GGUM-RANK Statement and Person Parameter Estimation With Multidimensional Forced Choice Triplets

Applied Psychological Measurement
2019, Vol. 43(3) 226–240
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618768294
journals.sagepub.com/home/apm

Philseok Lee¹, Seang-Hwane Joo², Stephen Stark³ and Oleksandr S. Chernyshenko⁴

Abstract

Historically, multidimensional forced choice (MFC) measures have been criticized because conventional scoring methods can lead to ipsativity problems that render scores unsuitable for interindividual comparisons. However, with the recent advent of item response theory (IRT) scoring methods that yield normative information, MFC measures are surging in popularity and becoming important components in high-stake evaluation settings. This article aims to add to burgeoning methodological advances in MFC measurement by focusing on statement and person parameter recovery for the GGUM-RANK (generalized graded unfolding-RANK) IRT model. Markov chain Monte Carlo (MCMC) algorithm was developed for estimating GGUM-RANK statement and person parameters directly from MFC rank responses. In simulation studies, it was examined that how the psychometric properties of statements composing MFC items, test length, and sample size influenced statement and person parameter estimation; and it was explored for the benefits of measurement using MFC triplets relative to pairs. To demonstrate this methodology, an empirical validity study was then conducted using an MFC triplet personality measure. The results and implications of these studies for future research and practice are discussed.

Keywords

multidimensional forced choice, noncognitive assessment, parameter recovery, item response theory, Markov chain Monte Carlo, ideal point, faking

To control response biases and rater errors, multidimensional forced choice (MFC) measures have been proposed as an alternative to Likert-type scales for noncognitive assessment (Stark, Chernyshenko, & Drasgow, 2005). MFC measures commonly present statements in blocks of two, three, or four. Within the blocks, statements representing different dimensions are matched on social desirability so the best answers are not obvious. The respondent's task is to rank the

¹South Dakota State University, Brookings, USA

²Katholieke Universiteit Leuven, Belgium

³University of South Florida, Tampa, USA

⁴Nanyang Technological University, Singapore

Corresponding Author:

Philseok Lee, Psychology Department, South Dakota State University, Scobey Hall 332, Brookings, SD 57007, USA.
Email: Philseok.Lee@sdstate.edu

statements in each block from most to least “like me.” Historically, MFC measures have been criticized because conventional scoring methods can lead to ipsativity problems that render scores unsuitable for interindividual comparisons (Hicks, 1970). However, over the last decade, advances in item response theory (IRT) have made it possible to derive normative information from the MFC format (e.g., Brown & Maydeu-Olivares, 2011; de la Torre, Ponsoda, Leenen, & Hontangas, 2012; Stark et al., 2005).

Stark et al. (2005) proposed the multi-unidimensional pairwise preference (MUPP) model for constructing and scoring pairwise preference responses. They estimated discrimination (α), location (δ), and threshold (τ) parameters for individually presented statements using the generalized graded unfolding model (GGUM, Roberts, Donoghue, & Laughlin, 2000) and then used those statement parameters (in conjunction with social desirability ratings) to construct and score multidimensional pairwise preference tests designed to reducing faking (i.e., two-step estimation process). More recently, 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). In their article, which explored person parameter estimation assuming GGUM statement parameters were known, Hontangas et al. referred to this model as the RANK model. This model is henceforth referred as the GGUM-RANK (generalized graded unfolding-RANK) to differentiate it from possible non-GGUM adaptations.

Although these modeling developments have helped to advance MFC research and practice, there remain many unexplored issues concerning parameter estimation. By scaling a statement pool using a unidimensional model prior to MFC testing, Stark et al.’s method avoids the complexity of estimating statement parameters directly from forced choice responses via marginal maximum likelihood (MML) methods, which require complicated derivatives, or Markov chain Monte Carlo (MCMC) methods that require long run-times. This approach also facilitates MFC computerized adaptive testing (CAT), as well as nonadaptive testing with many forms, because any number of MFC tests can be created and scored once a statement pool has been calibrated. On the contrary, this approach may be considered impractical if just one MFC form is needed. In such situations, it may be better to construct a single MFC form having some extra items based on expert judgment, administer the form to a selected sample of respondents, estimate statement parameters directly from the MFC responses, and cull any items identified as problematic before scoring for assessment purposes. Importantly, such an approach would allow a test developer to evaluate the actual items (i.e., pairs, triplets, or tetrads) presented to respondents. In addition, by using simulation methodology, one could systematically explore how statement and person parameter estimation depends on the combinations of statements forming the MFC blocks.

