77	.98	.63
800	6	4	ABS	.19	.17	.15	.23	.19	.20
			RMSE	.24	.22	.20	.28	.25	.25
			PSD	.25	.24	.21	.30	.28	.71
			CORR	.85	.99	.83	.80	.99	.65
		6	ABS	.16	.13	.12	.19	.15	.21
			RMSE	.20	.18	.16	.24	.20	.26
			PSD	.21	.17	.15	.25	.21	.70
			CORR	.90	.99	.90	.86	.99	.65
		8	ABS	.15	.13	.12	.18	.15	.20
			RMSE	.19	.19	.17	.22	.21	.24
			PSD	.19	.17	.16	.23	.21	.71
			CORR	.90	.99	.88	.87	.99	.65
	12	4	ABS	.19	.17	.15	.27	.24	.20
			RMSE	.24	.22	.20	.33	.32	.25
			PSD	.25	.24	.21	.33	.36	.72
			CORR	.84	.99	.84	.69	.97	.64
		6	ABS	.16	.14	.12	.21	.17	.21
			RMSE	.20	.19	.17	.26	.24	.26
			PSD	.21	.18	.16	.26	.23	.70
			CORR	.91	.99	.90	.84	.98	.65
		8	ABS	.15	.13	.12	.18	.17	.19
			RMSE	.19	.19	.17	.23	.22	.24
			PSD	.19	.17	.16	.23	.21	.71
			CORR	.90	.99	.88	.85	.99	.66

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

Table C4. Person Parameter Recovery Results for the 6D and .30 Correlated Dimensions Conditions.

Estimation Approach	Sample Size	IPD	Recovery Statistics	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Overall
Direct	400	4	ABS	.46	.45	.48	.45	.45	.45	.46
			RMSE	.60	.58	.63	.59	.59	.58	.60
			PSD	.57	.56	.60	.56	.58	.57	.57
			CORR	.81	.81	.78	.81	.81	.81	.80
		6	ABS	.39	.39	.40	.37	.38	.38	.38
			RMSE	.52	.51	.53	.49	.50	.50	.51
			PSD	.49	.49	.49	.48	.49	.49	.49
			CORR	.85	.86	.85	.87	.87	.87	.86
		8	ABS	.34	.33	.34	.34	.34	.33	.34
			RMSE	.44	.42	.44	.44	.44	.43	.44
			PSD	.43	.42	.42	.42	.42	.42	.42
			CORR	.89	.91	.90	.90	.90	.90	.90
	800	4	ABS	.45	.44	.47	.44	.45	.44	.45
			RMSE	.59	.58	.61	.58	.59	.57	.59
			PSD	.58	.56	.59	.55	.56	.55	.56
			CORR	.80	.81	.79	.82	.81	.82	.81
		6	ABS	.38	.38	.38	.37	.38	.37	.38
			RMSE	.51	.50	.50	.49	.49	.49	.50
			PSD	.48	.48	.49	.47	.48	.47	.48
			CORR	.86	.87	.87	.88	.87	.87	.87
		8	ABS	.33	.33	.33	.33	.33	.33	.33
			RMSE	.43	.42	.43	.43	.43	.42	.43
			PSD	.42	.41	.42	.42	.42	.41	.41
			CORR	.90	.91	.90	.90	.90	.91	.90
Two-Step	400	4	ABS	.45	.44	.48	.45	.45	.45	.45
			RMSE	.59	.58	.63	.59	.59	.58	.59
			PSD	.57	.54	.58	.54	.55	.55	.55
			CORR	.81	.81	.78	.81	.81	.82	.81
		6	ABS	.38	.38	.40	.38	.38	.38	.38
			RMSE	.51	.51	.53	.50	.50	.50	.51
			PSD	.48	.47	.48	.47	.47	.47	.47
			CORR	.86	.86	.85	.87	.87	.87	.86
		8	ABS	.34	.33	.34	.34	.34	.33	.34
			RMSE	.44	.43	.44	.44	.44	.43	.44
			PSD	.42	.41	.41	.42	.41	.41	.41
			CORR	.90	.90	.90	.90	.90	.90	.90
	800	4	ABS	.45	.44	.47	.44	.45	.44	.45
			RMSE	.59	.57	.61	.58	.59	.58	.59
			PSD	.57	.54	.58	.53	.55	.54	.55
			CORR	.80	.82	.79	.82	.81	.82	.81
		6	ABS	.38	.38	.38	.37	.38	.37	.38
			RMSE	.51	.49	.50	.49	.49	.49	.50
			PSD	.47	.47	.47	.46	.47	.46	.47
			CORR	.86	.87	.87	.88	.87	.88	.87
		8	ABS	.33	.33	.33	.33	.33	.33	.33
			RMSE	.43	.42	.43	.43	.43	.42	.43
			PSD	.41	.41	.41	.41	.41	.40	.41
			CORR	.90	.91	.90	.91	.90	.91	.90

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

Table C5. Person Parameter Recovery Results for the 12D and .30 Correlated Dimensions Conditions.

