Measurement: Interdisciplinary Research and Perspectives

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hmes20

From Likert to Forced Choice: Statement Parameter Invariance and Context Effects in Personality Assessment

Jianbin Fu, Patrick C. Kyllonen & Xuan Tan

To cite this article: Jianbin Fu, Patrick C. Kyllonen & Xuan Tan (01 Mar 2024): From Likert to Forced Choice: Statement Parameter Invariance and Context Effects in Personality Assessment, Measurement: Interdisciplinary Research and Perspectives, DOI: 10.1080/15366367.2023.2258482

To link to this article: https://doi.org/10.1080/15366367.2023.2258482

MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
https://doi.org/10.1080/15366367.2023.2258482

From Likert to Forced Choice: Statement Parameter Invariance and Context Effects in Personality Assessment

Jianbin Fu*, Patrick C. Kyllonen*, and Xuan Tan

Educational Testing Service

ABSTRACT

Users of forced-choice questionnaires (FCQs) to measure personality commonly assume statement parameter invariance across contexts – between Likert and forced-choice (FC) items and between different FC items that share a common statement. In this paper, an empirical study was designed to check these two assumptions for an FCQ assessment measuring interpersonal and intrapersonal skills. We compared parameters of common statements between two Likert forms and two FCQ forms with a block size of two (statement pairs) and among five FCQ pair forms. In three of the five FCQ forms, statements were paired only with a statement they had appeared with in a triplet block. In the other two FCQ forms, statements were paired with statements they had not been paired with in a triplet block. This design allows us to evaluate statement parameter changes due to changes in context. The results do not support the statement parameter invariance assumption between Likert and FC items or the assumption between FC items when recombining statements form new items. However, the assumption between FC items was generally held for pairs formed by dropping a statement from a triplet item. There were some suggestions for sources of context effects, but the analyses were not definitive. Implications of the findings for test practice are discussed.

KEYWORDS

Forced-choice item; Likert item; parameter invariance; context effect; item response theory

Introduction

The most popular item format for noncognitive assessments, that is, assessments of individuals' attitudes, values, beliefs, or behaviors (hereafter, personality), is the Likert item, in which a test taker selects a graded response (e.g., strongly disagree, disagree, agree, strongly agree) to a statement. However, Likert items are susceptible to various response biases and distortions such as those due to differential understanding of the rating scale, individual response styles (e.g., central or extreme tendency, acquiescence, and socially desirable responding), and faking, all of which are well documented in the literature (Friedman & Amoo, 1999; Griffith et al., 2007; King et al., 2004; Paulhus & Vazire, 2007). Forced-choice (FC) items may mitigate these biases and distortions and thus are an attractive alternative to Likert items. A forced-choice item is a block of two or more statements measuring a common trait or different traits, and test takers are asked to rank the statements within a block based on how well the statements describe their personality. Studies have shown that the FC item format is more resistant to various response biases (e.g., Cheung & Chan, 2002; Christiansen et al., 2005) and to faking (e.g., Lee & Joo, 2021; Martínez & Salgado, 2021; Wetzel et al., 2021) compared to Likert items.

CONTACT Jianbin Fu jfu@ets.org Educational Testing Service, Messick Hall, 660 Rosedale Road, Princeton, NJ 08541, USA
*Jianbin Fu and Patrick C. Kyllonen made equal contributions to this paper.

Item response theory modeling and scoring of forced-choice personality assessments

Various item response theory (IRT) models have been proposed to calibrate FC items and generate normative scores from forced-choice questionnaires (FCQs; sets of forced-choice items) (Brown, 2016). The notable models include the Thurstonian IRT (TIRT) model (Brown & Maydeu-Olivares, 2011) and various Rank models (de la Torre et al., 2012). All these models estimate probabilities of ranking sequences of statements within FC items. Their probability functions are functions of the dichotomous response function of each statement in an FC item. Rank models can be classified based on the statement response functions they employ and the block size (number of statements in an item) they apply to. The commonly used item response functions are the two-parameter logistic IRT model (2PLM; Birnbaum, 1968) and the dichotomous version of the generalized graded unfolding model (GGUM; Roberts et al., 2000). For pairs, there is the multi-unidimensional pairwise preference (MUPP) with 2PLM (MUPP-2PLM; Morillo et al., 2016) and MUPP-GGUM (Stark et al., 2005); for triplets, there are the Triplet-2PLM (Fu et al., 2023a) and Triplet-GGUM (Joo et al., 2018). The TIRT model uses the two-parameter normal ogive IRT model (McDonald, 1997) and is estimated by the limited information approach under structural equation modeling. The TIRT model applies to pairs; however, it can also be applied to items with larger block sizes by separating statements into pairs, and under structural equation modeling, it can model dependence between statements in blocks with more than two statements.

The fact that the response functions of the IRT models for FCQs build on the response function of each statement in an item has important implications for test assembly. Current standard practice is to field-test statements in Likert scales and calibrate them by 2PLM or GGUM. Then, the statement parameters based on the Likert item responses can be used to assemble FCQ forms[1] and treated as statement parameters for the above choice models to generate IRT scores from block item responses from the FCQs. The underlying assumptions are that (a) IRT statement parameters are invariant between Likert and FC format items, and (b) IRT statement parameters are invariant between FC blocks, that is, between two blocks sharing a statement. These two assumptions must be scrutinized because item format and context could affect statement responses (Lin & Brown, 2017).

Evaluating statement parameter invariance

However, these two assumptions have rarely been examined. Rather, statement parameter invariance is generally assumed (Morillo et al., 2019). Morillo et al. (2019) is the only published study we could find that investigated statement parameter invariance between Likert and FC items. As Morillo et al. (2019) admitted, their study had limitations on data, design, and estimation. They claimed that the parameter invariance assumption holds reasonably well based on their study results. However, they found that 26% to 38% of their reported intercept parameters were noninvariant (in Table 2 of their paper), a significant violation of the parameter invariance assumption.

Lin and Brown (2017) used the historical operational data of the Occupational Personality Questionnaire (OPQ) to check TIRT parameter invariance between FC items with four statements (quads) (which appear in the OPQ32i) and the corresponding FC triplets (which appear in the OPQ32r). The triplets were just the quads with one statement removed in each item (i.e., the statement that contributed least to item information; Brown & Bartram, 2009–2011). The triplets and quads were separated into statement pairs, and the TIRT model was used to estimate pairs. Therefore, their study did not directly inspect parameter invariance between FC items at the individual statement level; rather, it was at the statement pair level. However, this is the only paper we could find relevant to parameter invariance between FC items. They found that the TIRT item parameter invariance largely holds between triplets and quads. However, their study has a serious limitation. Because no common test takers took both triplets and quads, they used all common pairs as anchors to equate TIRT parameters between both item sets. Their justification was that they assume most of the TIRT item parameters are invariant between both sets of items. At the same time, this was the same assumption they were trying to evaluate. In addition, both sets of items were constructed carefully by experts, with only one statement difference in each item. This

represented minimum item contextual changes while, in practice, test developers desire to shuffle statements more freely into FC items so that a statement could appear with any number of different statements in different blocks. Nevertheless, Lin and Brown (2017) provided interesting qualitative analyses of item contexts on the outlier items in the equating (i.e., items with noninvariant parameters).

