Heliyon 10 (2024) e26884

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Review article

Multidimensional IRT for forced choice tests: A literature review

Lei Nie [a], Peiyi Xu [b,*], Di Hu [c]

[a] School of Public Administration, East China Normal University, China
[b] Department of Educational Psychology, Faculty of Education, East China Normal University, China
[c] School of Education and Social Policy, Northwestern University, USA

ARTICLE INFO

Keywords:
Multidimensional forced choice test
Ipsative data
TIRT
MUPP
Parameter estimation

ABSTRACT

The Multidimensional Forced Choice (MFC) test is frequently utilized in non-cognitive evaluations because of its effectiveness in reducing response bias commonly associated with the conventional Likert scale. Nonetheless, it is critical to recognize that the MFC test generates ipsative data, a type of measurement that has been criticized due to its limited applicability for comparing individuals. Multidimensional item response theory (MIRT) models have recently sparked renewed interest among academics and professionals. This is largely due to the development of several models that make it easier to collect normative data from forced-choice tests. The paper introduces a modeling framework made up of three key components: response format, measurement model, and decision theory. Under this paradigm, four IRT models were chosen as examples. Following that, a comprehensive study is carried out to compare and characterize the parameter estimation techniques used in MFC-IRT models. This work then examines empirical research on the concept by analyzing three distinct domains: parameter invariance testing, computerized adaptive testing (CAT), and validity investigation. Finally, it is recommended that future research initiatives follow four distinct paths: modeling, parameter invariance testing, forced-choice CAT, and validity studies.

1. Introduction

Non-cognitive psychological tests frequently employ Likert rating scales such as organization (e.g., "I am organized"). There are no correct or incorrect answers to the items. Participants must identify the item that best fits their situation on a five-point Likert scale ranging from little likeness (1) to maximum resemblance (5). People may purposefully distort their answers in important assessments such as employment and selection, particularly those related to traits such as responsibility and optimism. This strategic behavior attempts to appear more aligned with organizational expectations, even if it does not reflect the true nature of the organization. The phenomenon under investigation is "faking," which occurs when an assessment fails to separate people based on their abilities, thereby undermining its objectivity.

To reduce faking, pro-preventing or post-detecting are used [1]. To avoid mistaking an honest person as a fraud, post-control approaches must provide high recognition accuracy. The post-detecting approaches prevents cheating before or during questioning in order to obtain pollution-free data. There are warnings, bogus pipeline, and forced-choice test. Individual faking is unaffected by warnings, and bogus pipeline defrauds people, which is unethical [2]. The forced-choice test requires participants to select items of comparable quality. They cannot provide good options for all objects. Because the products' desirability is equal and none is superior,

[*] Corresponding author.
E-mail address: xpyecnu@gmail.com (P. Xu).

https://doi.org/10.1016/j.heliyon.2024.e26884
Received 21 September 2023; Received in revised form 11 February 2024; Accepted 21 February 2024
Available online 22 February 2024

social desirability is less likely to influence selection or faking. Researchers have explored the use of items involving several statements that are similar in social desirability but represent different dimensions [3].

However, traditional forced-choice test scoring will yield ipsative data. Dimension scores are interdependent in the forced-choice test. All dimensions will have different high and low scores. This information is ipsative. Internal score dependence of ipsative data violates one of classical test theory's basic assumptions, the independence of error variance, which affects statistical analysis and interpretation of forced-choice test scores [4,5], such as reliability, variance, and regression analysis, and increases the probability of type I error. It also has an impact on statistical test power [6]. The ipsative data's distortion of the dimension relationship pollutes the test's construct validity and criterion-related validity [7] and prevents it from being used for factor analysis [8]. Finally, comparing individuals and normalizing ipsative data by scores may affect their veracity. Self-comparison, for example, only displays interest test participants' preference rankings. According to Closs [8], direct comparisons will overvalue or undervalue individual interests.

The number of dimensions and their interrelationships have a great impact on ipsative data. More test dimensions, according to previous research [4,9,10], narrow the gap between ipsative and normality scores. When dimensions are positively or negatively correlated, the difference between the ipsative score and the normality score decreases [11]. As a result, increasing the number of test dimensions is one of the more effective traditional methods for resisting ipsative data, but it is only a compromise.

In conclusion, the ipsative data limits the application of the forced-choice test; while it can resist the ipsative problem by adding dimension methods, it does not reflect the individual's psychological decision process. To address the issue of ipsative data, it is necessary to abandon the traditional scoring method and adopt a modern measurement model to reflect individuals' decision process when answering forced-choice tests [7] and obtain the latent trait scores underlying the decision process from the explicit comparison results. Individual score normalcy can thus be restored.

The primary goal of this work is to provide a comprehensive and structured overview of the MIRT model for forced-choice tests. We introduce forced-choice IRT models and provides a concise summary of their three basic components. We also discuss the parameter estimation techniques used in forced-choice MIRT models. Following that, we demonstrate the application's current research progress. Finally, based on the practical implications of the forced-choice paradigm, this study suggests potential future research directions.

1.1. *Search methods*

In the Web of Science Core Collection database, the keyword "forced choice" yielded 135 papers. Following an examination of the abstracts, studies that did not include the forced choice model and only used forced choice tests as empirical experiments were excluded, leaving 45 papers to complete the paper search for this literature review. Then, in order to supplement the literature, we searched for the keyword "forced choice" in mainstream journals in the field of forced choice. Such as Educational and Psychological Measurement, Applied Psychological Measurement, Multivariate Behavioral Research, Journal of Educational, Behavioral Statistics and so on. Finally, we traced the sources of existing literature and included 97 references in total.

2. IRT model for multidimensional forced choice test

2.1. *Three key elements of multidimensional forced choice models*

Various MIRT-based scoring models for forced-choice tests have been developed over the last decade. These models relate explicit responses to underlying features in order to obtain latent trait scores with normal characteristics and to compare scores across individuals. These models are comprised of three major components: response format, measurement model, and decision theory. The response format reflects the format of the forced-choice response data, the measurement model reflects the relationship between item response intensity and dimensions, and the decision theory reflects the process by which participants choose between items. The decision theory acts as a link between the explicit response and the favorability of the items, and is then linked to the personal latent trait level by the measurement model, forming an overall forced-choice IRT model.

2.1.1. *Response format*

The forced-choice test typically consists of a number of item blocks of varying dimensions. The item block is made up of a fixed number of statements with different or identical dimensions and social desirability levels. The statements are explicit indicators of the dimension (i.e., latent trait).

According to Hontangas et al. [12], there are three common forms of forced-choice item blocks: Pick, Rank and MOLE. This classification is mainly reflected in the types of instructions. Pick (Table 1) requires individuals to choose the item that best matches them. Rank (Table 2) requires individuals to fully rank the items from most agreeable to least agreeable. MOLE (Table 3) requires individuals to choose the item that best fits themselves MOst and the item that is LEast suited to themselves. These three are equivalent

Table 1
Pick question types.

Instruction: Choose the one that best suits you from the following two descriptions	
Item block	Most
A Lack of finding things	✓
B Explore unfamiliar territory	

in a pair, while Pick and MOLE are equivalent in a triple.

The number of statements contained in an item block determines its size, with two to four questions being the most common. The size of the item block influences the individual's load on the selection task. The more items there are, the more times the individual must compare them. The cognitive complexity of the selection task is increased by a large item block. It may be harmful to people with limited education or poor reading skills [13].

For anti-fraud efficacy, the alignment of item desirability is the most important factor in forced-choice test construction, followed by explicit factors such as item size and instructions. In general, the matching degree is calculated by calculating the average absolute difference in desirability between items. The greater the disparity, the more mismatched it is. However, judging only by the mean ignores differences in the desirability evaluations of the same item by different evaluators. Pavlov et al. [14] proposed an alternative index, the IIA (Inter-item Agreement) index, which incorporated the BP and AC indexes [15] into the matching of desirability of items to better match those items with no difference in mean value of desirability. Practitioners can calculate the IIA index and automatically compose the paper using the R [16] package autoFC [17].

