Behavior Research Methods (2024) 56:600–614
https://doi.org/10.3758/s13428-022-02037-6

Item selection methods in multidimensional computerized adaptive testing for forced-choice items using Thurstonian IRT model

Wang Qin[1] · Yi Zheng[2] · Liu Kai[1] · Cai Yan[1] · Peng Siwei[1] · Tu Dongbo[1]

Accepted: 24 November 2022 / Published online: 7 February 2023

Abstract

Multidimensional computerized adaptive testing for forced-choice items (MFC-CAT) combines the benefits of multidimensional forced-choice (MFC) items and computerized adaptive testing (CAT) in that it eliminates response biases and reduces administration time. Previous studies that explored designs of MFC-CAT only discussed item selection methods based on the Fisher information (FI), which is known to perform unstably at early stages of CAT. This study proposes a set of new item selection methods based on the KL information for MFC-CAT (namely MFC-KI, MFC-K^B, and MFC-KLP) based on the Thurstonian IRT (TIRT) model. Three simulation studies, including one based on real data, were conducted to compare the performance of the proposed KL-based item selection methods against the existing FI-based methods in three- and five-dimensional MFC-CAT scenarios with various test lengths and inter-trait correlations. Results demonstrate that the proposed KL-based item selection methods are feasible for MFC-CAT and generate acceptable trait estimation accuracy and uniformity of item pool usage. Among the three proposed methods, MFC-K^B and MFC-KLP outperformed the existing FI-based item selection methods and resulted in the most accurate trait estimation and relatively even utilization of the item pool.

Keywords MFC-CAT · Thurstonian IRT model · Fisher information · Kullback–Leibler information · Forced-choice items · Item selection methods

Personality assessments that rely on respondent self-report have been widely used for personnel selection. Such assessments typically adopt single-statement formats, such as Likert-type items, where respondents are presented with one statement at a time and are required to choose one among several alternatives (e.g., agree/disagree). However, especially for high-stakes testing, this format is vulnerable to faking and other types of response biases, such as central tendency, acquiescence, socially desirable responding, halo effects, leniency, and impression management (Brown & Maydeu-Olivares, 2011; Cheung & Chan, 2002; Morrison & Bies, 1991). To address these concerns, one alternative is multidimensional forced-choice (MFC) item formats (Brown & Maydeu-Olivares, 2011). Instead of evaluating each statement separately, respondents are presented with blocks consisting of two or more similarly attractive statements, in which each statement is assumed to measure only one personality trait. Respondents are required to make comparative judgments, choosing between statements according to the extent to which the statements describe their preferences or behavior (Brown & Maydeu-Olivares, 2013). While comparative judgments may reduce response biases, the MFC item formats have also met controversy (Brown & Maydeu-Olivares, 2013; Walton et al., 2020). One commonly cited problem is that the traditional scoring approaches of MFC items produce ipsative data, that is, the total score of a test is constant for all respondents. Ipsative scoring distorts individual profiles (i.e., it is impossible to achieve all high or all low scores), and creates challenges in estimating construct validity, criterion-related validity, and reliability (Brown & Maydeu-Olivares, 2013; Dueber et al., 2019). To address such issues, a series of MFC item response theory (IRT) models have been proposed (e.g., Andrich, 1995; Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; Stark et al., 2005; Wang et al., 2017; Zinnes & Griggs, 1974) to model comparative responses generated via forced-choice items. For example, Stark et al. (2005)

✉ Cai Yan
cy1979123@aliyun.com

✉ Tu Dongbo
tudongbo@aliyun.com

[1] Jiangxi Normal University, Nanchang, China

[2] Arizonal State Univerity, Tempe, AZ, USA

developed the multi-unidimensional pairwise-preference (MUPP) model for blocks only containing two statements, and Brown and Maydeu-Olivares (2011) developed the Thurstonian IRT (TIRT) model, which can model blocks with more than two statements.

Recently, the integration of MFC item formats and computerized adaptive testing (MFC-CAT) has gained increasing attention as studies demonstrate great advantages, such as reducing testing time, obtaining more information with a shorter test, and improving measurement accuracy (e.g., Joo et al., 2020; Stark et al., 2012). A few studies explored adaptive testing of personality using forced-choice IRT models, but most of them have focused exclusively on ideal-point models. For example, Borman et al. (2001) compared a unidimensional forced-choice CAT with other CAT rating scales in terms of reliability, validity, and accuracy of performance ratings. Stark et al. (2012) implemented simulation studies based on the MUPP model (Stark et al., 2005), where they examined the effects of dimensionality, test length, inter-trait correlations, and other test design specifications on latent trait estimation accuracy in nonadaptive and adaptive situations. Since then, most studies and applications for MFC-CAT have used pairwise preference forced-choice items, and these studies have shown more efficient trait estimation than nonadaptive tests of an equal length (e.g., Aon Hewitt, 2015; Drasgow et al., 2012; Stark et al., 2012, 2014). To explore the benefits of MFC-CAT with more than two statements in a block, Joo et al. (2020) compared the accuracy of latent trait estimation with MFC pair, triplet, and tetrad tests using adaptive item selection based on the GGUM-RANK (generalized graded unfolding-RANK) model (Hontangas et al., 2015; Joo et al., 2018).

While the above studies all used ideal-point models, another group of IRT models developed for MFC items is dominance models (Wang et al., 2017), such as Maydeu-Olivares and Brown's (2010) TIRT model, Wang et al.'s (2017) Rasch ipsative model (RIM), and a polytomous extension of RIM (Qiu & Wang, 2016). We chose to focus on the TIRT model in this study. The TIRT model can be used to model a variety of forced-choice scales and has demonstrated efficacy in accommodating many combinations of traits and block sizes, which makes it widely applicable to many existing forced-choice questionnaires, such as the Survey of Interpersonal Values (Gordon, 1976), the Customer Contact Styles Questionnaire (SHL, 1997), and the Occupational Personality Questionnaire (SHL, 2006) (Brown & Maydeu-Olivares, 2011). Therefore, developing an adaptive testing approach based on the TIRT model presents a promising gateway towards further applications of MFC-CAT in personality tests that saves substantial cost of test administration.

In an adaptive test, the method used to select items from the item pool for each test-taker adaptively as the test progresses exerts a significant influence on measurement accuracy, test validity, and uniformity of item pool usage. Among the existing item selection methods for CAT, a group of methods developed for single-statement multidimensional CAT (MCAT; e.g., Chang & Ying, 1996; Mulder & van der Linden, 2009, 2010; Segall, 1996; Veldkamp & van der Linden, 2002) provides the foundation for this study, because MFC items measure multidimensional latent traits.

Among studies of item selection methods for single-statement MCAT, Mulder and van der Linden (2009) compared several methods based on the Fisher information (FI) and found that the estimation accuracy of the A-optimality method was slightly better than that of the D-optimality method, and the E-optimality method was the most unstable method. Although the FI-based item selection methods have achieved great popularity, several problems need to be addressed. For example, one assumption of the FI-based item selection methods is that the estimated trait levels are close to their true values, which is often violated at an early stage of CAT when few items have been administered, namely the attenuation paradox issue (e.g., Chang & Ying, 1996; Wang & Chang, 2011). When items with high FI are selected to match inaccurate trait estimates, the adaptive test loses efficiency and item exposure rates become uneven (Chang & Ying, 1996; Lin, 2012). As a global information index, the Kullback–Leibler (KL) information (Chang & Ying, 1996) has been proposed as an alternative to the FI to be used for CAT item selection. Veldkamp and van der Linden (2002) extended the KL information index (KL index, KI) method to multidimensional scenarios and proposed the posterior expectation KL information method (the K^B method), and illustrated that the KL-based item selection methods performed better in estimation accuracy than the FI-based item selection methods.

Note that research on item selection methods for single-statement MCAT so far has mainly concentrated on single-statement items. Although several studies have explored multidimensional forced-choice IRT (MFC-IRT) under nonadaptive testing (e.g., Brown & Maydeu-Olivares, 2011; Hontangas et al., 2015; Joo et al., 2018; Stark et al., 2005; Wang et al., 2017), only two studies so far discussed adaptive item selection methods in MFC-CAT contexts (Joo et al., 2020; Stark et al., 2012). Stark et al. (2012) conducted four simulation studies to explore the effects of test length, dimensionality, inter-trait correlations, and the advantages of adaptive item selection on the accuracy and precision of latent trait estimates for pairwise preference testing. Joo et al. (2020) conducted simulations of MFC-CAT with pair, triplet, and tetrad formats using the FI-based item selection methods, specifically the A-optimality method for MFC items (MFC-A-optimality). In contrast to FI-based item selection item selection methods, methods based on the KL information have not been studied in the MFC-CAT contexts. Hence, this article focuses on the extension and application of item selection methods based on the KL information for MFC-CAT.

To achieve the above goals and provide a foundation for MFC-CAT research using KL-based item selection methods,

this article is organized as follows: First, a brief summary of the TIRT model is presented. Second, we provide an introduction of the FI-based item selection methods that have been used in MFC-CAT contexts and present the proposed extension of the proposed KL-based item selection methods from single-statement MCAT to MFC-CAT. Third, we describe two Monte Carlo simulation studies to explore statistical properties and feasibility of these methods in MFC-CAT. We also discuss how test length, dimensionality, and inter-trait correlation affect the estimation accuracy and uniformity of item pool usage of MFC-CAT. Next, we present a simulation study based on real data using the item pool of the Big-Five factor marker questionnaire with forced-choice items to examine the empirical efficiency of the proposed item selection methods in a personality assessment application. We compare the latent trait estimation accuracy and uniformity of item pool usage of the new methods with the existing methods. Finally, we discuss limitations and recommendations.