What could cause context effects (why might statement parameters be noninvariant)?

Context effects play an important role in decision theories and theories of choice and preferences (Trueblood, 2022). Hence, reviewing key concepts and findings from that literature is useful. In classic decision theories from behavioral sciences, it is assumed that choices between two or more statements[2] are based on latent preferences or utilities associated with each statement. Luce's (1959) choice axiom states that the probability of selecting a statement is based on the utility for that statement over the sum of the utilities of all statements in the choice set, and the utility of a statement is independent of the utilities of other statements. Context effects are violations of Luce's (1959) choice axiom, in which the utility of a statement depends on the utilities of other statements; thus, the probability of selecting a statement is affected by what statements appear with it. Choosing a statement with respect to its self-descriptiveness differs from expressing a preference among different consumption options or other decision choices, but context effects are general to decision-making (Trueblood et al., 2013). Considering the three prominent context effects – attraction, compromise, and similarity – may be useful to anticipate where we might find parameter invariance violations.

In a choice between X and Y, where two salient dimensions govern the choice (e.g., X is higher quality and Y is cheaper), context effects occur when introducing a third option affects the choice. An *attraction* effect occurs when the third option is inferior to X in quality and cost, making X more attractive as a choice over Y. A *compromise* effect occurs when the third option is much higher than X on quality and cost, and then X becomes the compromise choice between the third option and Y. A *similarity* effect occurs when the third option is similar to option Y, and then X becomes more preferred.

Translated to the choice task in forced-choice personality assessment, Luce's choice axiom provides the basis for assuming statement parameter invariance across all contexts (Likert to forced choice, subset pairs from triplets, and recombined pairs), but context effects could occur in several ways. An *attraction* effect could occur in the change from Likert ratings to pairs, for example, if the paired statement (Y) were worse in some way (e.g., social desirability) than the Likert statement (X), making X more attractive. A *similarity* effect could occur in the change from pairs (X, Y) to triplets (X, Y, Z) if the third option (Z) was more similar (e.g., in the words used in the item) to Y than to X, making X more desirable. A compromise effect could occur if the third option were superior to both X and Y in such a way that X became the middle, compromise item. The statements used in this study vary on many dimensions – the intended personality construct being measured, other personality constructs, specific wording, and social desirability – and context effects (attraction, similarity, and compromise) could operate through any of them.

Lin and Brown (2017) conducted a qualitative analysis to identify themes associated with item parameter changes. They listed only a few examples of their findings, but at least one appears to be based on an item wording similarity effect. Adding a statement pertaining to "conversation" (Z) to a triplets block that already included a statement using the word "talking" (Y) led to lower discrimination of the "talking" (Y) statement (i.e., a lower factor loading on the dimension that the "talking" statement intended to measure), a similarity effect (the fact that this effect occurs through the lowering of the discrimination parameter for the Y statement means that the effect was limited to the upper part of the trait distribution). A similarity effect might also have been observed on another item in which a statement with high social desirability ("I finish things on time") was added to a triplets block that already contained a statement with high social desirability ("I consider what motivates people"), that is, there were now two statements that were similar in social desirability, leading to a reduction in the intercept parameter of the repeated statement (i.e., it was harder to endorse the statement in the quad than in the triplet). These examples are suggestive. Similar context effects might be revealed in the present study through exploratory analyses which we conducted.

Practical implications of statement parameter noninvariance

If the statement parameter invariance between Likert and FC items does not hold, then it is not appropriate to use calibrations from a field test of Likert items for FC item scoring (the calibrations might still be adequate for use in block assembly to create FC items). Instead, all proposed FC items would have to be field-tested in FCQs. Then, items would be selected into operational forms based on item analysis. Under the traditional use of the MUPP scoring method (Drasgow et al., 2012; Stark et al., 2005), statement parameters from a field test using Likert items are treated as fixed for the IRT model to generate test scores from an FCQ administration. If the statement parameter invariance between FC items is not supported, then statement parameters are item-dependent. In such a case, each item must be field-tested in FCQs to obtain its item parameters. This requirement would limit the item pool (allowing only intact blocks to be used) and significantly increase test costs. This approach would also make computer adaptive tests (CAT) infeasible because CATs require a large item pool. One way to do this would be with a statement pool that can be assembled into items on the fly, an approach adopted by Drasgow et al. (2012). Therefore, the two assumptions mentioned above (i.e., statement parameter invariance assumptions between Likert and FC items and between two FC items) are vital for efficient test assemblies and administrations, which may be why some test developers are willing to make the two assumptions despite little evidence to support them.

In the current study, we designed and administered specific FCQ forms, with pairs measuring interpersonal and intrapersonal skills to examine the two assumptions. In this study, we investigate the context effects of pairs (blocks with two statements) under two scenarios – subset pairs derived from triples (blocks with three statements) and reconstructed pairs. The issue to be evaluated is statement parameter invariance across item types (Likert vs. forced choice) and choice contexts (subset pairs from triplets and recombined pairs). We used 2PLM and MUPP-2PLM to calibrate Likert and FC items, respectively. We fitted unrestricted models where the statement parameters were noninvariant across items. Then, we used Wald tests (Fahrmeir et al., 2013, p. 663) to check whether a statement on different items had equal parameters.

The rest of this paper is organized as follows. First, in the Method section, we describe the research design, data collection, IRT models, Wald tests, and estimation program used for checking the two assumptions. Then, we present the results. Finally, we discuss the results, implications for practice, and the current study's limitations.

Method

In this section, we first present the form assemblies based on the study design and the form administrations, followed by the general strategy of data analyses. We then go into the details of the data analyses: the IRT models used to calibrate the Likert and FCQ forms, the Wald tests for parameter invariance, the correlations of latent trait scores, and the computer program used in the analyses.

Study design and form assemblies

This study was based on an assessment measuring interpersonal and intrapersonal skills essential to higher education and career success. First, 600 statements were developed measuring 15 traits such as perseverance, leadership, creativity, curiosity, responsibility, and self-discipline (identification and interpretation of these dimensions are described in a separate report). Forty statements were developed for each trait. The process by which these statements were assembled into the various forms for this study and the logic by which these forms were compared to evaluate context effects is shown in Figure 1 and described below.