The consistency of MFC item blocks in the ideal rank suggests that if item desirability is not well matched, the forced-choice test lacks an anti-fraud effect [18]. Well-matched blocks, on the other hand, produce more uniform response data. As a result, response data can be used to forecast item block faking ability in order to create a Faking Mixture Model [19].

2.1.2. Measurement model

The item is an explicit measure of the trait, and the item's relationship to the latent trait must be linked using a measurement model. Dominance Models and Unfolding Models (or Ideal-Point Models) are the two types of measurement models. According to dominance models, an individual's characteristic level increases their likelihood of responding yes to the item. The Rasch model and the Two-Parameter Logistic Model (2PLM) all assume that the individual answers the item in accordance with the dominance measurement model. The unfolding models assume that the item's proximity to the characteristic level being evaluated increases the likelihood of a positive response. Individuals who are too introverted, for example, may disagree with the item "I enjoy chatting quietly with a friend in a café" because they are uncomfortable in public places, whereas individuals who are extremely extroverted may disagree because they prefer more exciting settings [20]. Individuals at the intermediate level are more likely to agree with the item, and their item response function curve is single-peaked and bell-shaped; that is, the higher the probability of a positive answer, the closer the individual's trait level is to the item position. The Generalized Graded Unfolding Model (GGUM) is the unfolding model's representative model [21].

There is disagreement in the literature about whether models are better at reflecting individuals' responses to non-cognitive items [6,22,23]. Simulation and empirical studies, such as those conducted by Chernyshenko et al. [24] and Tay et al. [25], have aided in the unfolding model. According to these studies, unfolding response items are just as effective as dominance response items in assessing attitude qualities. The unfolding model is more flexible since it might be similar to the dominance model when the item's positional parameter is in its final place. However, studies have shown that this superiority is not universal in practice, and the psychometric properties of scales made entirely of unfolding response items are significantly inferior to scales made entirely of dominance response items, including lower reliability and criterion correlations [26]. Furthermore, the unfolding model cannot directly convert the scoring of reverse items [27]. The dominance model is generally more parsimonious and has fewer parameters than the unfolding model in terms of model complexity. Except where there is clear evidence to examine the superiority of the complex model [28], the more parsimonious model should be considered first. Furthermore, writing unfolding response items is more difficult, and defining the exact meaning reflected by the items is also difficult. More information on the dominance model and the unfolding model can be found in the work of Drasgow et al. [20].

The measurement model is an item-level feature and has nothing to do with the format of forced-choice items. Items from any measurement model can be used when combining items into forced-choice item blocks because they can all measure the same latent trait and the distribution of latent trait is constant for the same population. In practice, researchers must combine the characteristics of the item or the data to choose one of the dominance or unfolding models as the measurement model between the item and the latent trait, and there is currently no situation in which the two models are mixed in the same test.

2.1.3. Decision theory

Instead of evaluating each item independently, forced-choice tests require participants to make comparative judgments on a group of items and then make decisions on how to answer them. The absolute evaluation of the items serves as the foundation for determining an individual's trait level. According to Brown [13], the basis for individuals making comparative judgments on a set of items is their absolute evaluation level of each item being compared. To model forced-choice data, decision theory must explain the relationship between explicit response and absolute evaluation, allowing the individual's latent trait level to be assessed.

Table 2
Rank question types.

Instruction: Sort the following descriptions	
Item block	Rank
A Lack of finding things	3
B Explore unfamiliar territory	1
C Make decisions based on data analysis	2

Table 3
MOLE question types.

Instruction: Choose from the following descriptions the one that best fits you and the one that doesn’t fit you the most		
Item block	Most	Least
A Lack of finding things		
B Explore unfamiliar territory	✓	
C Make decisions based on data analysis		
D Do work that focuses on precision		✓

Thurston’s Law of Comparative Judgment

Utility is a latent variable that can be thought of as the psychological value of an item to an individual. Thurstone [29] believed that the individual’s consideration of the item was essentially a consideration of utility value. y_{ij} represent the explicit results after comparing the item i and j, and $y_{ij} = 1$ represent that individual selected the item i as the most consistent, otherwise $y_{ij} = 0$. The relationship between utility difference y_{ij}^* and explicit response y_{ij} can be sorted as formula (1):

$$y_{ij} = \begin{cases} 1, y_{ij}^* \geq 0 \\ 0, y_{ij}^* < 0 \end{cases} \tag{1}$$

Where $y_{ij}^* = t_i - t_j$ represent the utility difference between item i and item j. t represent the utility value of item.

The utility on item i can be divided into two parts: systematic and random. The systematic part $f(\theta_a)$ can be a response function related to the individual’s latent trait level, and the random part is random error ε_i. Thurstone assumed that they are independent of each other among different items and obey a normal distribution. Therefore, the relationship between utility and latent trait can be expressed as formula (2):

$$t_i = f(\theta_a) + \varepsilon_i \tag{2}$$

Where θ_a is the individual’s level on the latent trait i measured by the item a.

Luce’s Choice Axiom

Luce [30,31] extended the Bradley-Terry model [32] from binary choice situations, which used v_i to represent an individual item i-related response intensity. The set of alternative items is called S, then the probability $P(i[S])$ of choosing i from S is proportional to v_i as formula (3):

$$P(i[S]) = \frac{v_i}{\sum\limits_{k \ in \ S} v_k} \tag{3}$$

Luce describes the ranking process of a group of items as a series of independent steps to make the best choice: firstly, select the most suitable item i from the item set S, and then select the second most suitable item j from the remaining set $S - 1$, until the selection of the last two items is completed, thus realizing the ranking of all alternative items [12]. The probability of the ranking result is the result of the multiplication of the probability of each step.

When this decision theory is applied to the forced-choice model, v_i can be derived from the item response function related to the latent trait. The MUPP framework proposed by Stark [33] extends Luce’s Choice Axiom, which greatly promoted the development of forced-choice models [7]. In the MUPP framework, it is assumed that the individual’s assessment of each item is independent and that the items are unidimensional. Items in a block can come from the same or different dimensions, so it is called the Multi-Unidimensional Pairwise Preference Model (MUPP). Assuming an item block contains the item i, j, k, l, and the latent trait $\theta_a, \theta_b, \theta_c, \theta_d$ are measured separately. $P(i)$ represents the individual’s probability of accepting the item i and $Q(i)$ represents the probability of rejecting the item i, and $Q(i) = 1 - P(i)$. Put them into formula (3), $v_i = P(i)Q(j)Q(k)Q(l)$.

When it is Pick item format, the probability $P(i[ijkl])$ of an individual choosing i from the set $[ijkl]$ can be expressed as formula (4):

$$P(i[ijkl]) = \frac{P(i)Q(j)Q(k)Q(l)}{P(i)Q(j)Q(k)Q(l) + Q(i)P(j)Q(k)Q(l) + Q(i)Q(j)P(k)Q(l) + Q(i)Q(j)Q(k)P(l)} \tag{4}$$

Taking the Rank item format as an example, assuming that the ranking result of an individual is $i > j > k > l$, then $P(ijkl)$ is like formula (5):

$$P(ijkl) = P(i[ijkl]) \times P(j[jkl]) \times P(k[kl]) \tag{5}$$

Taking MOLE as an example, the rank of the two unselected items cannot be determined, so the two possible ranks are combined as the probability of the selection result of this item format. Using $P(i**l)$ to represent the probability that the subjects chose i and l as the most and the least in line with their probability, then we can get formula (6):

$$P(i**l) = P(ijkl) + P(ikjl) \tag{6}$$

In conclusion, determining the response probability of an individual item will allow us to determine the response probability of a block.

In addition, Thurston's Law of Comparative Judgment and Luce's choice axiom are equivalent in a pair.