TIRT

Thurstone (1927) proposed the law of comparative judgment to describe comparative choices made between statements in a forced-choice item block. This law assumes that each of the two statements (i.e., i and m) in a block elicits a corresponding utility (i.e., t_i and t_m). A respondent prefers to choose the statement with the larger utility. Let $\mathcal{Y}_l$ denote the observed binary outcome and $\mathcal{Y}_l^*$ denote the unobserved difference of utilities for a pairwise comparison, $l=\{i,m\}$, within a forced-choice item block.

$$\mathcal{Y}_l = \begin{cases} 1, & \textit{if statement } i \textit{ is preferred to statement } m, \\ 0, & \textit{if statement } m \textit{ is preferred to statement } i. \end{cases} \tag{1}$$

$$\mathcal{Y}_l^* = t_i - t_m. \tag{2}$$

Then, Thurstone's (1927) law can be written as the relationship between the observed binary outcome $\mathcal{Y}_l$ and the unobserved difference of utilities $\mathcal{Y}_l^*$:

$$\mathcal{Y}_l = \begin{cases} 1, & \textit{if } \mathcal{Y}_l^* \geq 0, \\ 0, & \textit{if } \mathcal{Y}_l^* < 0. \end{cases} \tag{3}$$

Based on Thurstone's (1927) law of comparative judgment, Brown & Maydeu-Olivares (2010, 2011) developed the TIRT model, which can be used to model a variety of forced-choice scales and has demonstrated efficacy in accommodating many combinations of traits and block sizes. When comparing statement i measuring latent trait η_a and statement m measuring the latent trait η_b, the item characteristic function (ICF) of the binary outcome $\mathcal{Y}_l$ can be described as

$$P(\mathcal{Y}_l = 1 | \eta_a, \eta_b) = \Phi\left(\frac{-\gamma_l + \lambda_i\eta_a - \lambda_m\eta_b}{\sqrt{\psi_i^2 + \psi_m^2}}\right), \tag{4}$$

where $\Phi(x)$ denotes the cumulative probability function of the standard normal distribution evaluated at x, γ_l is the threshold parameter for binary outcome γ_l, λ_i and λ_m are the statements' factor loadings, and ψ_i^2 and ψ_m^2 denote the statements' uniqueness.

Now, let

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_m^2}}, \beta_i = \frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_m^2}}, \beta_m = \frac{\lambda_m}{\sqrt{\psi_i^2 + \psi_m^2}}, \tag{5}$$

then the TIRT model (defined by Eq. 4) can be written in an intercept/slope form as

$$P(\mathcal{Y}_l = 1 | \eta_a, \eta_b) = \Phi(\alpha_l + \beta_i\eta_a - \beta_m\eta_b), \tag{6}$$

where α_l is the intercept parameter for binary outcome γ_l, and β_i and β_m are the slope parameters for statement i and statement m, respectively.

To help readers better understand the TIRT model and facilitate computations, in this study, we replaced the cumulative probability function of the standard normal distribution in the TIRT model with a logistic function by referring to the processing method adopted by Morillo et al. (2016):

$$P(\mathcal{Y}_l = 1 | \eta_a, \eta_b) = \frac{1}{1 + \exp\left[-(\beta_i\eta_a - \beta_m\eta_b + \alpha_l)\right]}. \tag{7}$$

Note that if a forced-choice item block contains more than two statements, there exist more than one pairwise comparison (e.g., three pairwise comparisons for a block with three statements). For example, the comparisons between three statements A, B, and C for a block can be presented as follows.

Three-statement block			Pairwise comparisons		
A	B	C	A vs. B	B vs. C	A vs. C

Note. A vs. B = the statement A is compared with the statement B; *B vs. C* = the statement A is compared with the statement B; *A vs. C* = the statement A is compared with the statement B.

Extension of item selection methods from MCAT to MFC-CAT

In order to facilitate the presentation, several notations will be introduced here. d denotes the measured trait dimensions in tests, $z \in \{1, \cdots, d\}$ denotes the component of latent trait vector $\boldsymbol{\eta}$ ($\boldsymbol{\eta}$ is a d-dimensional vector of latent traits), R represents the item pool, S_{k-1} denotes the set of the first $k-1$ administered blocks, $\boldsymbol{U}_{k-1}$ denotes the response vector of

the $k-1$ administered blocks, j_k denotes the block administered as the k_{th} block in the test, and R_k denotes the set of blocks remaining in the item pool after the $(k-1)th$ block is administered.

Under the framework of single-statement MCAT, a group of item selection methods have been developed (e.g., Chang & Ying, 1996; Mulder & van der Linden, 2009, 2010; Segall, 1996; Veldkamp & van der Linden, 2002). At present, only the MFC-A-optimality method, which is based on the FI, has been applied to MFC-CAT (Joo et al., 2020). The FI-based item selection methods assume that the intermediate trait estimates are close to their true values, which is often violated at the beginning of CAT due to few items having been administered (Mulder & van der Linden, 2009; Segall, 1996). One alternative to the FI to be used for CAT item selection is the global KL information (Chang & Ying, 1996), which is a measure of discrepancy between two probability distributions. It does not require that the estimated latent trait, $\hat{\boldsymbol{\eta}}$, be close to the true value, $\boldsymbol{\eta}$, and it is more robust than FI against early-stage estimation instability (Lima Passos et al., 2007). Several studies have demonstrated that the performance of KL-based item selection methods is more stable, efficient, and precise in terms of trait estimation, especially at an early stage of CAT or for a short CAT (Chang & Ying, 1996; Veldkamp & van der Linden, 2002; Wang et al., 2011). Therefore, with this study, we propose an extension of KL-based item selection methods from the single-statement MCAT context to the MFC-CAT context. We then explore whether the properties of the KL-based item selection methods continue to hold true in the MFC-CAT context. In the following sections, we first describe the FI-based item selection methods for MFC-CAT and then introduce the proposed KL-based item selection methods for MFC-CAT.

FI-based item selection methods for MFC-CAT

Under the framework of MFC-CAT, the FI is given as a matrix. With the TIRT model employed, the FI matrix for Block j can be defined as

$$I_j^*(\boldsymbol{\eta}) = P_j(\boldsymbol{\eta})Q_j(\boldsymbol{\eta})\begin{bmatrix} \beta_{1j}^2 & \beta_{1j}\beta_{2j} & \cdots & \beta_{1j}\beta_{dj} \\ \beta_{2j}\beta_{1j} & \beta_{2j}^2 & \cdots & \beta_{2j}\beta_{dj} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{dj}\beta_{1j} & \beta_{dj}\beta_{2j} & \cdots & \beta_{dj}^2 \end{bmatrix} =$$

$$\begin{bmatrix} \beta_{1j}^2 P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) & \beta_{1j}\beta_{2j}P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) & \cdots & \beta_{1j}\beta_{dj}P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) \\ \beta_{2j}\beta_{1j}P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) & \beta_{2j}^2 P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) & \cdots & \beta_{2j}\beta_{dj}P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{dj}\beta_{1j}P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) & \beta_{dj}\beta_{2j}P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) & \cdots & \beta_{dj}^2 P_j(\eta_a,\eta_b)Q_j(\eta_a,\eta_b) \end{bmatrix} \tag{8}$$

where d denotes the number of dimensions measured by the test. $P_j(\eta_a,\eta_b)$ denotes the probability of preferring the first statement measuring trait η_a over the second statement measuring trait η_b in a pairwise comparison, which is the shorthand notation for $P(\mathcal{Y}_l = 1|\eta_a,\eta_b)$ in Eq. 7, and $Q_j(\eta_a,\eta_b) = 1-P_j(\eta_a,\eta_b)$. Note that a single pair block only involves statements pertaining to two of the d dimensions, and hence the information matrix has only four nonzero elements, and all other elements equal 0. Likewise, a single triplet block only involves three of the d dimensions, and the information matrix has only nine nonzero elements. Also note that different blocks have different nonzero entries depending on the dimensions measured by the block respectively. However, these information matrices can be summed up across different blocks as in Eqs. 9 and 10 below because they share the same structure.

Under the conditional independence assumption of the responses given $\boldsymbol{\eta}$, the information matrix of a test is equal to the sum of the block information matrices. Therefore, the FI matrix of the test can be expressed as

$$I^*(\boldsymbol{\eta}) = \sum_{j=1}^{J} I_j^*(\boldsymbol{\eta}). \tag{9}$$

Then, the FI matrix of a set of S_{k-1} blocks could be computed by

$$I_{S_{k-1}}^*(\boldsymbol{\eta}) = \sum_{j\in S_{k-1}} I_j^*(\boldsymbol{\eta}). \tag{10}$$

Based on the FI, three popular optimality methods, namely the D-optimality method, the A-optimality method, and the E-optimality method, have been developed for single-statement MCAT (Mulder & van der Linden, 2009). The MFC-A-optimality method has been used in previous MFC-CAT studies but without being expressed with formulation (Joo et al., 2020). Mulder and van der Linden (2009) found that E-optimality lacks robustness in applications with sparse data. Therefore, we present the formulas for MFC-A-optimality and MFC-D-optimality as the following.