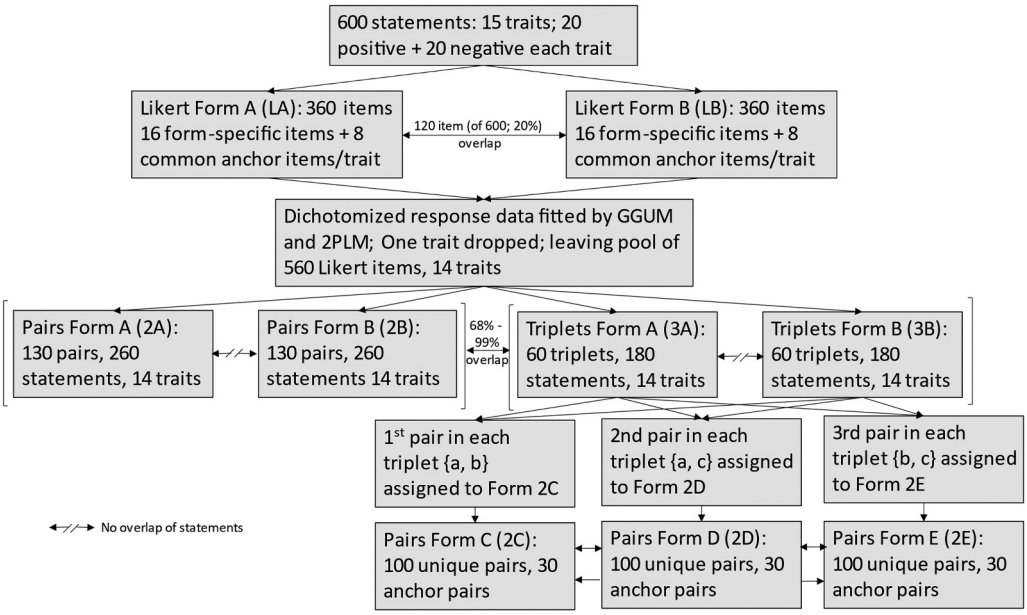

Figure 1. Steps taken to create the forms for the study design.

Creating initial Likert forms (LA and LB)

The 600 statements were randomly based on traits separated into two Likert forms (LA and LB); each statement was paired with a four-category rating scale (i.e., strongly disagree, disagree, agree, and strongly agree). In each form, each trait was assessed with 24 statements, 8 of which served as anchors between two forms – thus, eight anchor statements and 16 form-specific statements per trait per form. The two Likert forms were administered with randomly ordered statements between participants to mitigate order effects.[3] The rating scores were dichotomized (i.e., coded strongly disagree and disagree to 0 and others to 1) and fitted by the 2PLM.[4] (One trait with 40 statements was dropped from further administrations due to bad statement statistics).

Creating initial forced-choice forms (pairs forms 2A and 2B; triplets forms 3A and 3B)

From the remaining 560 statements measuring 14 traits, experts assembled two parallel FCQ forms with pairs (pair Forms 2A and 2B) and two parallel FCQ forms with triplets (triplets Forms 3A and 3B). Forms were assembled by an automatic assembly program following content requirements, item statistics, and best practices in measurement (form and block assembly is described in a separate report). The automatic assembly program is a linear programming solver that maximizes the overall test information of each form under some constraints, for example, the tolerance of the differences of the item location parameters and social desirability ratings among the statements in an item, no enemy statements in an item, the tolerance of the differences of the overall trait information across traits within a form and across forms for each trait, and so on. Each pair form comprised 130 items (260 statements), and each triplet form comprised 60 items (180 statements). Each item in each form comprised statements measuring different traits. All items within a form were multidimensional forced-choice items. Within the forms, statements only appeared once. Statements were not repeated across forms, except that statements appearing in a pairs form could appear once in a triplet form: 355 statements appeared once in a pairs form and once in a triplets form; 99% of triplet statements appeared in pair forms; 68% of pair statements appeared in triplet forms (the asymmetry is due to pairs forms using more statements [520] than triplet forms [360]).

Context change between subset pairs from triplets (forms 2C vs. 2D, 2C vs. 2E, and 2D vs. 2E)

We constructed a set of pair forms (Forms 2C, 2D, and 2E) from the two triplet forms (3A and 3B) by a simple rule. First, for 100 out of the total 120 triplet items, each triplet item with statements a, b, c was separated into three pair items, {a, b}, {a, c}, and {b, c}, which were assigned to Forms 2C, 2D, and 2E, respectively. Then, the remaining 20 triplets were used to create 30 pairs, which appeared in all three forms and served as anchor items (Figure 1 provides details). The anchor set represented a mini set of all test forms with respect to test content specifications. Thus, each of the newly created pair forms also had 130 items, and there were no repeated statements within each form.

Context change between new pairings (2A, 2B vs. 2C, 2D, 2E)

There were 355 common statements in Forms 2A and 2B that also appeared in Forms 2C, 2D, and 2E. Most were paired with different statements – only 17 items were unchanged between the two form groups. The scale linking between the two form groups was done by administering two forms to each respondent (2A and 2C or 2B and 2D). Forms administration order was randomized.

Context change between likert and forced-choice items (LA, LB vs. 2A, 2B)

Statement parameter invariance between Likert and FCQ items was examined by using the Likert data (Forms LA and LB) and the FCQ data on Forms 2A and 2B.[5] There were anchor items between the two Likert forms; however, there were neither common items nor common respondents between Likert and FCQ items and between the two FCQ forms. Rather, the scale linking was based on the assumption of random equivalence among the three groups of participants who took either of the Likert forms, Form 2A or Form 2B.

Form administrations and participants

A group of $N = 1005$ participants was administered one of the two Likert-scale questionnaires on the crowdsourcing website Amazon Mechanical Turk (AMT). A second group of $N = 2409$ participants (after data cleaning) received one of several forced-choice forms or combinations of two forms from the online survey site, Prolific. All FC items with a response time shorter than 5 seconds for pairs and 10 seconds for triplets were set to missing, and any participants with more than 20% missing items in a form

Table 1. Demographic distributions of test data (in percentages).

	Likert form (%)			Forced-choice form (%)					
Value	LA ($N = 505$)	LB ($N = 500$)	Total ($N = 1005$)	2A2C ($N = 489$)	2B2D ($N = 492$)	2C ($N = 478$)	2D ($N = 478$)	2E ($N = 472$)	Total ($N = 2409$)
Ethnicity									
White	76	71	73	75	70	75	67	69	71
Black	6	7	6	10	9	7	8	9	9
Asian	9	9	9	4	6	5	8	10	7
Hispanic	6	8	7	7	12	8	10	8	9
Multirace	4	4	4	4	3	4	5	4	4
Other	0	1	0	1	0	1	2	1	1
Education									
High-school diploma	1	2	1	31	35	27	34	31	32
Associate's degree	17	17	17	16	17	16	14	14	15
Bachelor's degree	56	60	58	38	34	42	36	39	38
Postgraduate degree	26	21	23	15	13	14	15	17	15
Other	0	0	0	0	0	0	1	0	0
English proficiency									
Professional working proficiency	1	1	1	1	1	1	1	1	1
Full professional proficiency	4	4	4	3	3	3	4	5	3
Native/multilingual proficiency	96	95	95	96	95	96	95	93	95
Other	0	0	0	0	1	0	0	0	0

Table 2. Fixing first statement's intercept for model identification: concurrent calibration of pair forms 2A to 2E.