Other types of decision theories include Coombs's Unfolding Preference Model and Andrich's Forced Endorsement Model. The former is a special case of Thurstone's Law of Comparative Judgment, and the latter is equivalent to the Bradley-Terry model after simplification.

2.2. IRT models for forced-choice test

Table 4 presents a concise overview of the prevailing forced-choice models [23,33–36] based on the three fundamental components of FC models. Due to the equivalence of decision theory under pairs, pair will be listed separately. This section elucidates the underlying distinctions among models by presenting a comprehensive overview of four specific examples, namely TIRT, MUPP-2PL, ZG-MUPP, and MUPP-GGUM.

2.2.1. TIRT model

Brown and Maydeu-Olivares [34] proposed TIRT, a MIRT model for dominance response items based on Thurstone's Law of Comparative Judgment.

TIRT assumes that the psychological process of individual selection or ranking is to make independent pairwise comparison judgments on n items in an item block in turn, and produces $\widetilde{n} = n(n-1)/2$ comparison results. Before modeling the data, binary coding the response to obtain the comparison results of the pairwise items are needed.

Taking a Rank-3 item block as an example, the items in the item block are $\{i,j,k\}$, i, j and k respectively represent a project and measure the latent feature s of an independent dimension. Assuming that the individual's selection result is $i > k > j$, the encoding result is $\{i,j\} = 1$, $\{i,k\} = 1$, $\{j,k\} = 0$, which represents $i > j$, $i > k$ and $j < k$. Taking $i > j$ as an example, $P(i > j)$ is like formula (7) :

$$P(i > j|\theta_a, \theta_b) = \Phi_N\left(\frac{\gamma_{ij} + \lambda_i\theta_a - \lambda_j\theta_b}{\sqrt{\psi_i^2 + \psi_j^2}}\right) \tag{7}$$

Where $\mu_i - \mu_j = \gamma_{ij}$, μ_i is the mean of the latent utility t_i, λ_i is the factor loading of the item i on the latent trait θ_a, assumming latent trait and error obey the normal distribution. The variances of error ε_i and ε_j are ψ_i^2, ψ_j^2, then the variance of the difference value is $\psi_i^2 + \psi_j^2$, and Φ_N represents the cumulative normal distribution function.

TIRT's applicability in various settings has been tested using simulations and empirical studies by many scholars [5,7,37–41]. On the one hand, these studies indicated that TIRT has overcome the ipsative issue in conventional scoring to some extent, has improved measurement accuracy compared to conventional scoring, and is closer to the results of the Likert single stimulus scale [42]; on the other hand, they also indicated that in order to show better properties than conventional scoring, TIRT requires more restrictions on the test design [34].

2.2.2. MUPP-GGUM model

The MUPP-GGUM model, proposed by Stark [33], is a multidimensional model for unfolding response items that is based on Luce's Choice Axiom. As the first forced-choice model used in computerized adaptive testing, it has been extensively used in the development of numerous personality assessments for the purpose of military personnel selection in the United States. It also provides consistent guidance for the development process.

Stark [33] used the binary scoring version of GGUM, which follows the dominance response model, to calculate the response probability of a single item, that is, the $P(i)$ and $Q(i)$ in formula (4). Hontangas et al. [22] developed MUPP-GGUM model suitable for the Rank and MOLE item formats and used the MCMC joint estimation algorithm to evaluate the statement and personal parameters. Based on the presented model, the likelihood of an individual selecting a particular item i can be determined as formula (8):

$$P(i) = \frac{\exp\{\alpha_i[(\theta_a - \delta_i) - \tau_i]\} + \exp\{\alpha_i[2(\theta_a - \delta_i) - \tau_i]\}}{1 + \exp\{\alpha_i[3(\theta_a - \delta_i)]\} + \exp\{\alpha_i[(\theta_a - \delta_i) - \tau_i]\} + \exp\{\alpha_i[2(\theta_a - \delta_i) - \tau_i]\}} \tag{8}$$

Among them, α_i represents the discrimination parameter of the item i, τ_i is the intercept parameter of the item i, δ_i is the location parameter of the item i, and θ_a represents the latent trait measured by item i.

For this model, Joo et al. [43] created two informative indices: OII (Overall Item Information) and OTI (Overall Test Information). When selecting items similar to OII, Joo et al. offered a method to draw graphs of conditional OII so that researchers can further

Table 4
Model summary.

		Pair	Pick	Rank	MOLE
Dominant model	Thurston's Law of Comparative Judgment	TIRT/RIM/MUPP-2PL	TIRT/BRB-IRT	TIRT/BRB-IRT	TIRT/BRB-IRT
	Luce's Choice Axiom		–	2PLM-RanK/ELIRT	–
Unfolding model	Thurston's Law of Comparative Judgment	ZG-MUPP/MUPP-GGUM	–	–	–
	Luce's Choice Axiom		–	GGUM-Rank/FCRM	–

compare and select the item block that provides the greatest amount of information within the target ability interval. The development of information indices also lays the groundwork for CAT [44].

2.2.3. MUPP-2PL model

According to Morillo et al. [24], the items used in the dominance measurement model were superior to those used in the unfolding measurement model in terms of item writing difficulty and model parsimony. Therefore, on the basis of the MUPP framework, Morillo et al. replaced the item response function calculated $P(i)$ and $Q(i)$ in formula (4) with the classical dominance response model 2PLM, and called it the MUPP-2PL model. According to this model, the probability of an individual choosing an item i is formula (9):

$$P(i) = P(i|\theta_a) = \frac{1}{1 + \exp[-(\alpha_i\theta_a - \beta_i)]} \tag{9}$$

Among them, α_i represents the discrimination parameter of the item i, β_i is the intercept parameter of the item i, and θ_a represents the latent trait measured by item i.

Morillo et al. [24] discovered that the length of the test influenced the recovery of relationships between item parameters, ability parameters, and traits; the longer the test, the more accurate the estimated results. Furthermore, sample size has a significant impact on parameter estimation accuracy, and this method can estimate difficult parameters more accurately than discrimination parameters. Finally, Morillo et al. discovered in an empirical study that MUPP-2PL's estimation results of the relationship between some latent traits were quite different from previous studies, but it was unclear whether the difference was due to the respondent population or a change in the test situation.

2.2.4. ZG-MUPP model

The ideal point model is used as the measurement model in the ZG-MUPP model [45]. The model extends the MUPP framework's decision theory from Luce's Choice Axiom to Thurstone's Law of Comparative Judgment.

Assuming the participant needs to choose between two items s and t, these items assess latent features in different dimensions. The ZG-MUPP model defines X_s as the latent feature variable of item s, and Z_s as the statement variable of item s. The latent feature variable follows a bivariate normal distribution, while the statement variables are independent. Individual choices between items s and t are made by comparing the distances between the latent features measured by the two items and the statement variables. If the distance $T_s = X_s - Z_s$ in item s is smaller than the distance $T_t = X_t - Z_t$ in item t, the participant is inclined to choose item s. The ZG-MUPP model calculates the probability of an individual choosing item s among items s and t as formula (10) to formula (12):

$$P(s > t) = 1 - \Phi(a_{st}^*) - \Phi(b_{st}^*) + 2\Phi(a_{st}^*)\Phi(b_{st}^*) \tag{10}$$

$$a_{st}^* = \frac{1}{\sqrt{2}}[\lambda_s(\theta_s - \mu_s) + \lambda_t(\theta_t - \mu_t)] \tag{11}$$

$$b_{st}^* = \frac{1}{\sqrt{2}}[-\lambda_s(\theta_s - \mu_s) + \lambda_t(\theta_t - \mu_t)] \tag{12}$$

Among them, λ_s represents the discrimination parameter of the item s, μ_s is the location parameter of the item s, and θ_s represents the latent trait measured by item s.