The D-optimality method for MFC items (MFC-D-optimality) The MFC-D-optimality method seeks to select the next item that maximizes the determinant of the information matrix, and this method can be expressed as

$$j_k = \underset{j\in R_k}{\operatorname{argmin}}\left\{ \det\left[I_{S_{k-1}}^*\left(\hat{\boldsymbol{\eta}}_{k-1}\right) + I_j^*\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right]\right\}, \tag{11}$$

where $\left[I_{S_{k-1}}^*\left(\hat{\boldsymbol{\eta}}_{k-1}\right) + I_j^*\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right]$ denotes the sum of the information matrix after the $k-1$ blocks already administered and the information matrix for candidate block j.

The A-optimality method for MFC items (MFC-A-optimality) This method seeks to select the next block that minimizes the sum of the (asymptotic) sampling variances of the

trait estimators, which is equivalent to minimizing the trace of the inverse of the information matrix. Its formulation is

$$j_k = \arg\min_{j\in R_k}\left\{ trace\left[\left(I^*_{S_{k-1}}\left(\hat{\boldsymbol{\eta}}_{k-1}\right)+I^*_j\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right)^{-1}\right]\right\} = \arg\max_{j\in R_k}\left\{\frac{\det\left[I^*_{S_{k-1}}\left(\hat{\boldsymbol{\eta}}_{k-1}\right)+I^*_j\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right]}{\sum_{z=1}^{d}\det\left(\left[I^*_{S_{k-1}}\left(\hat{\boldsymbol{\eta}}_{k-1}\right)+I^*_j\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right]_{[z,z]}\right)}\right\}, \tag{12}$$

where $\hat{\boldsymbol{\eta}}_{k-1}$ denotes the trait estimator after the first $k-1$ blocks are administrated, and $\left[I^*_{S_{k-1}}\left(\hat{\boldsymbol{\eta}}_{k-1}\right)+I^*_j\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right]_{[z,z]}$ is the submatrix after deleting the *zth* row and column of the information matrix $\left[I^*_{S_{k-1}}\left(\hat{\boldsymbol{\eta}}_{k-1}\right)+I^*_j\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right]$.

The proposed extension of KL-based item selection methods for MFC-CAT

Several adaptive selection methods based on KL information have been developed for single-statement MCAT (Chang & Ying, 1996; Mulder & van der Linden, 2010; Veldkamp & van der Linden, 2002; Wang & Chang, 2011), such as the KL index (KI) method, posterior expected KL information method (K^B), and the KL distance between subsequent posteriors (KLP) method. To adapt the above KL-based item selection methods to MFC-CAT, we propose to modify the classical KL information as

$$KL^*_j(\hat{\boldsymbol{\eta}} \parallel \boldsymbol{\eta}) = \sum_{c=1}^{C_j} L_{cj}(\hat{\boldsymbol{\eta}}) \log\left[\frac{L_{cj}(\hat{\boldsymbol{\eta}})}{L_{cj}(\boldsymbol{\eta})}\right], \tag{13}$$

where $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$ denote the unknown and estimated latent trait vectors, respectively; j denotes the *jth* block, C_j is the number of possible scoring patterns for Block j (e.g., a block with three statements, such as A, B, and C, has six possible scoring patterns; see Table 1); $c(c=1,2,\cdots,C_j)$ indicates the *cth* scoring pattern.

$L_{cj}(\boldsymbol{\eta})$ and $L_{cj}(\hat{\boldsymbol{\eta}})$ refer to the block response probability, namely the likelihood of pairwise comparison response probability, for latent traits $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$, respectively, given the *cth* scoring patterns of Block j. The expression of $L_{cj}(\boldsymbol{\eta})$ and $L_{cj}(\hat{\boldsymbol{\eta}})$ are respectively given by

$$L_{cj}(\boldsymbol{\eta}) = \prod_{a=1}^{K_{j-1}}\prod_{b=a+1}^{K_j} P_j\left(\eta_a,\eta_b\right)^{\mathcal{Y}_l}\left[1-P_j\left(\eta_a,\eta_b\right)\right]^{(1-\mathcal{Y}_l)}, \tag{14}$$

and

$$L_{cj}(\hat{\boldsymbol{\eta}}) = \prod_{a=1}^{K_{j-1}}\prod_{b=a+1}^{K_j} P_j\left(\eta_a,\eta_b\right)^{\mathcal{Y}_l}\left[1-P_j\left(\eta_a,\eta_b\right)\right]^{(1-\mathcal{Y}_l)} \tag{15}$$

Table 1 All possible scoring patterns in a block with three statements

Scoring pattern	*A vs. B*	*B vs. C*	*A vs. C*	*Computation of* $L_{cj}(\boldsymbol{\eta})$
1	1	0	0	$[P_{(A>B)}(\boldsymbol{\eta})][P_{B<C}(\boldsymbol{\eta})][P_{(A<C)}(\boldsymbol{\eta})]$
2	0	1	0	$[P_{(A<B)}(\boldsymbol{\eta})][P_{B>C}(\boldsymbol{\eta})][P_{(A<C)}(\boldsymbol{\eta})]$
3	1	0	1	$[P_{(A>B)}(\boldsymbol{\eta})][P_{B<C}(\boldsymbol{\eta})][P_{(A>C)}(\boldsymbol{\eta})]$
4	0	1	1	$[P_{(A<B)}(\boldsymbol{\eta})][P_{B>C}(\boldsymbol{\eta})][P_{(A>C)}(\boldsymbol{\eta})]$
5	1	1	1	$[P_{(A>B)}(\boldsymbol{\eta})][P_{B>C}(\boldsymbol{\eta})][P_{(A>C)}(\boldsymbol{\eta})]$
6	0	0	0	$[P_{(A<B)}(\boldsymbol{\eta})][P_{B<C}(\boldsymbol{\eta})][P_{(A<C)}(\boldsymbol{\eta})]$

A vs. B = statement A is compared with statement B; *B vs. C* = statement B is compared with statement C; *A vs. C* = statement A is compared with statement C. The observed binary outcomes are coded as 1 if respondents prefer the former statement over the later statement in the above pairwise comparisons: *A > B, B > C and B > C; otherwise, 0*

where K_j denotes the number of statements in Block j, $\mathcal{Y}_l$ is defined in Eq. 1, and $P_j(\eta_a,\eta_b)$ is defined in Eq. 7. Brown and Maydeu-Olivares (2011) proposed that effects of ignoring these dependencies on the latent trait estimates have been shown to be negligible in applications involving a single ranking task, and they are likely to be even smaller in forced-choice questionnaires where blocks are smaller and there are fewer local dependencies per item. So, throughout this article, we will use the simplifying assumption that the ICFs for the binary outcomes are locally independent.

The proposed extension of the KI method for MFC items (MFC-KI) The KL information as shown in Eq. 13 is a function of the true trait $\boldsymbol{\eta}$, but the true trait value is unknown. Therefore, Chang and Ying (1996) proposed to calculate the KL index (KI) by integrating the estimated trait $\hat{\boldsymbol{\eta}}$. The extended KI item selection method for MFC items (MFC-KI) can be defined as

$$j_k = \arg\max_{j\in R_k}\left\{KI_j\left(\hat{\boldsymbol{\eta}}_{k-1}\right)\right\} = \arg\max_{j\in R_k}\left\{\int_{\hat{\boldsymbol{\eta}}_k-\delta_{k-1}}^{\hat{\boldsymbol{\eta}}+\delta_{k-1}} KL^*_j\left(\hat{\boldsymbol{\eta}}_{k-1} \parallel \boldsymbol{\eta}\right)\partial\boldsymbol{\eta}\right\}, \tag{16}$$

where $\delta_k = d\sqrt{k-1}$ determines the size of the area on which the average is calculated, d usually takes a value of 3 (Chang & Ying, 1996; Veldkamp & van der Linden, 2002), and $k-1$ denotes the number of blocks that have been administered.

The MFC-KI method selects the blocks with the largest KI value among the remaining blocks R_k in the item pool.

The proposed extension of the KB method for MFC items (MFC-KB) By weighting KL through the posterior distribution of latent trait $\boldsymbol{\eta}$, Veldkamp and van der Linden (2002) proposed a Bayesian version of the KI method, that is, the multidimensional posterior expected KL information method

(K^B). Under the framework of MFC-CAT, the expression of the K^B method for MFC items (MFC-K^B) can be written as

$$j_k = \underset{j\in R_k}{\operatorname{argmax}} K_j^B\left(\hat{\boldsymbol{\eta}}_{k-1}\right) = \underset{j\in R_k}{\operatorname{argmax}} \int_{\eta} KL_j^*\left(\hat{\boldsymbol{\eta}}_{k-1} \parallel \boldsymbol{\eta}\right)\pi_{k-1}\left(\boldsymbol{\eta} \parallel \boldsymbol{U}_{k-1}\right)\partial\boldsymbol{\eta}$$
$$= \underset{j\in R_k}{\operatorname{argmax}} \int_{\eta} \left\{ \sum_{c=1}^{2^{\kappa_j}} L_{cj}\left(u_{jk}|\hat{\boldsymbol{\eta}}_{k-1}\right) \log\left[\frac{L_{cj}\left(u_{jk}|\hat{\boldsymbol{\eta}}_{k-1}\right)}{L_{cj}\left(u_{jk}|\boldsymbol{\eta}\right)}\right] \right\} \pi_{k-1}\left(\boldsymbol{\eta}|\boldsymbol{U}_{k-1}\right)\partial\boldsymbol{\eta}, \tag{17}$$

where $L_{cj}(u_{jk}|\boldsymbol{\eta})$ and $L_{cj}\left(u_{jk}|\hat{\boldsymbol{\eta}}_{k-1}\right)$ denote the response probability for $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}_{k-1}$ when selecting Block j as the kth administrated block of the test with the response score $u_{jk}(u_{jk}=0,1)$, respectively. $\pi_{k-1}(\boldsymbol{\eta}|\boldsymbol{U}_{k-1})$ indicates the posterior distribution for $\boldsymbol{\eta}$ after $k-1$ blocks have been administrated:

$$\pi_{k-1} = \left(\boldsymbol{\eta}\middle|\boldsymbol{U}_{k-1}\right) = \frac{g(\boldsymbol{\eta})L\left(\boldsymbol{U}_{k-1}|\boldsymbol{\eta}\right)}{\int g(\boldsymbol{\eta})L\left(\boldsymbol{U}_{k-1}|\boldsymbol{\eta}\right)\partial\boldsymbol{\eta}}, \tag{18}$$

where $g(\boldsymbol{\eta})$ denotes a prior distribution for $\boldsymbol{\eta}$, $\boldsymbol{U}_{k-1}$ denotes the response vector of the $k-1$ administered blocks, and $L(\boldsymbol{U}_{k-1}|\boldsymbol{\eta})$ denotes the likelihood associated with response vector $\boldsymbol{U}_{k-1}$.