Form	Common statements	Unique statements
2A	2C or 2D intercept	Likert intercept
2B	2C or 2D intercept	Likert intercept
2C	Likert intercept	Likert intercept
2D	2C intercept	
2E	2C intercept	

A first statement' intercept was fixed to either the precalibrated Likert intercept estimate or the value of the common statement from another FCQ form, as indicated in the table. Common statements are ones that appear in more than one form (e.g., 2A and 2C); A small percentage of unique statements appear only in one form (e.g., 2A only).

were removed from the form. Table 1 lists the distributions of ethnicity, highest academic degrees, and English proficiency levels on each dataset collected. The data from all forms had similar distributions on the three demographic variables, except that the test takers on the two Likert forms (from the AMT website) had higher educational attainment than those who took the forced-choice forms (from Prolific).

General analysis strategy

The two pair forms 2A and 2B were initially assembled based on data from the Likert statements in LA and LB. Forms 2C, 2D, and 2E were assembled from triplet items (see Figure 1). All the data from the five forms (2A to 2E) were then combined and concurrently calibrated by the MUPP-2PLM to investigate statement parameter invariance between items. Forms 2A and 2C were linked by common test takers; Forms 2B and 2D were also linked by common test takers, and Forms 2C, 2D, and 2E were linked by anchor items; thus, all forms were linked to a common scale by a concurrent calibration.[6] The combined Likert and Forms 2A and 2B data were calibrated together with Likert items fitted by the 2PLM and pair forms fitted by the MUPP-2PLM. The linking was conducted based on the assumption of randomly equivalent groups. All item parameters, including those of common statements across forms, were freely estimated except for the constraints needed for model identification. Then, Wald tests (Fahrmeir et al., 2013, p. 663) were applied to check the statistical significance of parameter equality. This process is explained in more detail in the following sections.

IRT models

IRT models for Likert items

The precalibrations of the Likert forms (Forms LA and LB) were carried out one trait at a time by the 2PLM. Intertrait correlations were not considered in the precalibrations. For each statement k, the probability of endorsement based on 2PLM is

$$P_k = \frac{\exp(a_k\theta_k + b_k)}{1 + \exp(a_k\theta_k + b_k)}, \tag{1}$$

and the probability of rejection is

$$Q_k = 1 - P_k(1|\theta_k), \tag{2}$$

where θ_k is a score of the latent trait measured by statement k, a_k is the discrimination parameter of θ_k, and b_k is the intercept.

For the concurrent calibration of the Likert and FCQ forms (i.e., Forms LA, LB, 2A, and 2B), the dichotomized Likert statement data were fitted by the compensatory two-parameter logistic multidimensional IRT model (M2PLM; Moustaki, 2000; von Davier, 2008). Because each statement only

measured one trait (i.e., the Likert scale had a simple structure), the Likert statements were calibrated by the 2PLM with the intertrait correlations considered.

IRT model for forced-choice items

The FC pairs in Forms 2A to 2E were calibrated by the MUPP-2PLM. In MUPP-2PLM, the probability of selecting the first statement in a pair is represented by

$$P^* = \frac{P_{1,2}(1,0)}{P_{1,2}(1,0) + P_{1,2}(0,1)}, \tag{3}$$

where $P_{1,2}(1,0)$ denotes the joint distribution of endorsing the first statement and rejecting the second statement, while $P_{1,2}(0,1)$ denotes the opposite. Assuming test takers respond to the two statements independently, then

$$P_{1,2}(1,0) = P_1 Q_2, \tag{4}$$

$$P_{1,2}(0,1) = Q_1 P_2, \tag{5}$$

where P_1 and P_2 follow Equation 1, and Q_1 and Q_2 follow Equation 2. Then, by substituting with Equations 1, 2, 4, and 5, Equation 3 can be simplified as

$$P^* = [1 + \exp(a_2\theta_2 + b_2 - a_1\theta_1 - b_1)]^{-1}. \tag{6}$$

The probability of selecting the second statement is

$$Q^* = 1 - P^*. \tag{7}$$

In Equation 6, for b_1 and b_2, only one of them can be estimated. We may estimate the difference between the two intercepts, $b_{12} = b_1 - b_2$. Then, Equation 6 becomes

$$P^* = [1 + \exp(a_2\theta_2 - a_1\theta_1 - b_{12})]^{-1}, \tag{8}$$

which is just a special case of M2PLM. However, to check the state parameter invariance, both intercepts need to be kept in the model. Thus, we can fix b_1 instead to identify the model.

In the current study, for the concurrent calibration of Pairs Forms 2A to 2E, we fixed the intercept parameters of the first statements in all items in Form 2C and the first statements in Forms 2A and 2B that did not appear in Forms 2C, 2D, or 2E to their precalibrated 2PLM intercept parameters in the Likert scales (Forms LA and LB; see Table 2). For the other first statements in Forms 2A and 2B, we fixed their intercept parameters to the intercepts of their common statements in Forms 2C and 2D. For the first statements in Forms 2D and 2E, we fixed their intercept parameters to the intercepts of their common statements in Form 2C.

For the current calibration of the Likert scales and Forms 2A and 2B, the intercepts of the first statements in Forms 2A and 2B were fixed to their intercepts in the Likert scales (Forms LA and LB; see Table 3). In all calibrations, all θ_k are assumed to follow a standard multivariate normal distribution to identify the models. The intertrait correlation matrix in the standard multivariate normal

Table 3. Fixing first statement's intercept for model identification: concurrent calibration of Likert and pairs forms LA, LB, 2A, and 2B.

Form	Intercept
LA	Free
LB	Free
2A	Likert intercept
2B	Likert intercept

distribution was fixed during the calibrations to stabilize the estimations. The fixed correlation matrix came from the estimations of trait correlations based on the precalibration of Likert Forms LA and LB.