The ZG-MUPP model was created in response to criticism of the MUPP-GGUM model, which has too many parameters, making parameter estimation difficult. Each item in the MUPP-GGUM model includes three types of parameters: discrimination, location, and threshold, which appear to be cumbersome and complex. This complexity makes parameter estimation difficult and increases the need for sample size [21]. As a result, reducing model parameters is necessary. Previous research has shown that when estimating item and latent feature parameters directly, the MUPP model's threshold parameters are more difficult to estimate than other parameters [35]. Moreover, threshold parameters provide little information for MFC testing [45]. Therefore, the ZG-MUPP model removes threshold parameters and retains two model parameters: discrimination λ and location μ. Joo also derived the information function for this model to facilitate its application in CAT and the calculation of SE [45].

All along, the unfolding model has been considered more flexible compared to dominant models [25,26]. However, it has a higher level of complexity and requires a larger sample size. The ZG-MUPP model simplifies the unfolding model and greatly increases its competitiveness.

3. Parameter estimation methods

To obtain the parameters of the forced choice model in complex scenarios with multidimensional data, some parameter estimation algorithms must be used. Based on the estimation process, these approaches can be divided into joint estimation and two-phase estimation strategies. The least squares algorithm, which is based on the maximum likelihood estimation (MLE) approach, and the Markov Chain Monte Carlo (MCMC) algorithm, which is based on the Bayesian method, are the two main algorithms used in joint estimation.

3.1. Two-phase estimation strategy

MUPP-GGUM uses a two-phase strategy: the item parameters needed to calculate $P(i)$ and $Q(i)$ were pre-calibrated by Likert scale data in steps 2–3 and estimated by the GGUM2000 computer program (Roberts, Donoghue, & Laughlin, 2000b). In step 7, forced-choice response data were used to estimate ability using MUPP-GGUM. Stark et al. [33,46] achieved the Maximum A Posteriori (MAP) for high-dimensional latent trait estimation using a BFGS (Broyden-Fletcher-Goldfarb-Shanno) method similar to Newton-Raphson iterations. Expected A Posteriori (EAP) or MLE can also be used to estimate latent traits. Because an increase in the number of dimensions leads to an exponential increase in the number of nodes for numerical integration in EAP, EAP is best suited for 1–2 dimensions, whereas MAP and MLE are best suited for a large number of dimensions.

Stark implemented the BFGS algorithm in DFPMIN [47], but it can also be implemented in R by specifying the method parameter as L-BFGS-B in the function "optim." In the area of item parameter calibration, GGUM has made many breakthroughs in parameter estimation in recent years [48], supported by the related R packages GGUM [49], mirt [50], and Bmggum [51].

This model makes an implicitly strong assumption that item parameters are consistent across test formats, which may not be correct. However, this process is very beneficial to the management of the item bank and then facilitates the development of forced-choice adaptive tests.

3.2. Joint estimation strategy

3.2.1. Least squares algorithm

TIRT is developed based on the structural equation model. The structural equation modeling software Mplus [52] or the Lavaan package [53] can estimate item parameters using unweighted least squares or diagonally weighted least squares. MLE, MAP, and EAP methods can also be used to estimate latent traits.

Brown and Maydeu-Olivares [54] provide an Excel macro (http://annabrown.name/software) that can export Mplus statements after entering the test design for the convenience of practitioners. Bürkner provides functions for data simulation in the thurstonianIRT package and serves as an interface for users to select the Lavaan package [53] or Mplus as the intrinsic processing of model fit methods, and can automatically generate codes based on the method selected by the user [55].

Obviously, the development of the TIRT software kit provides great convenience for practitioners, which is one of the reasons why TIRT is widely used, but there are some reservations. For example, Bürkner et al. [55] discovered serious model failure to converge when using Mplus and Lavaan to fit TIRT, particularly in the case of large tests (for example, a 5-dimension test with 27 items in each dimension, the model convergence rate is only about 0.3). Furthermore, a large amount of RAM is required (for example, for a 30-dimension test, where each dimension has nine item blocks and the model requires 32 GB of RAM); otherwise, it is necessary to specify in the code not to calculate the chi-square, standard error, and other fitting indices to reduce operating time and operating pressure. The most common error is a negative variance, which often necessitates specifying inter-dimensional relationships or factor loadings to facilitate convergence, but the estimation results also heavily rely on these fixed values. Given the sensitivity of TIRT in model identification, if TIRT is considered for use in a test with high dimensions, it is necessary to fully ensure the quality of the items during test development, such as through unidimensionality testing of the items to ensure the characteristics of unidimensionality. The issue of RAM must be considered when selecting an estimation method. Otherwise, the model does not converge or the memory is insufficient to obtain any estimation results, reducing the test developer's confidence in the test quality and the model.

3.2.2. MCMC algorithm

Unlike TIRT, the proposers of the later models all based the parameter estimation algorithm on MCMC. It is a probabilistic, full-information parameter estimation method that does not necessitate complex mathematical derivation but only requires researchers to construct a reasonable posterior probability distribution function and can achieve estimation accuracy comparable to frequentist algorithms (maximum likelihood estimation, etc.). The Metropolis-Hasting MCMC algorithm is used by the MUPP-2PL, GGUM-Rank, RIM, and BRB-IRT models to estimate item and ability parameters based on forced-choice data.

Table 5
Summary of model parameter estimation methods.

Parameter Estimation Methods	Software Implementations	Advantage	Disadvantage
Two steps: 1. Pre-calibrate item parameters based on Likert scale data 2. BFGS estimation power	1. R package: GGUM/mirt/bmggum 2. DFPMIN/R package: stats	Pre-calibration of item parameters is convenient for self-adaptive item bank management	Using Likert item parameter on forced-choice data to estimate ability have the risk of inconsistent item parameter across test formats
Weighted Least Squares/ Diagonally Weighted Least Squares	Mplus R package: thurstonianIRT (Mplus/Lavaan method)	Estimated time is short, easy to use	not easy to converge in high-dimensional situations, the memory usage is too high, and sometimes the calculation of the fitting index needs to be discarded
MCMC	Ox/WinBUGS/JAGS/OpenBUGS R package: thurstonianIRT (Stan method)	no convergence problem	Long estimated time, uneasy to use

Ox [56], OpenBUGS 3.2.3 [57], WinBUGS [58], JAGS [59], and other software can implement the MCMC algorithm. WinBUGS and OpenBUGS are relatively slow among these software programs, whereas the MCMC method developed by Bürkner et al. [55] for TIRT uses Stan [60] language, and the estimation speed is greatly improved by using the more advanced NUTS (No-U-Turn sampler) or HMC (Hamiltonian Monte Carlo) sampling methods. They all use the statistics $\widehat{R}$, proposed by Gelman & Rubin [61] in the model convergence evaluation criteria (less than 1.2 means the parameters have converged). Although these models do not have significant convergence issues, they do necessitate practitioners having a deeper understanding of MCMC-related knowledge and implementation steps, and the main disadvantage of MCMC methods is the long estimation time [62].

See Table 5 for a summary of various model parameter estimation methods.

4. Applied research

In the field of industrial and organizational psychology, the MFC-IRT model is widely used. For example, TIRT has been used to develop the Assessment of Work-Related Maladaptive Personality Traits [63], as well as the Occupational Personality Questionnaire (OPQ32r) and the Customer Contact Styles Questionnaire (CCSQ) [34,64]. In the 360-degree feedback test, it has also been suggested that using forced-choice tests and TIRT scoring has better construct validity and aggregate validity than using traditional Likert rating scales to score [65]. The Adaptive Employee Personality Test (Adept-15) [66] and the Tailored Adaptive Personality Assessment System [67] both use MUPP-GGUM. These two tests are also a ground-breaking attempt at a Computerized Adaptive Test (CAT) forced-choice test. Simultaneously, the test of item parameter invariance is an important part of the test development process, and the invariance test method for the forced-choice test is being developed and improved gradually. Considerable evidence has also accumulated in the field of validity research, which practitioners are increasingly interested in. As a result, this paper will summarize the current state of research in three areas: parameter invariance tests, CAT, and validity research.