The proposed extension of the KLP method for MFC items (MFC-KLP) An item should be selected to maximize the divergence between the posterior distributions of $\boldsymbol{\eta}$. One of the possible responses to the candidate item would move the posterior distribution of $\boldsymbol{\eta}$ toward the respondents' true trait, and the other would move it away from the respondents' true trait, and then this level of divergence between the response distributions generated by two different trait levels can be formalized by the KL information, i.e., the KLP (Mulder & van der Linden, 2010). The KLP method selects the item with the maximum expected KLP distributions $\pi_{k-1}(\boldsymbol{\eta}|\boldsymbol{U}_{k-1})$ and $\pi_k(\boldsymbol{\eta}|\boldsymbol{U}_{k-1},u_{jk})$ (Mulder & van der Linden, 2010; Tu et al., 2018). Under the framework of MFC-CAT, the expression of the KLP method for MFC items (MFC-KLP) can be defined as

$$j_k = \underset{j\in R_k}{\operatorname{argmax}} KLP_j$$
$$= \underset{j\in R_k}{\operatorname{argmax}} \sum_{c=1}^{Cj} \int_{\eta} L_{cj}\left(u_{jk}|\boldsymbol{U}_{k-1}\right) KL\left(\pi_{k-1}\left(\boldsymbol{\eta}|\boldsymbol{U}_{k-1}\right) \parallel \pi_k\left(\boldsymbol{\eta}\middle|\boldsymbol{U}_{k-1},u_{jk}\right)\right)\partial\boldsymbol{\eta}$$
$$= \underset{j\in R_k}{\operatorname{argmax}} \sum_{c=1}^{C_j} \int_{\eta} L_{cj}\left(u_{jk}|\boldsymbol{U}_{k-1}\right)\pi_{k-1}\left(\boldsymbol{\eta}|\boldsymbol{U}_{k-1}\right) \log\left[\frac{\pi_{k-1}\left(\boldsymbol{\eta}|\boldsymbol{U}_{k-1}\right)}{\pi_k\left(\boldsymbol{\eta}\middle|\boldsymbol{U}_{k-1},u_{jk}\right)}\right], \tag{19}$$

where the predictive probability and posterior distribution of the *kth* candidate block after $k-1$ blocks have been administrated can be defined as follows

$$L_{cj}\left(u_{jk}\middle|\boldsymbol{U}_{k-1}\right) = \int_{\eta} L_{cj}\left(u_{jk}|\boldsymbol{\eta}\right)\pi_{k-1}\left(\boldsymbol{\eta}\middle|\boldsymbol{U}_{k-1}\right)\partial\boldsymbol{\eta}, \tag{20}$$

$$\pi_k\left(\boldsymbol{\eta}\middle|\boldsymbol{U}_{k-1},u_{jk}\right) = \frac{g(\boldsymbol{\eta})L\left(\boldsymbol{U}_{k-1},u_{jk}|\boldsymbol{\eta}\right)}{\int g(\boldsymbol{\eta})L\left(\boldsymbol{U}_{k-1},u_{jk}|\boldsymbol{\eta}\right)\partial\boldsymbol{\eta}}, \tag{21}$$

where $L(\boldsymbol{U}_{k-1},u_{jk}|\boldsymbol{\eta}) = L(\boldsymbol{U}_{k-1}|\boldsymbol{\eta})L_{cj}(u_{jk}|\boldsymbol{\eta})$ denotes the likelihood of the *kth* candidate block after $k-1$ blocks have been administrated.

The R codes of the proposed MFC-KI, MFC-K^B, and MFC-KLP methods can be found at https://osf.io/bmg8r/.

Simulation studies

Two Monte Carlo simulation studies and a simulation based on real data were conducted to evaluate the proposed KL-based item selection methods for MFC-CAT. Study 1 and study 2 compared the performance of the newly developed KL-based item selection methods against the existing FI-based item selection methods in terms of trait estimation accuracy and uniformity of item pool usage in three-dimensional and five-dimensional MFC-CAT scenarios, respectively. Finally, the simulation based on real data (the Big-Five factor marker questionnaire response data) further investigated the feasibility of the proposed KL-based item selection methods in real MFC-CAT testing situations.

Simulation study 1

Simulation design

In this study, we were focused on the triplet test, where three latent trait dimensions (d=3) were measured in a block consisting of three statements, because it is more common in block matching. An item pool containing 100 triplet blocks were pre-assembled following methods used by Joo et al. (2020). Specifically, Joo et al. (2020) found that the percentage of unidimensional blocks had little influence on GGUM-RANK scoring. Therefore, we only considered the case that each statement in each block measures different traits from different dimensions. Item responses were simulated based on the TIRT model. The slope parameters β and the intercept parameters α were randomly sampled from a lognormal distribution and a normal distribution respectively. To compare item selection methods under a variety of test scenarios, we varied the correlations between traits (inter-trait correlations) at 0 and 0.5, and varied the test length at 5, 10, and 15 blocks. To simulate data for this study, 500 true latent trait vectors were randomly generated from a multivariate standard normal distribution with the abovementioned inter-trait correlations.

In sum, there were 5 (item selection method: MFC-A-optimality, MFC-D-optimality, MFC-KI, MFC-K^B, and MFC-KLP) × 2 (inter-trait correlation: 0, 0.5) × 3 (test

length: 5, 10, 15) = 30 simulation conditions. For each condition, 20 replications were performed. This study used the expected a posteriori (EAP; Bock & Mislevy, 1982) estimation for latent trait estimation, in which the trait prior distribution was set to a multivariate standard normal distribution. Gauss-Hermite numerical integration (Glas, 1992) was used for the parameter estimation and the integration was taken over the range of trait [−3, +3]. All simulation code was written in R.

Evaluation criteria The performance of each method was evaluated by trait estimation accuracy and uniformity of item pool usage. In this study, the indices to evaluate trait estimation accuracy were bias (BIAS), root mean squared error (RMSE) and the correlation between the generating and estimated traits (CORR), while the index to evaluate uniformity of item pool usage included chi-square (x^2).

The three trait estimation accuracy indices were computed as follows:

$$BIAS_d = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{\eta}_{nd} - \eta_{nd}\right), \tag{22}$$

$$RMSE_d = \sqrt{\frac{1}{N}\sum_{n-1}^{N}\left(\hat{\eta}_{nd} - \eta_{nd}\right)^2}, \tag{23}$$

$$CORR_d = \frac{\sum_{n=1}^{N}\left(\eta_{nd} - \bar{\eta}_d\right)\left(\hat{\eta}_{nd} - \bar{\hat{\eta}}_d\right)}{S_{\eta_d} S_{\hat{\eta}_d}}, \tag{24}$$

where N is the total number of respondents in the test, n denotes the *nth* respondent, and η_{nd} and $\hat{\eta}_{nd}$ are the true traits and the estimated traits of respondent n respectively. $\bar{\eta}_d$ and S_{η_d} are the mean value and standard deviation of the true traits of all respondents, while $\bar{\hat{\eta}}_d$ and $S_{\hat{\eta}_d}$ are the mean value and standard deviation of the estimated traits respectively. The smaller the BIAS and RMSE values and the larger the CORR values, the higher the trait estimation accuracy.

The x^2 index is employed to measure the overall exposure and it is defined as

$$x^2 = \sum_{j=1}^{J}\frac{\left[ER_j - E\left(ER_j\right)\right]^2}{E\left(ER_j\right)}, \tag{25}$$

where $ER_j = f_i/N$ is the exposure rate of block j, f_j is the number of times that block j is selected. $E(ER_j) = T/J$ is the expected exposure rate of block j, T is the test length, and J is the number of blocks in item pool (Chang & Ying, 1999). The smaller the x^2 is, the more evenly the whole item pool is used.