Wald test to evaluate statement parameter invariance

In the free models, all statement parameters, including those of the common statements, were free to estimate except for those necessary restrictions to identify the models (i.e., those presented in Tables 2 and 3). Once maximum likelihood estimates of the statement parameters and their asymptotic estimates of the covariance matrix in a model were obtained, a Wald test was performed to check if the discrimination or intercept parameter estimate of a common statement in two different FC items or item formats (Likert vs. FC) was statistically different. In the present study, we used the Wald test under two scenarios. First, we examined the intercept parameters of the second statements in Forms 2A and 2B that were used as the first statements in Form 2C. Since the intercepts of the first statements in Form 2C were fixed to their precalibrated 2PLM intercept parameters in the Likert scales, the null hypothesis of the Wald test for the intercept parameter of a second statement was that the intercept parameter estimate $\hat{b}_k$ was equal to a fixed value o. The Wald statistic is

$$W = (\hat{b}_k - o)^2 / \mathrm{var}(\hat{b}_k). \tag{9}$$

In the second set of Wald tests, used for all the other statement parameters, the null hypothesis was that the discrimination or intercept parameter estimates of a statement (denoted as $\hat{r}$ and $\hat{r}^*$) appearing in two different FC items or item formats were equal, i.e., $\hat{r} - \hat{r}^* = 0$. The second Wald statistic is

$$W = (\hat{r} - \hat{r}^*)^2 \left\{ [1, -1] \Psi_{\hat{r},\hat{r}^*} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\}^{-1}, \tag{10}$$

where $\Psi_{\hat{r},\hat{r}^*}$ is the asymptotic estimate of the covariance matrix of $\hat{r}$ and $\hat{r}^*$. Both Wald statistics follow an asymptotic chi-square distribution with one degree of freedom. If the p-value of a Wald test was not smaller than a threshold, then the test was not statistically significant. We conclude that the two parameter estimates are equal or the intercept estimate is equal to the fixed value. For the current calibration with the five FCQ forms, there were 1312 Wald tests, and for the concurrent calibration with the Likert forms and the two FCQ forms, 776 Wald tests were carried out. With a significance level of .05, the Bonferroni correction was applied to these tests, giving threshold values, $\alpha = .05/1312 = 3.81e-5$ and $\alpha = .05/776 = 6.44e-5$, respectively.

Computing trait scores using maximum a posteriori (MAP) estimates

Because the ultimate goal of this assessment is to provide accurate latent score estimates, we also fitted restricted models where the parameters of a common statement in different FC items and item formats were constrained to be equal. Based on the free and restricted models, we estimated the MAP (Maximum A Posteriori; Baker & Kim, 2004) trait scores on each FCQ form. The MAP estimate of the latent trait score vector for a test taker i, $\hat{\boldsymbol{\theta}}_i$, maximizes the following log likelihood function given item parameter estimates,

$$l_i = \sum_{j=1}^{J} \log \left[x_{ij} P^*_{ij} + (1 - x_{ij}) Q^*_{ij} \right] + \log \phi(\boldsymbol{\theta}_i | \Sigma), \tag{11}$$

where x_{ij} is test taker's binary response to FC item j in an FCQ form (total j items) with score 1 indicating the selection of the first statement in item j and score 0 the second statement; P^*_{ij} (Equation 6) and Q^*_{ij} (Equation 7) are test taker's probabilities of selecting the first and second statements in item j, respectively, and $\phi(\boldsymbol{\theta}_i|\Sigma)$ is the standard multivariate normal density of test

taker's trait score vector $\boldsymbol{\theta}_i$ with the intertrait correlation matrix Σ. Then, we calculate the correlations of the latent score estimates between free and restricted models to see if the latent score estimates shifted for each FCQ form.

Estimation program

The M2PLM model in the form of Equation 8 is implemented in many computer programs, including the *mirt* package (Chalmers, 2012) in the R program (R Core Team, 2022). However, to study the statement parameter invariance, we needed to use the M2PLM's item response function of Equation 6. This function has not been implemented in any programs that we could find. Fortunately, *mirt* provides a function (*CreatItem*) that allows users to define their own item response functions. Then, all the estimation and analysis modules in *mirt* can be used to estimate and further analyze the models. The functionality of restricting model parameters in *mirt* was convenient in setting up our models. The model estimation method was the maximum marginal likelihood estimation with an expectation-maximization algorithm (MML-EM; Bock & Aitkin, 1981; Fu, 2019). As part of the MML-EM, the joint latent trait distribution needed to be approximated. Since our models had 14 dimensions, it was quite challenging to make the approximation. We adopt the Metropolis-Hastings Robbins-Monro (MHRM) algorithm (Cai, 2010), which has been shown to perform better than other methods (Fu et al., 2023a; Garnier-Villarreal et al., 2021), such as the quasi-Monte Carlo integration method (Morokoff & Caflisch, 1995). The MAP latent trait scores were estimated by the *fscores* function, and the Wald tests were carried out by the *wald* function in *mirt*. All default settings in the *mirt* functions were kept. All estimations and analyses were conducted in the *mirt* 1.37 package and the 64-bit R 4.2.1.

Results

Statement parameter invariance between likert and FC items

As described previously, the statement parameter invariance assumption between Likert and FC items was checked based on the free model of the concurrent calibration of the Likert scales and FCQ Forms 2A and 2B. Table 4 shows the percentages of noninvariant discrimination parameters were 33% and 30% for Forms 2A and 2B, respectively, and for intercept parameters, they were 21% and 24%, respectively. We calculated the (Pearson) correlations of latent trait score estimates between the free and restricted models for the 14 traits in Forms 2A and 2B and plotted them in Figure 2. For both forms, most correlations were below .95 and could be as low as .78 (Trait 10) and .87 (Trait 3), respectively. Therefore, the statement parameter invariance assumption between Likert and FC items does not appear true for the Likert and FCQ forms under study.

Table 4. Statement parameter invariance between likert and FC items: Wald test result.

Form	Parameter	Total number of Tested Parameters	Number of noninvariant parameters	Percentage noninvariant parameters (%)
2A	Discrimination	256[a]	84	33
	Intercept	130	27	21
2B	Discrimination	260	78	30
	Intercept	130	31	24
Total	Discrimination	516	162	31
	Intercept	260	58	22

[a]The standard error estimates of four discrimination parameters were negative; thus, their Wald tests were not conducted.

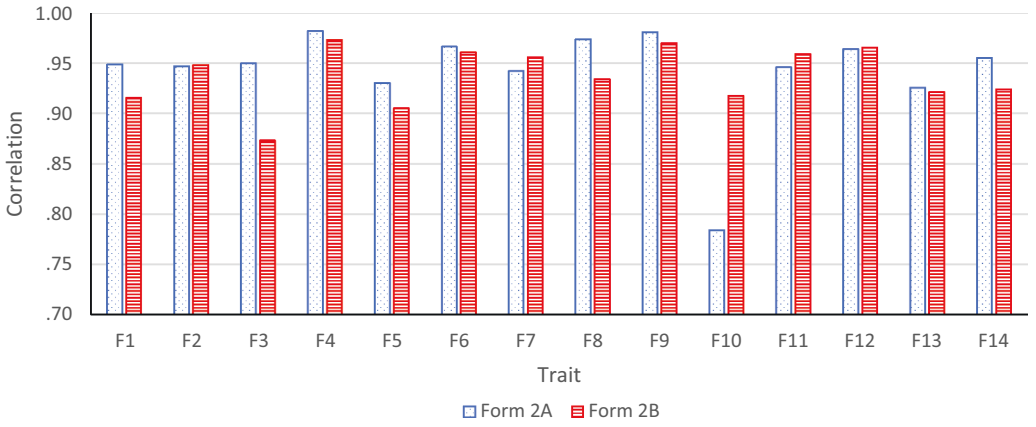

Figure 2. Statement parameter invariance between likert and FC items: FCQ forms 2A's and 2B's correlations of trait score estimates between free and restricted models.