4.1. Parameter invariance test

To ensure that all participants understand the item in the same way, test developers must run the measurement consistency test (item parameter invariance). In the forced-choice test, item parameter invariance can be classified as cross-block consistency or cross-population consistency. Items that lack parameter invariance indicate that their likelihood of answering is influenced by factors other than the measurement target.

The degree to which an item maintains parameter invariance when paired with different items across item blocks is measured by cross-block consistency. Block 1 (which contains items A, B, and C) and Block 2 (which contains items A, D, and E) are two item blocks that share item A. The estimation results for item parameters for item A in the two item blocks should be consistent, indicating cross-block parameter invariance. Lin and Brown [68] used the TIRT to compare the parameter invariance of two sets of Rank-3 and MOLE-4. Because the latter only added one new item to each of the former's item blocks, the proportion of common items between each pair of item blocks was 75%, and only a few items had significant deviations.

Cross-population consistency refers to whether an item has parametric invariance between people from different backgrounds (for example, people of different genders or different test situations). Differential Item Functioning (DIF) is another term for testing for such variability. If the item parameters differ significantly between groups, it indicates that the individual's background influences the likelihood of answering this item. If the test contains an excessive number of such items, the test's validity will be reduced and it will be unfair. Lee & Smith [69] tested the measurement invariance of TIRT using multiple group confirmatory factor analyses (CFA). It is suggested that ΔCFI $>$0.007 and ΔCFI $>$0.001 be the critical values of metric non-invariance and scalar non-invariance, respectively, but this method cannot be specific to the item to screen. The parameter inconsistency at the item level is DIF. DIF is the parameter inconsistency at the item level. P. Lee et al. [70] proposed an Omnibus Wald test for the discrimination and intercept indicators of the TIRT and suggested through simulation research that the detection efficiency was higher under the free baseline method: the detection rate was close to 1 and the type I error rate was close to 0.05 as sample size and DIF amount increased. Qiu & Wang [71] proposed three DIF test methods for RIM including EMD (equal-mean-difficulty), AOS (all-other-statement), and CS (constant-statement). Finally, it was found that the CS performed better than the other two methods in the test with DIF items.

4.2. Computerized adaptive testing

The measurement dimensions of personality assessment tools are typically high-dimensional due to the complexity of human personality. OPQ32r, for example, assesses 32 personality dimensions. The more dimensions there are, the more items are needed. Excessively large items will cause individual fatigue and boredom with the test, leading to careless answering. From the standpoint of measurement efficiency, when individuals in some dimensions have reached an acceptable measurement precision by a small number of items, which can be for individuals to have certain judgments on these dimensions, a subsequent focus on the items of the evaluation of uncertainty in the higher dimensions, a review of individuals in all dimensions as soon as possible can reach the level of reliability. Evaluation efficiency can thus be improved. One solution to the aforementioned problem is to create a CAT version of the forced-choice test.

The forced-choice CAT was first used to select US Navy personnel 15 years ago. Houston et al. [72] created the Navy Computer Adaptive Personality Scales, which assess 19 personality traits. Stark et al. [46] proposed a six-step forced-choice adaptive procedure for multi-dimensional and unidimensional pick-2 (single and multidimensional blocks) using MUPP-GGUM. The most significant difference from traditional CAT is that it must predetermine the proportion of unidimensional item blocks and the dimensional combination form of navigating and storing multidimensional item blocks. The two studies mentioned above imply that CAT improves efficiency more than non-CAT. The forced-choice CAT can be correctly only requiring half of the non-adaptive test questions. In addition, TAPAS, which is based on MUPP-GGUM, is an adaptable personality test for US military selection [67].

The ideal-point measurement model is currently used in the majority of computer-adaptive forced-choice tests, but dominant items have several practical advantages over ideal-point items. According to Brown and Maydeu-Olivares [27], creating ideal-point items is more difficult in content development. Ideal-point items have fewer analytic software options and are more difficult to estimate item parameters for [73]. Chen et al. [74] investigated the FC CAT with dominant items using the Rasch model, and Lin et al. [75] conducted the first empirical study on MFC CAT dominance items using TIRT model.

The assembly of blocks and the guidelines for block selection determine the validity of multidimensional FC-CAT. Fixed assembly and dynamic assembly are two methods of assembling blocks. A benefit of constant assembly is consistent block parameters. But fixed assembly produces fewer blocks from the same item pool than dynamic assembly. Dynamic assembly selects items in real time while the subject takes the test. Flexible matching generates more blocks, making item leakage harder. However, the stability of item parameters over these blocks may be an issue. And dynamically creating blocks requires aligning things with similar social desirability [14]. The genetic algorithm-based NHBSA (Node Histogram-Based Sampling Algorithm) [76] can help achieve this goal [77].

Block selection rules balance information for each dimension. Data from previous blocks determines the next block the subject should answer during the test. There are three FI-based (Fisher Information) selection rules for multi-dimensional CAT, which Mulder & van der Linden [78] categorized as A-optimality (trace), D-optimality (determinant), and E-optimality (eigenvalue). Among them, the A-optimality method has slightly better estimation accuracy than the D-optimal method, and the E-optimal method is the most unstable (Mulder & van der Linden, 2009). Veldkamp & van der Linden [79] proposed a posterior expectation KL information (K^B method) based on the KL information as an alternative to FI, which includes three selection rules: KL index (KI), posterior expected KL information (K^B), and posterior KL distance (KLP) between subsets [78–82].

Chen et al. [74] proposed three subpool selection strategies to improve the efficiency of item selection and control the exposure rate of items. The three strategies are the Sequential Strategy, the Multinomial Strategy, and the High-SE Strategy. The Sequential Strategy will choose items from each combination of items based on the amount of information until the termination standard is reached. The Multinomial Strategy solves the problem of sequential strategies by randomly selecting a sub-database based on the polynomial distribution. The high-SE Strategy first determines which dimensions an individual has the highest SE in, and then selects the item blocks of the corresponding dimension combination. In terms of overall performance, the Multinomial Strategy does well.

Furthermore, Chen et al. [74] proposed the Revised Sympson-Hetter Online (RSHO) to control the exposure rate of items. When selecting item blocks, first determine the most appropriate item blocks based on the amount of information and then select items with less exposure. The RSHO regulates the exposure rate of the items while slightly sacrificing measurement accuracy.

According Seybert & Becker [82], the retest reliability of the forced-choice CAT is lower than that of traditional Likert rating scales [83], but comparable to the retest reliability of duplicate Likert rating scales.

4.3. Validity studies

To answer this question, researchers focused on five areas to see if the IRT's latent trait scores could accurately reflect individuals' true characteristics. The first is to determine whether IRT scoring recovers better latent traits and their relationships than traditional scoring [12,22,84]. Using IRT to estimate trait scores can result in a significant improvement in measurement accuracy when compared to traditional scoring, which is almost the common conclusion of all studies in this direction, and it also gives researchers great confidence to develop more IRT. Some studies, however, have found that the results of forced-choice models are not always as good as those of traditional scoring models [40,85,86]. However, the extent to which the scores obtained from these models can be interpreted as traditional normality scores merits further investigation because it is directly related to whether these scores can be used as normality scores for personnel selection or for correlation analysis with external criteria.

To answer the above questions, the second direction attempted to investigate the relationship between IRT and latent trait scores obtained by the Likert single stimulus scale [42,64,87,88]. The score of a single stimulus scale is thought to be the most consistent with the true value of individual latent traits in these studies. If the score obtained by the forced-choice model maintains a high similarity with the score origin, size, and dimension relationships, the equivalence of the Likert scale and forced-choice scale will be proved.