Results of study 1

Trait estimation accuracy The trait estimation accuracies of the five compared item selection methods (MFC-A-optimality, MFC-D-optimality, MFC-KI, MFC-K^B, and MFC-KLP) under different inter-trait correlations and test lengths in the three-dimensional MFC-CAT scenarios are presented in Table 2. As shown, all average RMSEs ranged from 0.308 to 0.582, all CORRs ranged from 0.803 to 0.951, and all biases were around zero, which indicates that the trait estimation accuracy of MFC-CAT adaptive methods was relatively high across all three-dimensional conditions. Except for the MFC-KI method, all the other methods achieved satisfactory estimation accuracy, which demonstrates their applicability to MFC-CAT. Note that: (1) Among the existing FI-based item selection methods, MFC-A-optimality was comparable to MFC-D-optimality with slightly higher estimation accuracy of the latter. (2) Among the proposed KL-based item selection methods, MFC-KI performed noticeably worse than the other two methods, as it rendered the largest RMSE and BIAS, and the smallest CORR. (3) Among the five item selection methods, MFC-K^B performed similarly to MFC-KLP with higher trait estimation accuracy, which demonstrates that the proposed methods based on the KL information outperformed the existing methods based on the FI, especially when the test is short. These results are in line with the original expectations of this study.

Other factors held constant, the inter-trait correlations have a non-negligible influence on the trait estimation accuracy of MFC-CATs implemented in this study. The RMSEs increase and the CORRs decrease as the inter-trait correlations increase. In other words, the trait estimation accuracy of all methods reduces considerably as the inter-trait correlations increase, which is consistent with the results of Brown and Maydeu-Olivares's (2011) and Bürkner et al.'s (2019) study. For example, the average RMSEs of the MFC-K^B method ranged from 0.308 to 0.474 when the inter-trait correlation was 0, and took higher values ranging from 0.315 to 0.515 when the inter-trait correlation was 0.5 (see Table 2).

By contrast, the RMSEs of all methods decrease and the CORRs increase as the test length increases. It was evident that the estimation accuracy gradually improves as the test length increases. According to the results, the estimation accuracy of the 15-block tests performed better than the 5-block or 10-block tests. For example, in the conditions where the inter-trait correlation was 0, the average RMSEs of all methods for the 15-block tests ranged from 0.308 to 0.354, versus 0.474 to 0.531 for the 5-block tests.

Similarly, in the same condition, the CORRs of the 15-block tests ranged from 0.935 to 0.951, as opposed to 0.844 to 0.881 for the 5-block tests. As the test length increases, the difference of estimation accuracy between the proposed item selection methods and the existing MFC

Table 2 Trait estimation accuracy of the five compared item selection methods for three-dimensional MFC-CAT

r	Dimensions	Test length	BIAS					RMSE					CORR				
			MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP
0	Trait1	5	0.001	−0.001	−0.001	0.003	−0.001	0.478	0.475	0.520	0.470	0.469	0.879	0.879	0.853	0.883	0.882
		10	−0.004	−0.002	−0.004	0.002	−0.002	0.370	0.366	0.402	0.349	0.346	0.929	0.931	0.915	0.938	0.937
		15	−0.004	0.001	−0.007	0.001	−0.001	0.324	0.319	0.349	0.305	0.303	0.947	0.948	0.937	0.953	0.952
	Trait2	5	−0.007	−0.005	−0.001	−0.003	−0.008	0.504	0.494	0.548	0.470	0.476	0.861	0.868	0.833	0.881	0.877
		10	−0.005	−0.002	−0.003	−0.001	−0.004	0.378	0.368	0.409	0.354	0.354	0.926	0.929	0.911	0.935	0.934
		15	−0.004	0.002	−0.003	−0.003	−0.003	0.330	0.318	0.355	0.310	0.309	0.944	0.948	0.934	0.950	0.950
	Trait3	5	−0.001	−0.004	−0.002	−0.001	0.001	0.499	0.485	0.525	0.482	0.468	0.861	0.870	0.847	0.875	0.883
		10	−0.001	−0.001	−0.007	0.001	−0.005	0.378	0.371	0.411	0.357	0.351	0.924	0.927	0.909	0.934	0.936
		15	−0.002	0.001	−0.005	0.001	−0.004	0.327	0.321	0.357	0.310	0.307	0.944	0.947	0.933	0.950	0.951
	Mean	5	−0.002	−0.003	−0.001	0.001	−0.003	0.494	0.485	0.531	0.474	0.471	0.867	0.872	0.844	0.880	0.881
		10	−0.003	−0.002	−0.005	0.001	−0.004	0.375	0.368	0.407	0.353	0.350	0.926	0.929	0.912	0.936	0.936
		15	−0.003	0.001	−0.005	0.001	−0.003	0.327	0.319	0.354	0.308	0.306	0.945	0.948	0.935	0.951	0.951
0.5	Trait1	5	0.004	−0.004	−0.001	−0.003	0.001	0.528	0.512	0.580	0.512	0.515	0.847	0.859	0.810	0.860	0.859
		10	0.002	−0.003	0.001	0.003	−0.002	0.385	0.377	0.419	0.365	0.364	0.924	0.928	0.909	0.933	0.933
		15	0.004	−0.001	0.002	0.001	−0.002	0.331	0.322	0.354	0.312	0.312	0.945	0.948	0.937	0.951	0.952
	Trait2	5	0.006	0.004	0.003	−0.006	−0.001	0.554	0.543	0.596	0.521	0.525	0.825	0.835	0.791	0.852	0.848
		10	0.005	0.001	0.003	−0.001	−0.005	0.392	0.383	0.433	0.369	0.367	0.920	0.924	0.899	0.930	0.930
		15	0.005	0.006	0.002	0.001	−0.003	0.334	0.328	0.364	0.318	0.311	0.943	0.945	0.931	0.948	0.950
	Trait3	5	0.004	0.003	0.003	0.001	0.007	0.526	0.515	0.571	0.513	0.508	0.844	0.851	0.808	0.855	0.860
		10	0.007	0.003	−0.001	−0.003	−0.005	0.386	0.379	0.425	0.369	0.367	0.921	0.924	0.902	0.929	0.930
		15	0.005	0.004	0.005	−0.004	−0.005	0.330	0.324	0.361	0.315	0.312	0.943	0.946	0.931	0.949	0.951
	mean	5	0.005	0.001	0.002	−0.013	0.002	0.536	0.523	0.582	0.515	0.516	0.839	0.848	0.803	0.856	0.856
		10	0.005	0.001	0.001	0.001	−0.004	0.388	0.380	0.426	0.368	0.366	0.922	0.925	0.903	0.931	0.931
		15	0.005	0.003	0.003	−0.011	−0.003	0.332	0.325	0.360	0.315	0.312	0.944	0.946	0.933	0.949	0.951

r = inter-trait correlations, *Test length* = administered MFC items, *mean* = average values of traits, RMSE = root mean square error, *CORR* = correlation between generated and estimated traits

methods narrowed down. In sum, the proposed KL-based MFC-K^B and MFC-KLP methods performed better than the FI-based item selection methods in terms of trait estimation accuracy, especially when the test is short (or equivalently, at an early stage of MFC-CATs). However, the performance of the MFC-KI method needs to be further improved with lower trait estimation accuracy. The same pattern was consistently observed from other indices, as well.

Uniformity of item pool usage Item exposure control is an important component in CAT design and operation, especially for high-stake tests. Stocking and Lewis (1998) pointed out that in order to reduce the cost of item pool development, adaptive selection methods should also maximize the utilization of the item pool. Table 3 shows the results of the x^2 values. The results demonstrated that the proposed MFC-K^B method rendered the lowest x^2 values across five methods. Namely, MFC-K^B outperformed the existing FI-based MFC item selection methods in terms of uniformity of the item pool. The MFC-KLP method promoted greater utilization of the item pool and produced smaller x^2 values at the early stage. However, similar to the performance of estimation accuracy, MFC-KI performed the worst in item pool usage. For the FI-based item selection methods, MFC-A-optimality outperformed MFC-D-optimality in uniformity of item pool usage, though the former's accuracy was slightly worse than the latter. For example, the x^2 values of MFC-D-optimality was as high as 40.605 compared with the largest x^2 values of 39.313 when MFC-A-optimality was used. Overall, the use of the item pool was relatively more even when the KL-based item selection methods were used than when the FI-based item selection methods were used.

Simulation study 2

Simulation design Simulation study 1 mainly discussed the feasibility of all item selection methods under the three-dimensional MFC-CAT scenarios. In practice, however, MFC tests may need to measure more than three dimensions, namely higher-dimensional tests (e.g., TAPAS; Drasgow et al., 2012; Stark et al., 2014). Hence, study 2 intends to further explore the performance of all methods in relatively higher-dimensional (i.e., five-dimensional) MFC-CAT scenarios. At the same time, the performance of each method in the five-dimensional conditions is compared against study 1.

The simulation design of study 2 was mostly the same as study 1, except for the following aspects: first, five latent trait dimensions (d=5) were measured for triplet tests in this study. Furthermore, the number of MFC blocks administered were changed from 5, 10, and 15 blocks to 10, 15, and 20 blocks, respectively. In total, there were 5 (item selection method: MFC-A-optimality, MFC-D-optimality, MFC-KI, MFC-K^B, and MFC-KLP) × 2 (inter-trait correlation: 0, 0.5) × 3 (test length: 10, 15, 20) = 30 simulation conditions. For each condition, 20 replications were conducted. EAP estimation and Gauss-Hermite numerical integration were again utilized for trait estimation with the R program. Study 2 used the same evaluation criteria as study 1.

Results of study 2

Trait estimation accuracy For five-dimensional MFC-CATs, the RMSEs, biases, and CORRs of the five item selection methods are presented in Table 4. Overall, the average biases of all methods under various conditions were between [−0.014, 0.001]. The average RMSEs of each method under various conditions were between [0.341, 0.566], and the mean CORRs of each method were still acceptable, between [0.822, 0.936]. Therefore, the trait estimation accuracy was acceptable, which indicates that the proposed methods are also applicable to MFC-CATs under the higher-dimensional conditions.