Table 5. Statement parameter invariance between FC items: Wald test result.

Context (Forms)	Parameter	Total number of tested parameters	Number of noninvariant parameters	Percentage noninvariant parameters (%)
Recombination pairs (2A, 2B)	Discrimination	652	37	6
	Intercept	260	87	33
Triplet subset pairs (2C, 2D, 2E)	Discrimination	300	24	8
	Intercept	100	2	2
Total	Discrimination	952	61	6
	Intercept	360	89	25

Statement parameter invariance between FC items

Based on the free model for the concurrent calibration of the five FCQ forms, Wald tests were conducted on all parameters of a common statement appearing in different items. The exception was the intercept parameter of a common statement used as the first statement in both items because both intercepts were fixed to the same value for model identification. In addition, the intercept of the first statement in an item in Form 2D or 2E was constrained to be the same as that of the statement in Form 2C (as shown in Table 2), and the 60 statements in the 30 anchor items were the same across Forms 2C, 2D, and 2E. Thus, the Wald tests associated with these common item parameters were conducted only once in Form 2C. Table 5 shows the total number of tested parameters and the number and percentage of noninvariant parameters by discrimination and intercept for two form groups, (2A, 2B) and (2C, 2D, 2E). Of the tested discrimination parameters in the two form groups, 6% and 8%, respectively, were found to be noninvariant. In contrast, the form group (2A, 2B) had a much higher percentage of noninvariant intercepts than the form group (C, D, E) (33% vs. 2%). In Forms 2A and 2B, 18 second statements were used as the first statements in Form 2C. Since the intercepts of the first statements in Form 2C were fixed to those from the precalibrated Likert forms for model identification, the intercepts of the 18 second statements in Forms 2A and 2B were tested against the fixed values. The fixed values seemed arbitrary; however, these were the minimum assumptions we needed to make to identify the model. Among the 18 intercepts, only seven showed noninvariance. Removing these seven noninvariant intercepts, we still had 31% of the tested intercepts to be noninvariant in Forms 2A and 2B.

There were 17 common FC items between the two form groups, (2A, 2B) and (2C, 2D, 2E). In each common item, the two statements were the same across two forms. For the 34 statements in the 17

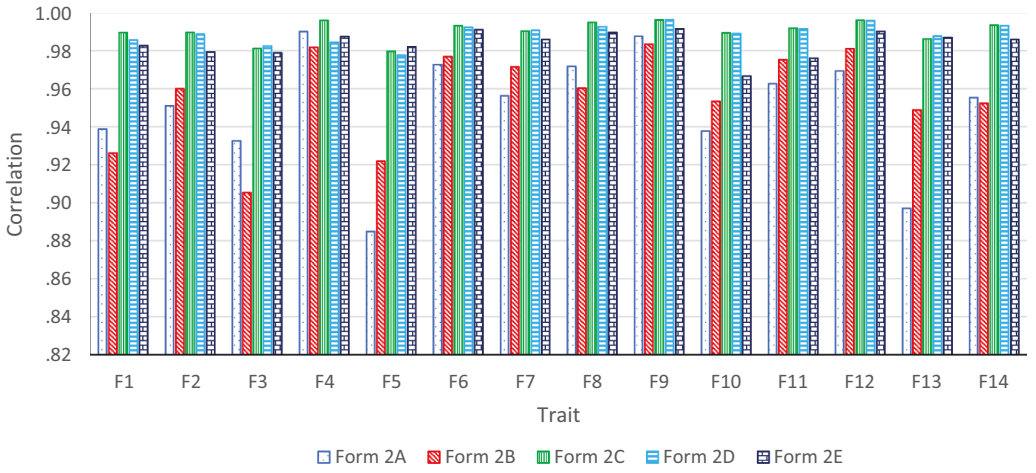

Figure 3. Statement parameter invariance between FC items: comparisons of correlations of trait score estimates between free and restricted models among five FCQ pairs forms.

common items, only one discrimination parameter changed statistically significantly between the two forms. Thus, it appears that statement choices in an item were not impacted by other items in a test form.

We next evaluate the impact of the statement parameter invariance assumption on latent score estimates. Figure 3 compares the (Pearson) correlations of the latent score estimates between the free and restricted models among the five FCQ forms for each trait. For Forms 2C, 2D, and 2E, all correlations were larger than .97, while for Forms 2A and 2B, many were smaller than .95 and could be as low as .88 (Trait 5) and .91 (Trait 3), respectively.

Therefore, the statement parameter invariance assumption appeared to hold within the form group (2C, 2D, 2E) in which context changes were subset pairs from triplets, but not between the two form groups (2A, 2B vs. 2C, 2D) in which context changes were recombined pairs. In the Discussion section, we explore some possible reasons for statement parameter drifts in the two contexts.

Analysis of items affected by context (in which statement invariance did not hold)

Likert vs. Forced-choice context effects

Forms 2A and 2B each comprised 130 pairs (260 statements; 520 statements altogether) that had also been administered as single Likert items. Of the 256 (Form 2A) and 260 (Form 2B) statements, when put in pairs, 31% showed context effects for discrimination, and 22% showed context effects for the intercept. There was a main effect for change in intercept from Likert ($M = .49, .47$) to forced choice ($M = .38, .41$) for Forms 2A and 2B, respectively. There was also a mean decrease in discrimination from Likert ($M = 1.49, 1.52$) to Forced Choice ($M = .87, .93$) for Forms 2A and 2B, respectively.

Attraction effect on intercepts. An attraction effect might be expected to increase the intercept of a target Likert statement, that is, increase target statement endorsement, if that target were paired with a lower-social-desirability decoy (Lin & Brown, 2017, found such an attraction effect). We found that for Form 2B, a decoy with lower social desirability was associated with a higher intercept (endorsement) of the target, $r\ (128) = -.18$, $p = .040$; we did not find this for the other form (2A), $r\ (128) = -11$, $p = .222$. When we evaluated only items flagged for showing parameter invariance violations, we found the direction of intercept differences to be

consistent with an attraction effect, but there were no significant differences, $r\ (25) = -.28$, $p = .164$, $r\ (29) = -.25$, $p = .181$, for Forms 2A and 2B, respectively.

Item length effects on discrimination. Reduction in item discrimination is sometimes attributed to factors that cause confusion (e.g., poor instructions, imprecision in wording) or comprehension difficulties (long, complex statements), as detailed in various item writing guideline documents (Haladyna & Downing, 1989; Osterlind, 1998). Statement length could be such a factor, and thus, we hypothesized that statement length would be correlated with a reduction in absolute item discrimination but found no relationship, $r = .05$, $r = .04$, for Forms 2A and 2B, respectively.