Estimation Approach	Sample Size	IPD	Recovery Statistics	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	Overall
Direct	400	4	ABS	.47	.46	.54	.46	.47	.45	.46	.45	.57	.45	.47	.46	.48
			RMSE	.62	.60	.70	.60	.62	.58	.61	.59	.74	.59	.62	.60	.62
			PSD	.63	.59	.70	.59	.61	.60	.62	.59	.74	.59	.63	.62	.62
			CORR	.79	.80	.73	.79	.78	.81	.79	.81	.68	.81	.80	.80	.78
		6	ABS	.39	.38	.39	.38	.39	.38	.39	.38	.39	.38	.39	.38	.39
			RMSE	.54	.51	.51	.50	.50	.50	.53	.50	.51	.49	.50	.50	.51
			PSD	.50	.49	.50	.50	.50	.49	.51	.49	.52	.48	.50	.52	.50
			CORR	.85	.86	.87	.87	.87	.87	.85	.87	.87	.87	.87	.87	.86
		8	ABS	.34	.33	.34	.33	.35	.34	.34	.33	.34	.34	.34	.34	.34
			RMSE	.44	.43	.44	.43	.45	.44	.44	.43	.44	.43	.45	.44	.44
			PSD	.43	.43	.43	.43	.44	.42	.44	.43	.43	.43	.44	.43	.43
			CORR	.90	.90	.89	.90	.89	.90	.90	.90	.90	.90	.89	.90	.90
	800	4	ABS	.47	.47	.52	.45	.48	.46	.50	.46	.51	.46	.48	.45	.48
			RMSE	.61	.61	.68	.60	.63	.60	.65	.61	.67	.60	.62	.59	.62
			PSD	.60	.60	.67	.59	.63	.59	.67	.59	.66	.59	.61	.59	.62
			CORR	.79	.80	.73	.80	.78	.80	.76	.79	.74	.80	.79	.81	.78
		6	ABS	.38	.38	.41	.38	.39	.38	.38	.37	.41	.38	.38	.37	.38
			RMSE	.52	.49	.54	.49	.50	.50	.51	.49	.54	.49	.50	.48	.50
			PSD	.50	.49	.54	.49	.49	.49	.50	.49	.54	.48	.49	.49	.50
			CORR	.85	.87	.86	.87	.86	.87	.85	.87	.85	.87	.87	.87	.86
		8	ABS	.34	.33	.33	.33	.34	.32	.34	.33	.33	.33	.33	.33	.33
			RMSE	.44	.42	.43	.43	.44	.42	.44	.42	.43	.43	.43	.43	.43
			PSD	.43	.42	.43	.43	.43	.42	.43	.42	.43	.43	.43	.42	.43
			CORR	.90	.90	.90	.90	.90	.91	.90	.90	.90	.90	.90	.91	.90
Two-Step	400	4	ABS	.46	.45	.47	.45	.45	.44	.45	.45	.47	.44	.45	.44	.45
			RMSE	.60	.59	.61	.59	.59	.57	.60	.59	.62	.58	.60	.58	.59
			PSD	.59	.55	.58	.55	.57	.55	.59	.55	.59	.55	.57	.56	.57
			CORR	.80	.81	.80	.80	.80	.81	.79	.81	.79	.82	.81	.81	.81
		6	ABS	.39	.38	.39	.38	.38	.38	.38	.38	.38	.38	.38	.37	.38
			RMSE	.53	.50	.51	.50	.50	.50	.51	.50	.50	.50	.50	.49	.50
			PSD	.48	.47	.48	.47	.47	.47	.49	.47	.48	.46	.47	.47	.47
			CORR	.85	.87	.86	.87	.87	.87	.86	.87	.87	.87	.87	.87	.87
		8	ABS	.33	.33	.34	.33	.34	.34	.34	.33	.34	.34	.34	.33	.34
			RMSE	.43	.43	.44	.43	.45	.44	.44	.43	.44	.44	.45	.43	.44
			PSD	.42	.41	.42	.41	.42	.41	.42	.41	.41	.41	.42	.41	.41
			CORR	.90	.90	.90	.90	.89	.90	.90	.90	.90	.90	.90	.90	.90
	800	4	ABS	.45	.44	.47	.44	.45	.45	.46	.44	.47	.44	.45	.44	.45
			RMSE	.59	.58	.62	.59	.59	.58	.60	.58	.62	.58	.59	.58	.59
			PSD	.57	.55	.60	.55	.57	.56	.58	.56	.59	.55	.57	.56	.57
			CORR	.81	.81	.78	.81	.80	.81	.80	.81	.79	.82	.81	.82	.81
		6	ABS	.38	.37	.39	.37	.38	.38	.38	.37	.38	.37	.37	.37	.38
			RMSE	.51	.48	.51	.49	.50	.50	.51	.49	.50	.49	.49	.48	.50
			PSD	.49	.47	.48	.47	.48	.47	.49	.47	.49	.46	.48	.47	.48
			CORR	.85	.87	.86	.88	.87	.87	.86	.87	.86	.88	.87	.88	.87
		8	ABS	.33	.33	.33	.33	.33	.32	.34	.33	.33	.33	.33	.33	.33
			RMSE	.43	.42	.43	.42	.43	.42	.43	.42	.42	.43	.43	.43	.43
			PSD	.42	.41	.42	.41	.42	.41	.42	.41	.41	.41	.42	.41	.41
			CORR	.90	.91	.90	.91	.90	.91	.90	.91	.91	.90	.90	.91	.90

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