Purpose of Research

An MCMC estimation algorithm was developed and evaluated for GGUM-RANK parameters, and its usefulness was demonstrated for empirical research. Unlike previously published studies that focused exclusively on scoring (e.g., Hontangas et al., 2015; Stark et al., 2005), this research examined the recovery of GGUM-RANK *statement and person* parameter estimation directly from the MFC responses (i.e., direct estimation process). It is noted that the direct estimation process is similarly effective or even better than two-step estimation process for person parameter recovery (Seybert, 2013). Three studies were conducted. Study 1 examined the recovery of GGUM-RANK parameters with MFC triplet measures while manipulating sample

size, test length, intrablock discrimination, and intrablock location parameter variability. Rather than 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. Study 2 compared GGUM-RANK parameter recovery for MFC pair and triplet measures of different test lengths and sample sizes. Because there is growing interest in the benefits of triplet measures relative to pairs for noncognitive assessment (e.g., Anguiano-Carrasco, MacCann, Geiger, Seybert, & Roberts, 2014; Guenole, Brown, & Cooper, 2016), it was particularly important to see how much the added complexity of triplets might improve person parameter estimation. “Study 3” was an empirical example exploring construct and criterion-related validities of an MFC triplet personality measure calibrated and scored using the new GGUM-RANK estimation method. Before discussing these studies and results, MUPP and GGUM-RANK models were reviewed briefly.

The MUPP Model

Stark et al. (2005) proposed the MUPP model. 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 or her, the respondent considers each statement separately. The probability of preferring statement j over statement k ($j > k$) given his or her scores on the respective dimensions (d) is given by,

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

where $P_{jk}\{1, 0\}$ = 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 θ_{d_j} , and $P_k\{1\}$, $P_k\{0\}$ = probability of endorsing/not endorsing statement k at θ_{d_k} .

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*. For details concerning use of the GGUM in connection with the MUPP model, see Stark (2002) and Stark et al. (2005).

The GGUM-RANK Model for MFC Triplet Measures

Following Luce (1959), who proposed that the probability of a set of ranks can be viewed as a 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 a triplet example, the probability of the hypothetical ranking $A > B > C$ would be given by the following sequence of PICK decisions:

$$\begin{aligned} P_{(A>B>C)}(\theta_{d_A}, \theta_{d_B}, \theta_{d_C}) &= P_{(A>B, C)} \times P_{(B>C)} \\ &= \frac{P_A(1)P_B(0)P_C(0)}{P_A(1)P_B(0)P_C(0) + P_A(0)P_B(1)P_C(0) + P_A(0)P_B(0)P_C(1)} \times \frac{P_B(1)P_C(0)}{P_B(1)P_C(0) + P_B(0)P_C(1)}, \end{aligned} \quad (2)$$

where $P_A(1)$, $P_B(1)$, and $P_C(1)$ represent the probabilities of endorsing statements A, B, and C, respectively; and $P_A(0)$, $P_B(0)$, and $P_C(0)$ represent the probabilities of not endorsing statements A, B, and C, respectively. 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 the GGUM for dichotomous responses. The same logic would apply to the remaining five possible ranking responses ($A > C > B$, $B > A > C$, $B > C > A$, $C > A > B$, and $C > B > A$). For further details on the RANK model, please see Hontangas et al. (2015).

Study 1

Study 1 investigated the accuracy of MCMC statement and person parameter recovery using a Metropolis-Hasting Within Gibbs (MHWG; Tierney, 1994) algorithm developed for GGUM-RANK triplet responses in the Ox programming language (Doornik, 2009). This algorithm is available in Appendix A of the Online Supplementary material.