Subset pairs from a triplet

Since there were very few instances of statements flagged for invariance violation due to context changes from triplets to pairs, no analyses of statements identified under this condition were conducted.

Pair recombinations

We analyzed this to determine whether particular dimensions were more likely to be associated with context effects. We found statistically significant context effects (indicated by the numbers of non-invariant discrimination and intercept parameters) due to dimension ($\chi^2\ (14) = 44.3$, $p < .005$) and Big Five composite ($\chi^2\ (5) = 33.3$, $p < .005$). Big Five composites group the 14 dimensions into five composites: Conscientiousness, Positive Emotionality, Extraversion, Agreeableness, and Openness. We also note that Openness dimensions are relatively more likely to be subject to context effects (violations of parameter invariance) and that items from Conscientiousness dimensions are relatively less likely to experience context effects.

Discussion

This study was designed to evaluate evidence for several context effects that could occur on forced-choice personality assessments. One is the change in context from a Likert rating scale measure to a ranking (forced-choice) measure. In principle, ratings and rankings are simply alternative approaches to achieving the same outcome – the identification of the location of an individual's trait level on a scale. However, they do so through different means, a set of ratings on a Likert scale standard or a set of statement comparisons. A second is a change in context due to dropping an option or, as implemented here, deleting a statement from a triplet to create a statement pair. A third was a change in context due to recombining pairs. Conventional choice models in behavioral sciences generally assume "independence of irrelevant alternatives," which in personality assessment means that statement agreement should not depend on the other statements in the choice set. In item response theory, this means that statement parameters, such as the slope and intercept parameters from the 2PLM, are assumed to be invariant across contexts. Context effects refer to a violation of these assumptions.

This study found evidence for context effects due to item type (Likert vs. forced-choice responding) and within forced-choice items due to recombining pairs. We found only limited context effects in cases of subset pairs from triplets. The evidence for context effects was based on statistical tests of the invariance of statement parameters (slope and intercept of the item response function) from an IRT model (2PLM) when those statements appeared in different contexts. For the change from Likert to forced choice, 31% of the statements showed a change in their discrimination parameters; 22% showed a change in their intercept parameters. For the subset pairs from triplets, we saw changes in only 8% (discrimination) and 2% (intercepts) of the statements. For the recombined pairs, changes were found in 6% (discrimination) and 33% (intercepts) of the statements. Thus, although there is substantial parameter invariance, there is also a large amount of parameter change due to context.

Thus, our study does not support the assumption of statement parameter invariance between all Likert and FC items because many statement parameters vary, and latent trait score estimates drift between the two item formats. Item format appears to impact test takers' responses to statements substantially. This is consistent with the finding in Fu et al. (2023b). Fu et al. (2023b) compared the MUPP-2PLM on the real test data from the FCQ forms with pairs with fixed statement parameters from the Likert scales and directly estimated statement parameters from the FC data. The same comparison was also made on the Triplet-2PLM on the real test data with triplets. In all cases, the models with directly estimated statement parameters performed much better than those with fixed statement parameters in terms of the criteria for model comparison, model fit, and item fit. The correlations of latent trait score estimates between the fixed and estimated models were only between .8 and .9 on average.

In FCQ forms 2C, 2D, and 2E, the study found that the statement parameter invariance between FC items generally held, as indicated by a small portion of drifted statement parameters and, more importantly, nearly perfect correlations (>.97) of latent trait score estimates between free and restricted models. In these forms, statements are paired in a restricted way. Specifically, the statement pool was deliberately separated into groups of three statements (i.e., triplets) by content experts following well-established test assembly guidelines. Then, the statement pairing was done within each triplet. On the other hand, the items between the form groups (A, B) and (C, D, E) represent statement recombinations. For these items, the statement parameter invariance assumption was frequently violated.

Following Lin and Brown's (2017) qualitative findings on potential themes underlying intercept (relative social desirability changes) and discrimination (changes in content similarity) drifts, we conducted various analyses to determine what might lead to context effects (parameter invariance violations). Drawing on the choice context literature, we found some but limited evidence for an attraction effect such that a decoy statement's low social desirability was associated with an increase in a target statement's intercept. However, we failed to find evidence for an effect on discrimination that we might have expected due to combining lengthy, multiclause statements. Such statements are often associated with greater confusion and lower discrimination (Alreck & Settle, 1994; Fowler, 2006; Tourangeau et al., 2000). We did find evidence that certain personality dimensions (e.g., openness) were more likely to be associated with context effects than others (e.g., conscientiousness). Nevertheless, our analyses of context effects generally yielded no clear evidence for other factors that might have contributed to context effects. It is clear from our brief exploratory analyses that our findings were merely suggestive, and further, more systematic research is needed to understand what item factors might contribute to context effects.

Our results show that in the 17 common FC items across forms, only one discrimination parameter was noninvariant, which generally suggests the local independence of statement choices at the FC item level. However, because the local independence of FC item responses is not the focus of this study, further elaborative research is needed to address this issue.

Limitations

The current study used convenience samples recruited from online survey sites. The samples may differ from the population the FCQ forms are intended for applicants for colleges and graduate and professional schools. In addition, the samples given the Likert forms differed from those given the FCQ forms on educational levels. These sample differences could affect the results of this study, especially the comparison between Likert and FC items, because the scale linking in this comparison assumed equivalent respondent groups. It will be prudent to conduct this study again once more samples from operational administrations of the FCQ forms are available and, if possible, administer both Likert and FCQ forms to a group of participants.

Another limitation concerns how pairs are formed. In the current study, the three subset pairs from a triplet generally showed statement parameter invariance between pairs. The triplets were constructed through an automated assembly process that constrained blocks so that statements within blocks were

similar in item location and social desirability and varied in dimension. Large blocks could alternatively have been used. For example, the statement pool can be divided into blocks of 10 or 20 statements, and within each block, all statements could be paired together. Large blocks of statements mean less restriction on statement pairing, which can provide more meaningful results for operational testing practice than small blocks.

The findings in this study pertain to the statements and their measured constructs used in the assessment. Caution should be exercised in generalizing the findings to other FCQ assessments. It would be prudent to conduct a similar study on a different assessment to verify whether the parameter invariance assumptions hold for that assessment.

Implications for practice

Statement parameter invariance between Likert and FC items and between FC items are two important assumptions to lower the cost of developing and administering FCQs, especially for computer adaptive assessments. However, these two assumptions are not self-evident; an appropriate study should be conducted for each assessment to verify the soundness of these two assumptions rather than taking them for granted.

FCQ assembly guidelines based on best practices limit the statements with which a statement can be paired. However, such a practice might still fail to ensure statement parameter invariance between FC items. More stringent rules may be needed. For example, a statement pool may be separated into large groups, with statement pairing permitted only within a group to achieve parameter invariance in general. Further research is needed to determine if this is possible.