The third direction is to investigate the ability of forced-choice tests to detect fraud. When the social desirability of the forced-choice item block is matched, the forced-choice test outperforms the Likert rating scales in terms of anti-fraud ability [89]. When compared to using TIRT to analyze the forced-choice test, using the Graded Response Model (GRM) to analyze the Likert rating scales cannot effectively distinguish high-ability individuals because participants tend to perform better, resulting in low discrimination of items that reflect high ability [90].

The fourth direction is to investigate the application of IRT in non-self-rating situations. Because Likert rating scales have a common method deviation, different raters' evaluations are influenced by their internal ideal behavior standards, resulting in low consistency and reliability among raters. Hung et al. [91] proposed Forced-Choice Ranking Models (FCRM), which quantify rater leniency and task difficulty as new indicators and are useful in non-self-rated scenarios.

The fifth direction is reliability research. Reliability affects validity, so reliability is a crucial indicator for evaluating the applicability of FC models. Many scholars have reported on reliability indicators when conducting IRT-FC studies [7,23,27,34,35,37,42,43,54,88,89,92]. According to Lin et al. [92], reliability indicators include theoretical reliability, empirical reliability, simulated true estimation reliability, and retest reliability. The evaluation score is influenced by four aspects of measurement errors [93]: random errors (due to random fluctuations in an individual's responses), transient errors (due to situational factors affecting a specific measurement occasion), item-specific errors (due to consistent interindividual differences in item interpretation), and scale-specific errors (due to different measurement operationalizations of the same psychological construct). Simulating the reliability of true estimation involves random, item-specific, and scale-specific errors, while simulated retest reliability only reflects random errors [92]. Theoretical and empirical reliability often overestimate the reliability of IRT information because of the local independence assumption between pairwise comparisons within the same group (containing more than two items), which is not actually rigorous [90].

5. Future research

IRT research has great potential as a form of assessment that can effectively resist fraud and response biases and improve individual response efficiency, particularly in the application of non-cognitive, high-stakes situation assessment. In addition to the unresolved problems of previous research, the following future research directions are proposed: new FC models, item parameter invariance research, forced-choice CAT research, and validity research.

5.1. New FC models

Current forced-choice models work for Pick, MOLE, and Rank item kinds. There are variations of the Pick-2, including the Adept 15 [66], which asks participants to choose their preferred item and their willingness to do so (see Table 6). Brown and Maydeu-Olivares [94] created a factor analysis model and information function for graded blocks based on Thurston's Law of Comparative Judgment. The DPM model was built to handle graded block data by Qiu et al. [95]. This type of item refines the individual's behavior and provides more information, but the cognitive load of the item was also increased. Only the dominance models RIM and TIRT have now been expanded to a polytomous level. MUPP-GGUM is a dichotomy iteration of GGUM with the potential for future development of polytomous forced ideal point models.

A new kind of IRT model with time has also been developed. As decision time increases, product preferences become less differentiated, making decision-making harder. Thus, their trait levels are likely to be similar. This method is used in the Thurstonian D Diffusion Model [96], the Linear Ballistic Accumulator Item Response Theory Model [97], and the Guo et al.'s [98] Log-Linear Model. Collecting response time data in this setting is effortless, therefore augmenting the quantity of information and improving the efficiency of the model without imposing additional cognitive load on the participants. Each traditional model has the capability to create a similar version that incorporates the element of time. In the future, more models may be taken into account for temporal extension.

Both polytomous models and models with time can obtain more information, which produces more accurate parameter estimation results. All models in Table 4 can be extended in these two directions. However, it should be noted that models with time do not increase cognitive load and are generally only suitable for computerized testing with simple data collection; the polytomous model increases cognitive load and thus has significant limitations on the size of blocks.

5.2. Research on parameter invariance based on each model

Following Lin and Brown's study [68] on TIRT, when the proportion of common items decreases, whether a higher proportion of items' parameters can still have cross-item block invariance remains to be studied. In addition, the cross-item block consistency of other models needs to be studied.

At present, there is only research on the parameter invariance of TIRT [69,70] and RIM [71]. Future studies should broaden the repertoire of differential item functioning (DIF) test methods for the forced-choice model and improve their sensitivity in detecting DIF from multiple sources.

5.3. Forced-choice CAT

Although forced-choice CAT has accumulated more experience in empirical research, the adaptive process for latent trait estimation has been developed using item parameters calibrated in advance with a single stimulus scale. The database used for items is the single-item database rather than the item block database. During item selection, items will be combined to form the forced-choice item blocks, so the impact of cross-item block consistency on latent trait estimation under this CAT process needs further study. In addition, the combination of item block dimensions and the test length will increase significantly in a high-dimensional situation, which brings challenges to content balance and test efficiency. In the future, we can further explore how to elaborate on the advantages of CAT in a high-dimensional situation. Although the subpool partition strategies for item selection and within-person statement exposure control procedures proposed by Chen et al. [74] do not involve scoring and can be extended to CAT based on other non-RIM models, the

Table 6
Pick-2 graded blocks.

	Slightly agree	Agree
A lack of finding things		✓
B explore unfamiliar territory		

specific performance still needs to be explored by research. In addition, control methods such as The Multinomial Strategy cannot be directly applied to the variable-length test. In the future, we can further explore how to construct a more appropriate item selection strategy.

5.4. Validity studies

A large amount of research compares forced-choice tests and Likert rating scales to see if they measure the same measurement content similarly. However, the difference between the two in the test form and the response biases caused by Likert rating scales will inevitably lead to some errors. It is worth exploring how to maximize the control of these biases in the future. In the form of forced choice, the larger the item block, the stronger the resistance to faking, but it also increases the cognitive load [89]. Future research can explore the balance between the anti-faking effect and cognitive load on the size of the item block. In addition, most of the existing validity studies focus on TIRT, and the validity studies of GGUM-Rank and other new models need to be explored.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Ethics requirements statement

This article does not contain any studies with human participants performed by any of the authors.
References

CRediT authorship contribution statement

Lei Nie: Writing – original draft, Resources, Investigation, Funding acquisition. **Peiyi Xu:** Writing – review & editing, Project administration, Conceptualization. **Di Hu:** Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Luo, H. Zhang, Methods of coping with faking of personality tests, Psychological Exploration 27 (4) (2007) 78–82.
[2] H. Aguinis, M.M. Handelsman, Ethical issues in the use of the bogus pipeline, J. Appl. Soc. Psychol. 27 (7) (1997) 557–573, https://doi.org/10.1111/j.1559-1816.1997.tb00647.x.
[3] L.A. White, M.C. Young, Development and Validation of the Assessment of Individual Motivation (AIM), Paper presented atthe annual meeting of the American Psychological Association, San Francisco, 1998, August.
[4] H. Baron, Strengths and limitations of ipsative measurement, J. Occup. Organ. Psychol. 69 (1) (1996) 49–56, https://doi.org/10.1111/j.2044-8325.1996.tb00599.x.
[5] S. Frick, A. Brown, E. Wetzel, Investigating the normativity of trait estimates from multidimensional forced-choice data, Multivariate Behav. Res. 58 (1) (2023) 1–29, https://doi.org/10.1080/00273171.2021.1938960.
[6] S. Wang, F. Luo, H. Liu, The conventional and the IRT-based scoring methods of Forced-Choice personality tests, Adv. Psychol. Sci. 22 (3) (2014) 549–557, https://doi.org/10.3724/SP.J.1042.2014.00549.
[7] A. Brown, A. Maydeu-Olivares, How IRT can solve problems of ipsative data in forced-choice questionnaires, Psychol. Methods 18 (1) (2013) 36–52, https://doi.org/10.1037/a0030641.
[8] S.J. Closs, On the factoring and interpretation of ipsative data, J. Occup. Organ. Psychol. 69 (1) (1996) 41–47, https://doi.org/10.1111/j.2044-8325.1996.tb00598.x.
[9] D. Bartram, The relationship between ipsatized and normative measures of personality, J. Occup. Organ. Psychol. 69 (1) (1996) 25–39, https://doi.org/10.1111/j.2044-8325.1996.tb00597.x.
[10] W.V. Clemans, An analytical and empirical examination of some properties of ipsative measures, Psychometric Monographs 14 (1966).
[11] P. Saville, E. Willson, The reliability and validity of normative and ipsative approaches in the measurement of personality, J. Occup. Psychol. 64 (3) (1991) 219–238, https://doi.org/10.1111/j.2044-8325.1991.tb00556.x.
[12] P.M. Hontangas, J. de la Torre, V. Ponsoda, I. Leenen, D. Morillo, F.J. Abad, Comparing traditional and IRT scoring of forced-choice tests, Appl. Psychol. Meas. 39 (8) (2015) 598–612, https://doi.org/10.1177/0146621615585851.
[13] A. Brown, Item response models for forced-choice questionnaires: a common framework, Psychometrika 81 (1) (2016) 135–160, https://doi.org/10.1007/s11336-014-9434-9.
[14] G. Pavlov, D. Shi, A. Maydeu-Olivares, A. Fairchild, Item desirability matching in forced-choice test construction, Pers. Indiv. Differ. 183 (2021) 111114, https://doi.org/10.1016/j.paid.2021.111114.
[15] K.L. Gwet, Handbook of inter-rater reliability, in: The Definitive Guide to Measuring the Extent of Agreement Among Raters, fourth ed., Advanced Analytics, LLC, Gaithersburg, MD, 2014.
[16] R Core Team, R: A Language and Environment for Statistical Computing, 2021. Vienna, Austria, https://www.R-project.org/.