Simulation Study Design

MFC test dimensionality was set at 10 dimensions. Four independent variables were fully crossed to produce 16 experimental conditions: (a) *sample size* (250, 500), (b) *test length* (30 triplets, 60 triplets), (c) *inrablock discrimination* (low = α parameters sampled randomly from a uniform distribution, $U(0.75, 1.25)$; high = α parameters sampled randomly from $U(1.75, 2.25)$), and (d) *inrablock location* (δ) *standard deviation* (SD ; small ≈ 0.3 , large ≈ 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 inrablock location SD conditions, half of the δ parameters were sampled from $U(-2, 0)$, and the other half were sampled from $U(0, 2)$. Then the requirements for large and small location SD conditions were fulfilled by mixing the generated δ parameters. The α , δ , and τ parameters represent item discrimination, item location, and the location of the subjective response category thresholds, respectively. Because pilot testing revealed that one replication would take 11 to 47 hr, depending on the condition, the number of replications in each condition was set at 20.

MFC Tests Constructed for This Simulation

Four 30-triplet and four 60-triplet MFC tests were created by crossing levels of inrablock discrimination and location SD . The test specifications and generating parameters are provided in the Online Supplemental Appendices B and C.

Data Generation

GGUM-RANK response data were generated using an Ox computer program (Doornik, 2009). For each respondent, a vector of 10 latent trait scores (θ) was sampled from a multivariate standard normal distribution with the covariances among dimensions set to zero. Six possible triplet rank responses were assigned and consecutive numerical codes from 1 to 6. Using the trait scores and statement parameters for each experimental condition, the GGUM-RANK probabilities were computed (see Equation 2) and used to demarcate segments of a cumulative probability continuum. To generate an MFC rank response, a random number was sampled from $U(0, 1)$ and the numerical code for the rank corresponding to the segment containing that random number was chosen.

Prior Distributions and Initial Parameter Values

Four-parameter beta priors $\{1.5, 1.5, .25, 3\}$, $\{2, 2, -4, 4\}$, and $\{2, 2, -3, 1\}$ were used for α , δ , and τ estimation, respectively. For the person parameters associated with each dimension (d), a $N(0,1)$ prior was used. To start MCMC estimation, all θ parameters were initialized to zero. δ parameters were initialized to -1 or 1 , in accordance with the sign of the true parameter. All α and τ parameters were initialized to 1 and -1 , respectively.

Convergence Checks and Indices of Estimation Accuracy

Convergence was checked using $\hat{R}$ statistics (Gelman & Rubin, 1992), which assess the ratio of between-chain variation to within-chain variation. If the chains have converged, the between-chain and within-chain variation will be near one; otherwise, larger ratios will be observed. Strict convergence is met when $\hat{R} < 1.20$ for all parameters (de la Torre et al., 2012). Each chain involved a total of 30,000 iterations; the first 15,000 iterations were treated as burn-in. Post-burn-in samples were used to compute the mean and *SD* of the posterior distributions. To evaluate the efficacy of GGUM-RANK parameter estimation, absolute biases (ABS), root mean square errors (RMSEs), correlations (CORR) between true and estimated parameters, and posterior standard deviations (PSDs) were computed for the parameter estimates across replications, as in previous research with related ideal point IRT models. Smaller ABS, RMSE, and PSD values indicate better parameter recovery.

Results

Convergence rates were generally high, ranging from .93 to 1.00 across conditions. In the few instances where a parameter did not meet the convergence criterion, the parameter estimate was excluded from the calculation of the recovery statistics. Table 1 presents the parameter recovery results for GGUM-RANK statement parameter estimation, averaged over replications. Across all conditions, ABS ranged from .13 to .23, .11 to .27, and .16 to .21 for α , δ , and τ , respectively. The corresponding RMSEs ranged from .16 to .31, .14 to .35, and .20 to .26. The 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. It was noted that the low CORRs between generating and estimated α parameters were due in part to the restricted range of the generating parameters, and the relatively poor recovery of τ parameters is consistent with previous simulation results using the dichotomous version of the GGUM (e.g., de la Torre, Stark, & Chernyshenko, 2006; Joo, Lee, & Stark, 2017). As expected, recovery of α and δ parameters was improved with test length (60-triplet better than 30-triplet), sample size (500 better than 250), and intrablock discrimination (high better than low), although the pattern of results for τ was inconsistent. Interestingly, intrablock location *SD* did not seem to influence the recovery results. It can be seen that ABS, RMSE, PSD, and CORR values were nearly the same across the small and large intrablock location *SD* conditions.