Suppose statement parameter invariance between Likert and FC items is difficult to support for an assessment. In that case, all statement parameters used for scoring should come from the calibration of FCQ data. Depending on the purpose of an assessment, if test scale invariance across test administrations is not critical, then initially, statement parameters from Likert scales may be used. Later, statement parameters can be periodically updated when more FCQ data are available.

If it is deemed that the statement parameter invariance between FC items does not hold for an assessment, then item-dependent statement parameter estimates from FCQ data should be used for scoring. For this case, using linear fixed forms may be the best choice. If so desired and feasible, computer adaptive testing can be carried out at the item level (rather than the statement level).

Notes

1. For example, blocks can be assembled so that within-block statements match on social desirability, dimension information is maximized and forms match on expected information by dimension.
2. We use the term statements here because that is the entity on which we conduct analyses; the wider choice literature typically refers to items, options, objects, or simply choices. We avoid items here because, for our context, an item in an FCQ refers to a block of statements.
3. Items were randomly ordered separately for each participant for all forms.
4. These data were also fit by the GGUM; generally, the 2PLM fit better. The results of a model comparison are documented elsewhere (Fu et al., 2023c).
5. Data from Pairs Forms 2C, 2D, and 2E data were not included in the Likert-to-FC comparison because including all the five pair forms (2A to 2E) would lead to a very large dataset with many items (1,150 forced-choice items plus 560 Likert statements) making model estimation slow and difficult due to the demand on computer memory.
6. Concurrent calibration is a planned missing design method to estimate item parameters from multiple forms in a single analysis (Kolen & Brennan, 2014).

Acknowledgement

Zhitong Yang helped build the three subset pair forms from the two triplet forms. Daniel Fishtein and Yuan Wang managed the administration of all test forms. Nimmi Devasia and Steven Holtzman prepared the test data used in this paper. We are grateful for their contributions to this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Jianbin Fu http://orcid.org/0000-0001-8026-242X
Patrick C. Kyllonen http://orcid.org/0000-0002-6517-4576
Xuan Tan http://orcid.org/0000-0002-8260-3504

Data availability statement

The data supporting this study's findings are properties of ETS and not publicly available. The data may be available on request from the corresponding author on a case-by-case basis.

References

Alreck, P. L., & Settle, R. B. (1994). *The survey research handbook: Guidelines and strategies for conducting a survey* (2nd ed.). McGraw Hill.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). CRC Press. https://doi.org/10.1201/9781482276725

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, *81*(4), 135–160. https://doi.org/10.1007/s11336-014-9434-9

Brown, A., & Bartram, D. (2009–2011). *OPQ32r technical manual*. SHL Group.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Cai, L. (2010). High-dimensional exploratory item factor analysis by a metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, *75*(1), 33–57. https://doi.org/10.1007/s11336-009-9136-x

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Cheung, M. W. L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling*, *9*(1), 55–77. https://doi.org/10.1207/S15328007SEM0901_4

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, *18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2012, April). *Examining the viability of recent models for forced-choice data* [Paper presentation]. American Educational Research Association, Vancouver, BC, Canada.

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.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods, and applications*. Springer. https://doi.org/10.1007/978-3-642-34333-9

Fowler, F. J. (2006). *Survey research methods* (3rd ed.). Sage.

Friedman, H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, *9*, 114–123. https://ssrn.com/abstract=2333648

Fu, J. (2019). *Maximum marginal likelihood estimation with an expectation-maximization algorithm for multigroup/mixture multidimensional item response theory models* (ETS Research Report No. 19–35). Educational Testing Service. https://doi.org/10.1002/ets2.12272

Fu, J., Tan, X., & Kyllonen, P. C. (2023a). *The Rank-2PL IRT models for forced-choice questionnaires: Maximum marginal likelihood estimation with an EM algorithm* [Manuscript submitted for publication]. Educational Testing Service.

Fu, J., Tan, X., & Kyllonen, P. C. (2023b). A comparison of item response theory models for multidimensional forced-choice questionnaires: Real data examples [Unpublished Manuscript]. Educational Testing Service.

Fu, J., Tan, X., & Kyllonen, P. C. (2023c). *Can the generalized graded unfolding model fit dominance statements?* [Manuscript submitted for publication]. Educational Testing Service.

Garnier-Villarreal, M., Merkle, E. C., & Magnus, B. E. (2021). Between-item multidimensional IRT: How far can the estimation methods go? *Psych*, *3*(3), 404–421. https://doi.org/10.3390/psych3030029

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, *36*(3), 341–355. https://doi.org/10.1108/00483480710731310

Haladyna, T. M., & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education*, *2*(1), 51–78. https://doi.org/10.1207/s15324818ame0201_4

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. https://doi.org/10.1111/jedm.12183

King, G., Murray, C. J. L., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191–207. https://doi.org/10.1017/S000305540400108X

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer. https://doi.org/10.1007/978-1-4939-0317-7

Lee, P., & Joo, S. (2021). A new investigation of fake resistance of a multidimensional forced-choice measure: An application of differential item/test functioning. *Personnel Assessment and Decisions*, *7*(1), Article 4. https://doi.org/10.25035/pad.2021.01.004

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement*, *77*(3), 389–414. https://doi.org/10.1177/0013164416646162

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology*, *12*, 732241. https://doi.org/10.3389/fpsyg.2021.732241

McDonald, R. P. (1997). Normal-Ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258–270). Springer.

Morillo, D., Abad, F. J., Kreitchmann, R. S., Leenen, I., Hontangas, P., & Ponsoda, V. (2019). The journey from Likert to forced-choice questionnaires: Evidence of the invariance of item parameters. *Journal of Work and Organizational Psychology*, *35*(2), 75–83. https://doi.org/10.5093/jwop2019a11

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., De la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *40*(7), 500–516. https://doi.org/10.1177/0146621616662226

Morokoff, W. J., & Caflisch, R. E. (1995). Quasi-Monte Carlo integration. *Journal of Computational Physics*, *122*(2), 218–230. https://doi.org/10.1006/jcph.1995.1209

Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement*, *24*(3), 211–223. https://doi.org/10.1177/01466210022031679

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats* (2nd ed.). Kluwer Academic Publishers.

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guilford Press.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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. https://doi.org/10.1177/01466216000241001

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. https://doi.org/10.1177/0146621604273988

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Trueblood, J. S. (2022). Theories of context effects in multialternative, multiattribute choice. *Current Directions in Psychological Science*, *31*(5), 428–435. https://doi.org/10.1177/09637214221109587

Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not just for consumers: Context effects are fundamental to decision making. *Psychological Science*, *24*(6), 901–908. https://doi.org/10.1177/0956797612464241

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 287–307. https://doi.org/10.1348/000711007X193957

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156–170. https://doi.org/10.1037/pas0000971