[17] M. Li, T. Sun, B. Zhang, autoFC: an R Package for Automatic Item Pairing in Forced-Choice Test Construction, Applied Psychological Measurement, Advance online publication, 2021, https://doi.org/10.1177/01466216211051726.

[18] A.W. Hughes, P.D. Dunlop, D. Holtrop, S. Wee, Spotting the “Ideal” personality response: effects of item matching in forced choice measures for personnel selection, J. Person. Psychol. 20 (1) (2021) 17–26, https://doi.org/10.1027/1866-5888/a000267.

[19] S. Frick, Modeling faking in the multidimensional forced-choice format: the faking mixture model, Psychometrika 87 (2022) 773–794, https://doi.org/10.1007/s11336-021-09818-6.

[20] F. Drasgow, O.S. Chernyshenko, S. Stark, 75 years after Likert: Thurstone was right, Industrial and Organizational Psychology 3 (4) (2010) 465–476, https://doi.org/10.1111/j.1754-9434.2010.01273.x.

[21] J.S. Roberts, J.R. Donoghue, J.E. Laughlin, A general item response theory model for unfolding unidimensional polytomous responses, Appl. Psychol. Meas. 24 (1) (2000) 3–32, https://doi.org/10.1177/01466216000241001.

[22] P.M. Hontangas, I. Leenen, J. de la Torre, V. Ponsoda, D. Morillo, F.J. Abad, Traditional scores versus IRT estimates on forced-choice tests based on a dominance model, Psicothema 28 (1) (2016) 76–82, https://doi.org/10.7334/psicothema2015.204.

[23] D. Morillo, I. Leenen, F.J. Abad, P. Hontangas, J. de la Torre, V. Ponsoda, A dominance variant under the multi-unidimensional pairwise-preference framework: model formulation and Markov chain Monte Carlo estimation, Appl. Psychol. Meas. 40 (7) (2016) 500–516, https://doi.org/10.1177/0146621616662226.

[24] O.S. Chernyshenko, S. Stark, K.Y. Chan, F. Drasgow, B. Williams, Fitting item response theory models to two personality inventories: issues and insights, Multivariate Behav. Res. 36 (4) (2001) 523–562, https://doi.org/10.1207/S15327906MBR3604_03.

[25] L. Tay, U.S. Ali, F. Drasgow, B. Williams, Fitting IRT models to dichotomous and polytomous data: assessing the relative model–data fit of ideal point and dominance models, Appl. Psychol. Meas. 35 (4) (2011) 280–295, https://doi.org/10.1177/0146621610390674.

[26] J. Huang, A.D. Mead, Effect of personality item writing on psychometric properties of ideal-point and Likert scales, Psychol. Assess. 26 (4) (2014) 1162–1172, https://doi.org/10.1037/a0037273.

[27] A. Brown, A. Maydeu-Olivares, Issues that should not be overlooked in the dominance versus ideal point controversy, Industrial and Organizational Psychology 3 (4) (2010) 489–493, https://doi.org/10.1111/j.1754-9434.2010.01277.x.

[28] F.L. Oswald, K.L. Schell, Developing and scaling personality measures: Thurstone was right—but so far, likert was not wrong, Industrial and Organizational Psychology 3 (4) (2010) 481–484, https://doi.org/10.1111/j.1754-9434.2010.01275.x.

[29] L.L. Thurstone, A law of comparative judgment, Psychol. Rev. 34 (4) (1927) 273–286, https://doi.org/10.1037/h0070288.

[30] R.D. Luce, On the possible psychophysical laws, Psychol. Rev. 66 (2) (1959) 81–95, https://doi.org/10.1037/h0043178.

[31] R.D. Luce, The choice axiom after twenty years, J. Math. Psychol. 15 (3) (1977) 215–233, https://doi.org/10.1016/0022-2496(77)90032-3.

[32] R.A. Bradley, M.E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons, Biometrika 39 (3/4) (1952) 324–345, https://doi.org/10.2307/2334029.

[33] S. Stark, O.S. Chernyshenko, F. Drasgow, An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: the multi- unidimensional pairwise-preference model, Appl. Psychol. Meas. 29 (3) (2005) 184–203, https://doi.org/10.1177/0146621604273988.

[34] A. Brown, A. Maydeu-Olivares, Item response modeling of forced-choice questionnaires, Educ. Psychol. Meas. 71 (3) (2011) 460–502, https://doi.org/10.1177/0013164410375112.

[35] P. Lee, S.-H. Joo, S. Stark, O.S. Chernyshenko, GGUM-Rank statement and person parameter estimation with multidimensional forced choice triplets, Appl. Psychol. Meas. 43 (3) (2019) 226–240, https://doi.org/10.1177/0146621618768294.

[36] C.J. Zheng, J. Liu, Y.L. Li, P.Y. Xu, B. Zhang, R. Wei, W.Q. Zhang, A 2PLM-RANK Multidimensional Forced-Choice Model and its Fast Estimation Algorithm Behav, Res. (2024), https://doi.org/10.3758/s13428-023-02315-x.

[37] P.-C. Bürkner, N. Schulte, H. Holling, On the statistical and practical limitations of Thurstonian IRT models, Educ. Psychol. Meas. 79 (5) (2019) 827–854, https://doi.org/10.1177/0013164419832063.

[38] H. Li, Y. Xiao, H. Liu, Influencing factors of Thurstonian IRT model in faking-resisting forced-choice questionnaire, J. Beijing Normal Univ. (Nat. Sci.) 53 (5) (2017) 624–630, https://doi.org/10.16360/j.cnki.jbnuns.2017.05.019.

[39] X. Lian, Q. Bian, S. Zeng, H. Che, The Fitting Analysis of MAP Occupational Personality Forced Choice Test Based on Thurston IRT Model, National Conference on Psychology, *Beijing*, China, 2014.

[40] N. Schulte, H. Holling, P.-C. Bürkner, Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? Educ. Psychol. Meas. 81 (2) (2021) 262–289, https://doi.org/10.1177/0013164420934861.

[41] P. Lee, S. Joo, S. Zhou, M. Son, Investigating the impact of negatively keyed statements on multidimensional forced-choice personality measures: a comparison of partially ipsative and IRT scoring methods, Pers. Indiv. Differ. 191 (2022) 1–15, https://doi.org/10.1016/j.paid.2022.111555.