Compared with the three-dimensional study (simulation study 1), the estimation accuracy of all methods, especially MFC-KI, decreased significantly with the increase of dimensionality. As can be seen from Table 2 and Table 4, under the three-dimensional conditions, except for MFC-KI, which generated the lowest estimation accuracy, the estimation accuracies of MFC-A-optimality and MFC-D-optimality were relatively high, and the estimation accuracies of

Table 3 The x^2 values of the five compared item selection methods for three-dimensional MFC-CAT

r	Indices	Test length	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP
0	x^2	5	35.370	36.789	44.893	24.740	32.038
		10	25.380	27.676	48.054	24.578	29.259
		15	21.799	24.076	49.569	23.046	26.167
0.5	x^2	5	39.313	40.605	49.935	29.884	36.125
		10	31.623	33.363	54.673	31.516	34.513
		15	28.095	30.264	56.181	29.913	31.453

r = inter-trait correlations, *Test length* = number of administered MFC items

Table 4 Trait estimation accuracy of the five compared item selection methods for five-dimensional MFC-CAT

r	Dimensions	Test length	BIAS					RMSE					CORR				
			MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP
0	Trait1	10	−0.002	0.002	−0.006	0.001	−0.007	0.475	0.441	0.471	0.429	0.428	0.878	0.897	0.881	0.896	0.905
		15	−0.005	−0.001	−0.004	−0.001	−0.008	0.415	0.381	0.412	0.369	0.371	0.909	0.924	0.911	0.924	0.930
		20	−0.005	−0.002	−0.005	0.001	−0.006	0.387	0.353	0.373	0.342	0.341	0.922	0.935	0.928	0.936	0.941
	Trait2	10	−0.007	0.002	−0.017	−0.003	−0.005	0.507	0.468	0.512	0.432	0.442	0.858	0.881	0.854	0.892	0.896
		15	−0.008	−0.004	−0.013	−0.004	−0.005	0.437	0.399	0.432	0.378	0.381	0.897	0.915	0.899	0.920	0.9244
		20	−0.007	−0.003	−0.011	−0.005	−0.005	0.403	0.363	0.393	0.348	0.349	0.913	0.931	0.917	0.933	0.937
	Trait3	10	−0.004	0.003	−0.002	−0.001	−0.007	0.471	0.440	0.485	0.428	0.421	0.879	0.895	0.870	0.892	0.907
		15	−0.003	0.001	−0.006	−0.001	−0.007	0.415	0.380	0.411	0.371	0.365	0.908	0.924	0.909	0.920	0.931
		20	−0.004	0.001	−0.007	−0.001	−0.008	0.386	0.347	0.375	0.339	0.335	0.922	0.937	0.926	0.935	0.942
	Trait4	10	−0.005	−0.001	−0.006	−0.005	−0.007	0.485	0.437	0.472	0.425	0.414	0.865	0.894	0.875	0.894	0.908
		15	−0.001	−0.005	−0.008	−0.003	−0.009	0.420	0.380	0.407	0.365	0.355	0.902	0.922	0.909	0.924	0.931
		20	−0.001	−0.005	−0.009	−0.003	−0.005	0.390	0.352	0.371	0.338	0.333	0.917	0.934	0.925	0.936	0.942
	Trait5	10	−0.001	−0.004	0.001	−0.004	−0.009	0.487	0.445	0.476	0.420	0.428	0.872	0.896	0.875	0.906	0.902
		15	0.001	−0.006	−0.001	−0.002	−0.006	0.426	0.385	0.409	0.366	0.369	0.905	0.924	0.912	0.930	0.929
		20	−0.002	−0.006	−0.006	−0.001	−0.003	0.394	0.354	0.370	0.337	0.343	0.920	0.937	0.930	0.941	0.939
	Mean	10	−0.004	0.001	−0.006	−0.002	−0.007	0.485	0.446	0.483	0.427	0.427	0.870	0.893	0.871	0.896	0.904
		15	−0.003	−0.003	−0.006	−0.002	−0.007	0.423	0.385	0.414	0.370	0.368	0.904	0.922	0.908	0.924	0.929
		20	−0.004	−0.003	−0.008	−0.002	0.005	0.392	0.354	0.376	0.341	0.340	0.919	0.935	0.925	0.936	0.940
.5	Trait1	10	0.002	−0.002	−0.006	−0.001	0.004	0.549	0.501	0.530	0.484	0.478	0.839	0.869	0.848	0.878	0.878
		15	−0.003	0.001	−0.002	0.004	0.001	0.472	0.425	0.454	0.410	0.408	0.885	0.910	0.893	0.915	0.913
		20	−0.001	0.001	0.002	0.002	0.003	0.433	0.384	0.407	0.370	0.373	0.905	0.927	0.916	0.931	0.928
	Trait2	10	0.008	0.001	−0.012	0.003	−0.004	0.597	0.548	0.586	0.505	0.496	0.802	0.838	0.808	0.866	0.868
		15	0.009	0.004	−0.003	0.002	−0.001	0.501	0.447	0.491	0.422	0.417	0.868	0.897	0.872	0.909	0.909
		20	0.010	0.002	0.003	−0.001	−0.001	0.451	0.399	0.433	0.377	0.378	0.895	0.919	0.903	0.929	0.926
	Trait3	10	0.002	−0.004	−0.017	−0.002	−0.008	0.560	0.514	0.567	0.483	0.487	0.823	0.857	0.821	0.876	0.873
		15	0.002	−0.006	−0.013	−0.002	−0.002	0.477	0.427	0.462	0.406	0.411	0.881	0.907	0.888	0.916	0.912
		20	0.002	−0.005	−0.007	−0.001	−0.0027	0.438	0.382	0.411	0.368	0.373	0.902	0.926	0.914	0.932	0.928
	Trait4	10	−0.006	−0.010	−0.017	−0.005	−0.007	0.564	0.508	0.546	0.484	0.466	0.823	0.860	0.833	0.875	0.884
		15	−0.004	−0.007	−0.010	−0.003	−0.005	0.481	0.424	0.464	0.406	0.393	0.877	0.906	0.884	0.915	0.920
		20	−0.003	−0.006	−0.006	−0.002	−0.003	0.438	0.382	0.412	0.371	0.357	0.901	0.925	0.911	0.929	0.934
	Trait5	10	−0.004	−0.008	−0.020	−0.001	−0.006	0.559	0.495	0.529	0.466	0.486	0.824	0.868	0.842	0.887	0.875
		15	−0.003	−0.006	−0.015	−0.001	−0.004	0.482	0.421	0.450	0.398	0.404	0.875	0.908	0.891	0.919	0.916
		20	−0.004	−0.005	−0.011	0.001	−0.002	0.440	0.382	0.403	0.363	0.365	0.898	0.926	0.916	0.933	0.932
	Mean	10	0.001	−0.005	−0.014	−0.001	−0.004	0.566	0.513	0.552	0.484	0.483	0.822	0.858	0.830	0.876	0.876
		15	0.001	−0.003	−0.009	0.001	−0.002	0.483	0.429	0.464	0.408	0.407	0.877	0.906	0.886	0.915	0.914
		20	0.001	−0.003	−0.004	0.001	−0.001	0.440	0.386	0.413	0.370	0.369	0.900	0.925	0.912	0.931	0.930

r = inter-trait correlations, *Test length* = administered MFC items, *mean* = average values of traits, *RMSE* = root mean square error, *CORR* = correlation between generated and estimated traits

MFC-K^B and MFC-KLP were higher than the other item selection methods. Similar to the three-dimensional study, in the case of five dimensions, MFC-K^B and MFC-KLP have similar performance. Among the five item selection methods, those two item selection methods have a higher estimation accuracy and a greater accuracy improvement over the other item selection methods. Moreover, the estimation accuracy of MFC-KI slightly improved, while MFC-A-optimality performed the worst. In conclusion, under both the three-dimensional and five-dimensional conditions, the proposed MFC-K^B and MFC-KLP methods not only had high estimation accuracy, but also were notably better than the existing FI-based item selection methods, while MFC-KI did not perform as well as the others.

The influence of the inter-trait correlations on trait estimation of item selection methods varied by the level of correlations. Other factors held constant, the trait estimation accuracy of the five methods decreases as the inter-trait correlations increase, which is consistent with study 1. Moreover, this performance pattern was more obvious in the five-dimensional conditions. For example, in the conditions in which the dimension correlation was 0 (see Table 4), the average RMSEs for MFC-K^B ranged from 0.341 to 0.427, versus 0.370 to 0.484 in which the inter-trait correlation was set to 0.5. The same pattern was consistently observed from other indices as well.

The test length also has a non-negligible impact on the estimation accuracy of methods in five-dimensional simulation. As expected, as the length of the MFC-CAT test increases, the estimation accuracy of all methods gradually improves. For example, in the conditions in which the inter-trait correlation was 0, the average RMSEs of all methods with 20-block tests ranged from 0.340 to 0.392, versus 0.427 to 0.485 for 10-block tests. This may be because the more blocks administrated in the tests, the more information was provided. Compared with the three-dimensional MFC-CAT, this trend was more notable in five-dimensional tests. When the test length increases from 10 blocks to 15 blocks, or from 15 blocks to 20 blocks, the estimation accuracy of each method significantly improved.