Table 2 presents parameter recovery statistics for latent trait scores (θ s), averaged over replications. The averages across the 10 dimensions are shown in the last column. In general, parameter recovery improved as test length and intrablock discrimination increased. However, there was no noteworthy improvement with larger samples or location *SD*s. 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

Table 1. Statement Parameter Recovery With MFC Triplet Tests in Study 1.

Sample	Discrimination	Location SD	Recovery statistics	30 blocks			60 blocks		
				Statement parameter			Statement parameter		
				α	δ	τ	α	δ	τ
250	High	Large	ABS	.22	.17	.18	.19	.15	.17
			RMSE	.28	.22	.22	.24	.19	.20
			PSD	.31	.22	.51	.23	.19	.52
			CORR	.37	.99	.80	.42	.99	.82
		Small	ABS	.21	.17	.19	.19	.16	.16
			RMSE	.27	.21	.23	.25	.20	.20
			PSD	.30	.23	.51	.23	.20	.51
			CORR	.41	.99	.78	.42	.99	.81
	Low	Large	ABS	.23	.27	.21	.21	.24	.19
			RMSE	.30	.34	.26	.26	.30	.24
			PSD	.29	.42	.53	.27	.35	.52
			CORR	.46	.96	.69	.58	.97	.72
		Small	ABS	.23	.27	.20	.20	.23	.20
			RMSE	.31	.35	.25	.25	.30	.24
			PSD	.29	.42	.53	.27	.35	.52
			CORR	.44	.96	.69	.58	.97	.74
500	High	Large	ABS	.16	.12	.19	.13	.11	.18
			RMSE	.20	.16	.23	.16	.14	.21
			PSD	.21	.16	.51	.16	.13	.51
			CORR	.46	.99	.79	.53	.99	.82
		Small	ABS	.16	.13	.20	.13	.11	.19
			RMSE	.20	.16	.23	.16	.14	.22
			PSD	.21	.17	.50	.16	.14	.51
			CORR	.54	.99	.80	.55	.99	.80
	Low	Large	ABS	.18	.23	.20	.16	.19	.19
			RMSE	.23	.29	.25	.20	.24	.23
			PSD	.24	.36	.52	.20	.26	.52
			CORR	.61	.97	.74	.69	.98	.76
		Small	ABS	.18	.23	.19	.15	.19	.19
			RMSE	.22	.30	.24	.19	.25	.23
			PSD	.23	.36	.53	.20	.26	.52
			CORR	.54	.97	.71	.68	.98	.74

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

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.

Study 2

Study 2 was conducted to examine the benefits of parameter estimation with MFC triplets relative to pairs. Statement and person parameter recovery were assessed in a simulation study that crossed two independent variables: *sample size* (250, 500) and *test type* (30-triplet, 30-pair, 90-pair).

Table 2. Person Parameter Recovery With MFC Triplet Tests in Study 1.

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

(continued)

Table 2. (continued)

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

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

Table 3. Statement Parameter Recovery for MFC Triplet and Pair Tests in Study 2.

Test type	Sample	Recovery statistics	Statement parameter		
			α	δ	τ
30-triplet	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	.24	.98	.69
90-pair	250	ABS	.32	.28	.17
		RMSE	.38	.33	.20
		PSD	.37	.28	.62
		CORR	.29	.98	.62
	500	ABS	.21	.20	.13
		RMSE	.32	.24	.16
		PSD	.29	.20	.61
		CORR	.38	.99	.71

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

MFC Test Design and Analyses

From Study 1, the 30-triplet test in the high intrablock discrimination, high intrablock location *SD* 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. The average generating parameters for each dimension in the 30-pair and 90-pair test type conditions were similar. The process for data generation, parameter estimation, and analysis was the same as in Study 1. The test specifications and generating parameters for the 30-pair and 90-pair tests are provided in the Online Supplemental materials (see Appendices D and E).

Results

Overall convergence rates approached 100% within 30,000 iterations. Table 3 presents the 30-pair and 90-pair parameter recovery results. (The results for the selected 30-triplet test in Study 1 are shown again for convenience.) 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 250 and 500 sample size conditions, the 30-triplet measure had higher CORR and lower ABS, RMSE, and PSD values.

Table 4. Person Parameter Recovery for MFC Triplet and Pair Tests in Study 2.

Test type	Sample	Recovery	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6	Dim7	Dim8	Dim9	Dim10	Overall
		statistics											
30-triplet	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. MFC = multidimensional forced choice; ABS = absolute bias; RMSE = root mean square error; PSD = posterior standard deviation; CORR = correlation between true and estimated parameter.

Table 4 presents the person parameter (θ) 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, θ s were estimated better with 90 pairs than with 30 pairs. And, perhaps most importantly, 90 pairs were needed to achieve levels of ABS, RMSE, PSD, and CORR similar to the 30-triplet condition. Specifically, with a sample size of 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.

Study 3: Empirical Example

Studies 1 and 2 examined statement and person parameter recovery using GGUM-RANK MFC triplet and pair measures. Although these simulations provided insights into MFC test construction practices, they provided no evidence concerning the comparability of GGUM-RANK MFC and Likert-type scale scores with real examinees. To address that issue, an empirical validity investigation was conducted using MFC triplet and single-statement (SS) personality measures and relevant SS criterion variables.

Procedure

A total of 60 statements measuring the Big Five factor markers (Goldberg, 1992) were selected from the International Personality Item Pool, translated into Korean, and two personality

measures were created. A total of 12 statements were selected to measure each of the five factors. The first was a 60-item SS Big Five measure that required respondents to indicate their level of agreement using a 5-point Likert-type format. The second was a 20-triplet MFC measure, in which each triplet measured three different personality dimensions. The statements forming the triplets were matched as closely as possible on social desirability, computed as the difference in the Likert-type item scores between “honest” and “fake good” administrations using a within-subjects design ($N = 205$). Respondents were instructed to rank the statements in each MFC triplet from 1 (*most like you*) to 3 (*least like you*). The SS and MFC personality measures were administered to 417 Korean college students, and a smaller subset of students ($N = 235$) also completed Korean versions of SS criterion measures that included life satisfaction (The Satisfaction With Life Scale; Diener, Emmons, Larsen, & Griffin, 1985), positive or negative feeling (Scale of Positive and Negative Experience; Diener et al., 2010), aggression (Buss-Perry Aggression Questionnaire; Bryant & Smith, 2001), and RIASEC vocational interests (Heo, 2011).

Analyses

The SS personality and SS criterion measures were scored using the conventional summative approach. To score the MFC triplet personality measure, the GGUM-RANK model was fitted to the rank responses using the Ox program developed for this research with the specifications described in simulations. Next, the correspondence between the MFC and SS personality scores was examined via a multitrait multimethod (MTMM) analysis and by comparing CORRs with the criterion measures.

Results

Tables 5 and 6 present the MTMM and criterion validity findings. Importantly, the MFC and SS personality measures exhibited convergent validity, with monotrait-heteromethod CORRs ranging from .67 to .86. These CORRs are similar to those reported by Chernyshenko et al. (2009) for MUPP MFC and SS Big Five measures. Furthermore, the MFC and SS measures exhibited a similar pattern of CORRs with criterion measures. However, MFC criterion CORRs were somewhat lower which could be due, in part, to differences in the magnitudes of the reliabilities or reduced response consistency bias. Overall, the CORRs in Table 6 were generally consistent with previous meta-analytic findings (e.g., Barrick, Mount, & Gupta, 2003; Miller, Lynam, & Leukefeld, 2003; Steel, Schmidt, & Shultz, 2008).