To confirm that our observed result patterns are also statistically significant, we performed a three-way factorial ANOVA on the RMSE outcomes, and the results are presented in Table 5. Although the two-way interactions are significant, based on Keppel and Wickens (2004), because these interaction effects are all noticeably smaller than the main effects as indicated by the smaller F values, it is meaningful to interpret the main effects as reflecting the general trends in the data. The main effect of the item selection method on RMSE was significant ($F(4, 180) = 297.3$, $p < .001$, $\eta^2 = 0.888$). Multiple comparisons revealed that the KL-based methods evoked smaller RMSE than those of the FI-based methods (all $p < .001$). The main effect of the correlation between traits on RMSE was significant ($F(1, 180) = 1005.492$, $p < .001$, $\eta^2 = 0.870$). Multiple comparisons revealed that the 0 inter-trait correlation evoked smaller RMSE than those in the 0.5 inter-trait correlation condition (all $p < .001$). The main effect of the test length on RMSE was significant ($F(2, 180) = 1694.602$, $p < .001$, $\eta^2 = 0.958$). Multiple comparisons revealed that the 10 block and 15 block conditions evoked smaller RMSE than those in the 5 block condition (all $p < .001$).

Uniformity of item pool usage The x^2 values of each method are shown in Table 6. In the five-dimensional MFC-CAT, except for MFC-KI, the x^2 values of all methods were relatively small. MFC-A-optimality had the most uniform exposure, while it had the lowest estimation accuracy. For the FI-based item selection methods, the higher the estimation accuracy was, the more uneven the utilization of the item pool. Among the KL-based item selection methods, MFC-KI has a relatively uneven item pool usage, while MFC-K^B and MFC-KLP had more even item pool usage. On the whole, the results indicated that the uniformity of item pool usage of the proposed KL-based item selection methods also better performed in five-dimensional study.

A simulation based on real data

The first two simulation studies provide evidence for the feasibility and effectiveness of the proposed KL-based item selection methods to measure various numbers of dimensions. The third simulation evaluates the proposed methods in real testing situations. This study used the Big-Five factor marker questionnaire with forced-choice items (Bunji & Okada, 2020), which measures five traits with 25 blocks, each block containing two statements measuring different traits. Based on the response data from 499 subjects provided by Bunji and Okada (2020), the Markov Chain Monte Carlo (MCMC) method was used to estimate the correlation matrix and item parameters (see Table 7), which were used as the true and generating correlation matrix and item parameters in this simulation. The real data can be found at https://osf.io/x92a3/.

For this study, five trait dimensions were measured, and the test length was fixed to 10, 15 and 20 blocks. A total of 1000 true latent trait vectors were randomly generated from a multivariate standard normal distribution with the correlation matrices of the NEO-PIR shown in Table 7. In sum, there were 5 (item selection method: MFC-A-optimality, MFC-D-optimality, MFC-KI, MFC-K^B, and MFC-KLP) × 3 (test length: 10, 15, 20) = 15 simulation conditions. For each condition, 20 replications were conducted. EAP estimation and Gauss-Hermite numerical integration were utilized for trait estimation with the R program.

Table 5 Main effects of item selection method, inter-trait correlation, and test length on RMSE

Independent variables	*SS*	*df*	*MS*	*F*
Item selection method	0.128	4	0.032	297.3***
Inter-trait correlation	0.108	1	0.108	1005.492***
Test length	0.365	2	0.182	1694.602***
Item selection method * Inter-trait correlation	0.003	4	0.001	6.675***
Item selection method * Test length	0.002	8	0.001	2.412*
Inter-trait correlation * Test length	0.007	2	0.004	34.351***
Item selection method * Inter-trait correlation * Test length	0.001	8	0.001	0.085

*$p < .05$, ** $p < .01$, ***$p < .001$

For the trait estimation accuracy evaluation, the RMSEs of each dimension are presented next (BIAS and CORR are omitted for this study as previous studies revealed similar patterns as RMSE). For item exposure, the x^2 index was computed.

Results

Table 8 summarizes the RMSEs and x^2 values of study 3. It is evident that the estimation accuracy and uniformity of item pool usage of five item selection methods were acceptable in real testing situations. Compared with the five-dimensional MFC-CAT simulation in study 2, the estimated RMSEs of each method were relatively high. This may be because the quality of blocks in the item pool was relatively low, and the inter-trait correlations in the real correlation matrix were relatively high. The performance pattern of five methods in real testing situations was similar with that in the previous two simulation studies. For example, the average RMSEs of MFC-K^B ranged from 0.723 to 0.772, which performed better than the FI-based methods. As shown in Table 8, MFC-K^B yielded the smallest RMSEs, while MFC-A-optimality produced the largest RMSEs. In general, the estimation accuracies of the KL-based item selection methods exceed that of the FI-based item selection methods in real testing situations.

The performance pattern of the five methods in terms of uniformity of item pool usage was also similar to the first two simulation studies. Among the five methods, the item pool usage of the KL-based item selection methods is relatively even with lower x^2 values, which outperformed the FI-based item selection methods. However, MFC-KI still had the worst performance.

In summary, from the perspective of the estimation accuracy and uniformity of item pool usage, MFC-K^B performed the best and the proposed KL-based item selection methods generally outperformed the existing FI-based item selection methods under the circumstance of the practical NEO-PIR item pool.

Summary and discussion

MFC-CAT is a promising new research area that has gained more and more attention given that it integrates MFC personality assessment with CAT. Compared with traditional tests, MFC-CAT not only greatly reduces test time, but also eliminates response bias, thus improving test efficiency and estimation accuracy. Currently, studies on MFC-CATs were mainly focused on the FI-based item selection methods using the GGUM-RANK model (e.g., Joo et al., 2020). However, studies found that the KL-based item selection methods can be an alternative to address the issue of attenuation paradox of FI-based item selection methods (Chang & Ying, 1996; Veldkamp & van der Linden, 2002). Moreover, the TIRT model is a promising alternative model for MFC-CAT as it was widely used to model a variety of forced-choice scales and has demonstrated efficacy in accommodating many combinations of traits and block sizes (Brown & Maydeu-Olivares, 2011, 2013).

Table 6 The x^2 values of the five compared item selection methods for five-dimensional MFC-CAT

r	Indices	Test length	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP
0	x^2	10	26.849	31.434	45.879	31.472	29.480
		15	22.064	27.026	46.734	30.367	26.238
		20	19.355	23.880	46.492	27.191	23.518
0.5	x^2	10	32.705	36.501	51.974	33.562	33.967
		15	29.540	33.161	53.380	34.836	31.082
		20	27.512	30.452	53.139	32.291	28.317

r = inter-trait correlations, *Test length* = number of administered MFC items

Table 7 The correlation matrices of the Big-Five factor marker questionnaire

Traits	*N*	*E*	*C*	*A*	*O*
N	1				
E	0.552	1			
C	0.371	0.526	1		
A	0.355	−0.209	−0.110	1	
O	0.476	0.616	0.498	−0.158	1

N = neuroticism; *E* =extraversion, *C* = conscientiousness, *A* = agreeableness, *O* = openness to experiences

Therefore, this study constructs the MFC-CAT procedures based on the TIRT model and proposes the MFC-KI, MFC-K^B, and MFC-KLP item selection methods based on the KL information for MFC-CAT. The results from three simulation studies confirmed that the proposed KL-based item selection methods outperformed the existing FI-based item selection methods, especially when the test is short (or equivalently, at an early stage of the CAT), generating greater trait estimation accuracies and utilization of the item pool. These findings are encouraging for applications of MFC-CAT to noncognitive personality evaluation in talent assessment.

More specifically, two Monte Carlo simulations and a simulation based on real data were conducted under three-dimensional, five-dimensional, and real testing settings. In these simulations, we manipulated several factors, including the number of dimensions, the inter-trait correlations, and the test length. The findings are summarized as the following.

First, the trait estimation accuracy and uniformity of item pool usage of all proposed item selection methods were acceptable. Among the five compared methods, the proposed MFC-K^B and MFC-KLP performed best and comparably in terms of estimation accuracy and uniformity of item pool usage. By using the posterior distribution, these two item selection methods extract more information from the respondents (Mulder & van der Linden, 2010; Veldkamp & van der Linden, 2002), resulting in more precise trait estimation than the other methods. Except for MFC-KI, which performed the worst among all five compared methods and resulted in lower trait estimation accuracy and relatively higher utilization of the item pool. It is consistent with previous studies in single-statement MCAT (e.g., Tu et al., 2018). This may be because MFC-KI prefers blocks with high discrimination parameters in both dimensions, while blocks with larger KI do not necessarily provide higher power to discriminate $\boldsymbol{\eta}$ from $\hat{\boldsymbol{\eta}}$. For example, a block j satisfying $\sum_{d=1}^{p} \alpha_{jd}\left(\hat{\eta}_d - \eta_d\right) = 0$ may has high KI, but it does not actually

Table 8 The results of the five compared item selection methods for MFC-CAT based on real data

Indices	Test length	MFC-A-optimality	MFC-D-optimality	MFC-KI	MFC-K^B	MFC-KLP
RMSE-trait1	10	0.831	0.748	0.725	0.720	0.817
	15	0.743	0.708	0.696	0.690	0.727
	20	0.707	0.685	0.676	0.679	0.688
RMSE-trait2	10	0.881	0.838	0.793	0.747	0.887
	15	0.825	0.736	0.700	0.707	0.762
	20	0.723	0.697	0.688	0.693	0.704
RMSE-trait3	10	0.873	0.873	0.884	0.875	0.925
	15	0.846	0.848	0.854	0.842	0.871
	20	0.823	0.829	0.836	0.821	0.831
RMSE-trait4	10	0.833	0.730	0.693	0.723	0.823
	15	0.730	0.692	0.678	0.694	0.716
	20	0.712	0.676	0.671	0.685	0.681
RMSE-trait5	10	0.982	0.858	0.845	0.796	0.856
	15	0.853	0.812	0.764	0.753	0.779
	20	0.811	0.752	0.743	0.739	0.752
Mean	10	0.880	0.809	0.788	0.772	0862
	15	0.799	0.759	0.738	0.737	0.771
	20	0.755	0.728	0.723	0.723	0.731
x^2	10	7.852	6.915	10.853	6.674	8.025
	15	3.795	4.404	7.910	4.045	3.833
	20	1.808	2.832	3.720	1.593	1.100

r = inter-trait correlations, *Test length* = number of administered MFC items, *mean* = average RMSE values of traits, *RMSE* = root mean square error, *CORR* = correlation between generated and estimated traits

provide discrimination power with respect to $\boldsymbol{\eta}$ and $\hat{\boldsymbol{\eta}}$ as $KL(\hat{\boldsymbol{\eta}} \parallel \boldsymbol{\eta}) = 0$ (Tu et al., 2018; Wang & Chang, 2011).