Conclusion and Discussion

The main findings and practical implications of these studies are as follows. First, larger sample size led to more accurate statement parameter estimation, but the effect on person parameter estimation was small. 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., 500) are recommended for statement parameter estimation when measures are developed for high-stakes decision making. Second, regarding test length, 30 triplets may be adequate for assessment with 10-dimension MFC measures, provided that the triplets are pre-tested to ensure adequate intrablock discrimination. 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.

Table 5. Correlations of Personality Scores Based on Single-Statement and MFC Triplet Responses in Study 3.

Format	Construct	SS					MFC				
		O	C	E	A	N	O	C	E	A	N
SS	O	(.84)									
	C	.09	(.89)								
	E	.11	.29	(.91)							
	A	.13	.12	.28	(.89)						
	N	.02	-.32	-.39	-.17	(.89)					
MFC	O	.70	.14	-.01	-.07	-.08	(.67)				
	C	.03	.83	.16	.01	-.30	.14	(.84)			
	E	.05	.23	.86	.27	-.37	-.08	.17	(.82)		
	A	-.04	.17	.25	.67	-.26	.02	.17	.27	(.67)	
	N	.02	-.30	-.34	-.11	.82	-.12	-.37	-.36	-.28	(.82)

Note. Bold values indicate monotrait-heteromethod correlations; values in the parentheses are reliability coefficients. The reliabilities of the single-statement (SS; “Likert-type”) measures were computed using coefficient alpha. The reliabilities for the MFC triplet measure were calculated using the marginal reliability equation provided by Brown and Croudace (2015). MFC = multidimensional forced choice; O = openness; C = conscientiousness; E = extraversion; A = agreeableness; N = neuroticism.

Table 6. Criterion-Related Validities Based on SS and MFC Triplet Responses in Study 3.

Criterion variables	SS					MFC				
	O	C	E	A	N	O	C	E	A	N
SWLS	.01	.14	.27**	.21*	-.42**	.09	.15	.21*	.25**	-.44**
PA	.03	.10	.27**	.22*	-.39**	-.02	.01	.21*	.10	-.33**
NA	.16	.05	-.15	-.09	.52**	.05	.03	-.20*	-.09	.47**
AGG	.09	-.11	.02	-.26**	.58**	.02	-.15	-.06	-.25**	.43**
HR	.13	.10	.02	-.07*	-.12	.17*	.24**	.05	.01	-.19*
HI	.17*	.18*	-.01	-.07	-.16	.27**	.27**	-.03	.18*	-.25**
HA	.61**	.01	.24**	.17*	.13	.46**	-.01	.19*	.14	.13
HS	.04	.02	.35**	.43**	-.03	-.05	-.04	.34**	.37**	-.02
HE	.04	.14	.47**	.05	-.22*	.09	.13	.44**	.11	-.32**
HC	-.04	.17*	-.11	.14	-.05	-.09	.16*	-.15	.07	-.04

Note. All criterion variables used the SS format. SS = single statement (“Likert-type”); MFC = multidimensional forced choice; O = openness; C = conscientiousness; E = extraversion; A = agreeableness; N = neuroticism; SWLS = Life Satisfaction Scale; PA = Positive affect; NA = negative affect; AGG = aggression; HR = Holland realistic; HI = investigative; HA = artistic; HS = social; HE = enterprising; HC = conventional.
* $p < .05$. ** $p < .01$.

Researchers and practitioners are, therefore, strongly encouraged to create MFC tests comprising highly discriminating statements to ensure sufficient measurement precision. Importantly, the GGUM-RANK MCMC direct estimation process accounts for context, or potential “interactions” between statements within MFC blocks, that may influence overall triplet quality. This method will thus enable practitioners to conduct more effective MFC item analysis and facilitate parallel MFC test construction. Fourth, intrablock location variability had little to no effect on parameter recovery when generating theta CORRs were zero. In accordance with a reviewer’s suggestion, a follow-up simulation was conducted for selected conditions (30 triplets

with high α and large δ *SD*, and high α and small δ *SD*) in Study 1 with correlated thetas (0.3). The results were just slightly better in the large δ *SD* conditions, and the CORRs between generating thetas had little adverse effect on item and person parameter recovery (e.g., .42 vs. .47 RMSEs for person parameter recovery; see Online Appendix F for the detailed results). Fifth, MFC triplet measures outperformed MFC pair measures of similar length and intrablock discrimination in terms of estimation accuracy. The 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 person parameter recovery. Finally, this research not only developed GGUM-RANK estimation methods but also illustrated their viability for applied use. The empirical study provided evidence of convergent and discriminant validity for an MFC measures with real participants.

This research had some limitations. The simulation studies considered limited numbers of replications and conditions out of all the possibilities that may be seen in MFC testing applications due to the very long run-times. Development of alternative estimation methods that can reduce computation time would be a worthy topic for future research. In addition, these simulations explored parameter recovery exclusively with 10-dimension tests, but MFC tests of higher (and lower) dimensionality are used in some applied settings (e.g., Brown & Bartram, 2009; Stark et al., 2014). Thus, additional simulation research is needed to explore the accuracy of GGUM-RANK scoring with measures involving greater (and fewer) than 10 dimensions.

This research also considered just two levels of intrablock location parameter variability. Future simulations should examine a wider variety of location *SD* conditions and deliberately explore estimation with all positively or negatively worded statements within MFC blocks. This would complement the research by Brown and Maydeu-Olivares (2011) suggesting that MFC items should be created by mixing positively and negatively worded statements to ensure more accurate estimation with their Thurstonian IRT model. 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.

Moreover, this research focused on single-sample statement and person parameter estimation. To facilitate applications, research is needed to develop methods for assessing GGUM-RANK model-data fit, linking parameters across subpopulations, and concurrent calibration. This line of research would provide a foundation for GGUM-RANK differential item functioning (DIF) detection that is important for answering questions about fairness in high-stakes settings, as well as score comparisons with partially overlapping forms that may be of interest in multinational assessment contexts.

In closing, there is increasing interest in the use of MFC measures for noncognitive measurement. This research expanded on previous investigations (Hontangas et al., 2015; Seybert, 2013; Stark et al., 2005) by introducing an MCMC algorithm for estimating GGUM-RANK statement and person parameters from MFC triplet responses. It is hoped that this research provides a solid foundation for field applications and a springboard for new psychometric developments.

Author's note

Oleksandr S. Chernyshenko is currently affiliated with the University of Western Australia

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

References

- 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.
- 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.
- Brown, A., & Bartram, D. (2009). *Development and psychometric properties of OPQ32r* (Supplement to the OPQ32 technical manual). Thames Ditton, UK: SHL Group.
- Brown, A., & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 307-333). New York, NY: Routledge/Taylor & Francis Group.
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71, 460-502.
- 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., 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.
- de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2012, April). *Some extensions of the multi-dimensional 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.
- 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. D., 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*. London, England: Timberlake Consultants.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4, 26-42.
- Guenole, N., Brown, A. A., & Cooper, A. J. (2016). Forced-choice assessment of work-related maladaptive personality traits preliminary evidence from an application of Thurstonian item response modeling. *Assessment*, 7, 1-14.
- Heo, C. G. (2011). Development of Holland vocational interest inventory (short form) and investigation of Holland's hypotheses. *Journal of Korean Psychology*, 24, 695-718.
- Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, 74, 167-184.
- 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*, 39, 598-612.

- Joo, S. H., Lee, P., & Stark, S. (2017). Evaluating Anchor-Item Designs for Concurrent Calibration With the GGUM. *Applied Psychological Measurement, 41*, 83-96.
- Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*, 81-95.
- Miller, J. D., Lynam, D., & Leukefeld, C. (2003). Examining antisocial behavior through the Five-Factor Model of personality. *Aggressive Behavior, 29*, 497-514.
- 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.
- Seybert, J. (2013). *A new item response theory model for estimating person ability and item parameters for multidimensional rank order responses* (Doctoral dissertation). University of South Florida, Tampa.
- 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.
- 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., 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.
- Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin, 134*, 138-161.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics, 22*, 1701-1728.