Second, the influence of the inter-trait correlations, test lengths, and dimensionality on various item selection methods for MFC-CAT was examined. We found that the lower the inter-trait correlations, the higher the estimation accuracy and the utilization of the item pool. These findings are consistent with similar studies (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019). The reason may be that, in forced-choice tests, as the correlation between the traits measured in each block increases, the uncertainty of the participants' responses increases, thus reducing the trait estimation accuracy. Similarly, consistent with the previous MFC-CAT studies (Bürkner et al., 2019; Joo et al., 2020), the more test items, the higher the estimation accuracy. From three to five dimensions, the performance pattern of the five MFC-CAT item selection methods as varying by inter-trait correlations and test lengths stays the same.

Lastly, a simulation based on real data was conducted to evaluate the proposed KL-based item selection methods in a practical setting. Results show that acceptable trait estimation accuracy (in terms of RMSEs) and acceptable uniformity of item pool usage (in terms of x^2 values) can also be rendered in a practical application of the proposed methods in MFC-CAT.

In sum, simulation results show that the proposed KL-based item selection methods are all viable to the MFC-CAT, and MFC-K^B and MFC-KLP are the best choices recommended.

The simulation studies conducted in this research are by no means exhaustive. This article represents a crucial step in the research of MFC-CAT by exploring CAT procedures and item selection methods applicable to forced-choice items based on the TIRT model. For future studies, it is interesting to investigate other adaptive methods for MFC-CAT. The item selection methods used in this paper are extended from the single-statement MCAT. New and more efficient methods and algorithms may be explored for MFC-CAT. To make MFC-CAT more applicable in real work contexts, it is necessary to discuss the nonstatistical factors, such as item exposure control, content constraints, and so on. Moreover, in order to further verify the practical applicability of the proposed methods, real empirical research is needed. Last but not least, while the MFC-CAT simulations in this study are fixed-length tests, future research can be conducted to explore termination strategies in variable-length MFC-CAT, which may further shorten the test length and improve the efficiency and fairness of the test.

Funding The work was supported by the National Natural Science Foundation of China (62167004, 32160203, 31960186, and 61967009).

References

Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement, 19*, 269–290. https://doi.org/10.1177/014662169501900306

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Borman, W. C., Buck, D. E., Hanson, M., Motowidlo, S. J., Stark, S., & Drasgow, F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using Computerized Adaptive Rating Scales. *Journal of Applied Psychology, 86*, 965–973. https://doi.org/10.1037/0021-9010.86.5.965

Brown, A. (2010). How IRT can solve problems of ipsative data (Doctoral dissertation). University of Barcelona, Spain. Retrieved from http://hdl.handle.net/10803/80006

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

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52. https://doi.org/10.1037/a0030641

Bunji, K., & Okada, K. (2020). Joint modeling of the two-alternative multidimensional forced-choice personality measurement and its response time by a Thurstonian D-diffusion item response model. *Behavior Research Methods, 52*(3), 1091–1107. https://doi.org/10.3758/s13428-019-01302-5

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 5*, 827–854. https://doi.org/10.1177/0013164419832063

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229. https://doi.org/10.1177/014662169602000303

Chang, H. H., & Ying, Z. (1999). A stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222. https://doi.org/10.1177/01466219922031338

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129–145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Cheung, M. W., & 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

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 selection and classification decisions (Tech. Rep. No. 1311)*. U.S. Army Research Institute for the Behavioral and Social Sciences.

Dueber, D. M., Love, A. M. A., Toland, M. D., & Turner, T. A. (2019). Comparison of Single-Response Format and Forced-Choice Format Instruments Using Thurstonian Item Response Theory. *Educational and Psychological Measurement, 79*(1), 108–128. https://doi.org/10.1177/0013164417752782

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement, 46*(3), 84–103. https://doi.org/10.1111/j.1745-3984.2009.01070.x

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In *Objective measurement: Theory into practice* (Vol. 1, pp. 236–258). Ablex.

Gordon, L. V. (1976). *Survey of interpersonal values (Revised manual)*. Science Research Associates.

Hewitt, A. (2015). *2015 Trends in global employee engagement report*. Aon Corp.

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. https://doi.org/10.1177/0146621615585851

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*, 357–372. https://doi.org/10.1111/jedm.12183

Joo, S. H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods, 52*(2), 761–772. https://doi.org/10.3758/s13428-019-01274-6

Keppel, G., & Wickens, T. D. (2004). Simultaneous comparisons and the control of type I errors. In *Design and analysis: A researcher's handbook* (4th ed). Pearson Prentice Hall.

Lima Passos, V., Berger, M. P. F., & Tan, F. E. (2007). The D-optimality item selection criterion in the early stage of CAT: A study with the graded response model. *Journal of Educational and Behavioral Statistics, 33*(1), 88–110. https://doi.org/10.3102/1076998607302631

Lin, H. (2012). Item selection methods in multidimensional computerized adaptive testing adopting polytomously-scored items under multidimensional generalized partial credit model. *Dissertations & Theses - Gradworks, 5*(4), 392–403.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley. https://doi.org/10.1037/013774

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*, 935–974. https://doi.org/10.1080/00273171.2010.531231

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354–368. https://doi.org/10.1037/1082-989X.9.3.354

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. *Applied Psychological Measurement, 40*(7), 500–516. https://doi.org/10.1177/0146621616662226

Morrison, E. W., & Bies, R. J. (1991). Impression management in the feedback-seeking process: A literature review and research agenda. *The Academy of Management Review, 16*(3), 522–541. https://doi.org/10.2307/258916

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*, 273–296. https://doi.org/10.1007/s11336-008-9097-5

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing, statistics for social and behavioral sciences* (pp. 77–101). Springer. https://doi.org/10.1007/978-0-387-85461-8

Qiu, X.-L., & Wang, W.-C. (2016). Item response theory models for ipsative tests with polytomous multidimensional forced-choice items. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45–58. https://doi.org/10.1177/014662169001400105

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354. https://doi.org/10.1007/BF02294343

SHL. (1997). *Customer Contact: Manual and User's Guide*. SHL Group.

SHL. (2006). *OPQ32 Technical Manual*. SHL Group.

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

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*(1), 25–39. https://doi.org/10.1037/0021-9010.91.1.25

Stark, S., Chernyshenko, O. S., & Guenole, N. (2011). Can subject matter experts' ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scales? *Organizational Research Methods, 14*, 256–278. https://doi.org/10.1177/1094428109356712

Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive Testing with Multidimensional Pairwise Preference Items. *Organizational Research Methods, 15*(3), 463–487. https://doi.org/10.1177/1094428112444611

Stark, S., Chernyshenko, O. S., Drasgow, F., White, L. A., Heffner, T., Nye, C. D., & 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. https://doi.org/10.1037/mil0000044

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 79*, 281–299.

Tu, D. B., Han, Y. T., Cai, Y., & Gao, X. L. (2018). Item Selection Methods in Multidimensional Computerized Adaptive Testing with Polytomous Scored Items. *Applied Psychological Measurement, 8*, 677–694. https://doi.org/10.1177/0146621618762748

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398–412. https://doi.org/10.3102/10769986024004398

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575–588. https://doi.org/10.1007/BF02295132

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates Inc.. https://doi.org/10.1023/A:1016834001219

Walton, K. E., Cherkasova, L., & Roberts, R. D. (2020). On the validity of forced choice scores derived from the Thurstonian item response theory model. *Assessment, 4*(27), 706–718. https://doi.org/10.1177/1073191119843585

Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing gaining information from different angles. *Psychometrika, 76*, 363–384. https://doi.org/10.1007/s11336-011-9215-7

Wang, C., Chang, H. H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika, 76*, 13–39. https://doi.org/10.1007/s11336-010-9186-0

Wang, W. C., Qiu, X. L., Chen, C. W., Ro, S., & Jin, K. Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement, 41*, 600–613. https://doi.org/10.1177/0146621617703183

Zinnes, J. L., & Griggs, R. A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika, 39*, 327–350. https://doi.org/10.1007/BF02291707

Open practice statement The codes of the proposed MFC-KI, MFC-KB, and MFC-KLP methods have been uploaded to https://osf.io/bmg8r/. The real data used in this article could be also found at https://osf.io/x92a3/.