[42] T. Joubert, I. Inceoglu, D. Bartram, K. Dowdeswell, Y. Lin, A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument, Int. J. Sel. Assess. 23 (1) (2015) 92–97, https://doi.org/10.1111/ijsa.12098.

[43] S.-H. Joo, P. Lee, S. Stark, Development of information functions and indices for the GGUM-Rank multidimensional forced choice IRT model, J. Educ. Meas. 55 (3) (2018) 357–372, https://doi.org/10.1111/jedm.12183.

[44] S.-H. Joo, P. Lee, S. Stark, Adaptive testing with the GGUM-Rank multidimensional forced choice model: comparison of pair, triplet, and tetrad scoring, Behav. Res. Methods 52 (2) (2020) 761–772, https://doi.org/10.3758/s13428-019-01274-6.

[45] S.-H. Joo, P. Lee, S. Stark, Modeling Multidimensional Forced Choice Measures with the Zinnes and Griggs Pairwise Preference Item Response Theory Model. Multivariate Behavioral Research, Advance online publication, 2021, https://doi.org/10.1080/00273171.2021.1960142.

[46] S. Stark, O.S. Chernyshenko, F. Drasgow, L.A. White, Adaptive testing with multidimensional pairwise preference items, Organ. Res. Methods 15 (3) (2012) 463–487, https://doi.org/10.1177/1094428112444611.

[47] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recipes: the Art of Scientific Computing, Cambridge University Press, New York, 1986.

[48] J.S. Roberts, V.M. Thompson, Marginal maximum a posteriori item parameter estimation for the generalized graded unfolding model, Appl. Psychol. Meas. 35 (4) (2011) 259–279, https://doi.org/10.1177/0146621610392565.

[49] J.N. Tendeiro, S. Castro-Alvarez, GGUM: an R package for fitting the generalized graded unfolding model, Appl. Psychol. Meas. 43 (2) (2018) 172–173.

[50] R.P. Chalmers, mirt: a multidimensional item response theory package for the R environment, J. Stat. Software 48 (6) (2012) 1–29.

[51] N. Tu, B. Zhang, L. Angrave, T. Sun, Bmggum: an R package for Bayesian estimation of the multidimensional generalized graded unfolding model with covariates, Appl. Psychol. Meas. 45 (7–8) (2021) 553–555.

[52] L. Muthén, B. Muthén, Mplus. The Comprehensive Modelling Program for Applied Researchers: User’s Guide, vol. 5, Author, Los Angeles, CA, 2015.

[53] Y. Rosseel, lavaan: an R package for structural equation modeling, J. Stat. Software 48 (2) (2012) 1–36.

[54] A. Brown, A. Maydeu-Olivares, Fitting a Thurstonian IRT model to forced-choice data using Mplus, Behav. Res. Methods 44 (4) (2012) 1135–1147, https://doi.org/10.3758/s13428-012-0217-x.

[55] P.-C. Bürkner, thurstonianIRT: Thurstonian IRT models in R, J. Open Source Softw. 4 (42) (2018) 1662, https://doi.org/10.21105/joss.01662.

[56] J.A. Doornik, An Object-Oriented Matrix Programming Language Ox 6, Timberlake Consultants Ltd, London, England, 2009.

[57] D. Lunn, D. Spiegelhalter, A. Thomas, N. Best, The BUGS project: evolution, critique and future directions, Stat. Med. 28 (25) (2009) 3049–3067, https://doi.org/10.1002/sim.3680.

[58] D. Spiegelhalter, A. Thomas, N. Best, *WinBUGS Version 1.4* [Computer Program], MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK, 2003.

[59] M. Plummer, JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling, International Workshop on Distributed Statistical Computing, Vienna, Austria, 2003. Paper presented at the 3rd.

[60] Stan Development Team, RStan: the R Interface to Stan, 2020. http://mc-stan.org/.

[61] A. Gelman, D. Rubin, Inference from iterative simulation using multiple sequences, Stat. Sci. 7 (4) (1992) 457–472, https://doi.org/10.1214/ss/1177011136.

[62] J.-S. Kim, D. Bolt, Estimating item response theory models using Markov chain Monte Carlo methods, Educ. Meas. 26 (4) (2007) 38–51, https://doi.org/10.1111/j.1745-3992.2007.00107.x.

[63] N. Guenole, A. Brown, A. Cooper, Forced-choice assessment of work-related maladaptive personality traits: preliminary evidence from an application of Thurstonian item response patterning, Assessment 25 (4) (2016) 513–526, https://doi.org/10.1177/1073191116641181.

[64] SHL, OPQ32r Technical Manual, SHL, 2018.

[65] A. Brown, I. Inceoglu, Y. Lin, Preventing rater biases in 360-degree feedback by forcing choice, Organ. Res. Methods 20 (1) (2017) 121–148, https://doi.org/10.1177/1094428116668036.

[66] Hewitt Aon, 2015 Trends in Global Employee Engagement Report, Aon Corp, Lincolnshire, IL, 2015.

[67] S. Stark, O.S. Chernyshenko, F. Drasgow, C.D. Nye, L.A. White, T. Heffner, W.L. Farmer, From ABLE to TAPAS: a new generation of personality tests to support military selection and classification decisions, Mil. Psychol. 26 (3) (2014) 153–164, https://doi.org/10.1037/mil0000044.

[68] Y. Lin, A. Brown, Influence of context on item parameters in forced-choice personality assessments, Educ. Psychol. Meas. 77 (3) (2017) 389–414, https://doi.org/10.1177/0013164416646162.

[69] H. Lee, W.Z. Smith, Fit indices for measurement invariance tests in the Thurstonian IRT model, Appl. Psychol. Meas. 44 (4) (2020) 282–295, https://doi.org/10.1177/0146621619893785.

[70] P. Lee, S.-H. Joo, S. Stark, Detecting DIF in multidimensional forced choice measures using the Thurstonian item response theory model, Organ. Res. Methods 24 (4) (2020) 739–771, https://doi.org/10.1177/1094428120959822.

[71] X.-L. Qiu, W.-C. Wang, Assessment of differential statement functioning in ipsative tests with multidimensional forced-choice items, Appl. Psychol. Meas. 45 (2) (2021) 79–94, https://doi.org/10.1177/01466216209657.

[72] J. Houston, W. Borman, W. Farmer, R. Bearden, *Development of the Navy Computer Adaptive Personality Scales (NCAPS)* (NPRST-TR-06-2), Navy Personnel Research, Millington, TN, 2006. Studies, and Technology.

[73] C.G. Forero, A. Maydeu-Olivares, Estimation of IRT graded response models: limited versus full information methods, Psychol. Methods 14 (3) (2009) 275–299, https://doi.org/10.1037/a0015825.

[74] C.-W. Chen, W.-C. Wang, M.M. Chiu, S. Ro, Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items, J. Educ. Meas. 57 (2) (2020) 343–369, https://doi.org/10.1111/jedm.12252.

[75] Y. Lin, A. Brown, P. Williams, Multidimensional forced-choice CAT with dominance items: an empirical comparison with optimal static testing under different desirability matching, Educ. Psychol. Meas. 83 (2) (2023) 322–350, https://doi.org/10.1177/00131644221077637.

[76] S. Tsutsui, Node histogram vs. edge histogram: a comparison of probabilistic model-building genetic algorithms in permutation domains, IEEE International Conference on Evolutionary Computation (2006) 1939–1946, https://doi.org/10.1109/CEC.2006.1688544.

[77] R.S. Kreitchmann, F.J. Abad, M.A. Sorrel, A genetic algorithm for optimal assembly of pairwise forced-choice questionnaires, Behav Res 54 (2022) 1476–1492, https://doi.org/10.3758/s13428-021-01677-4.

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

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

[80] H.H. Chang, Z. Ying, A global information approach to computerized adaptive testing, Appl. Psychol. Meas. 20 (1996) 213–229, https://doi.org/10.1177/014662169602000303.

[81] Q. Wang, Y. Zheng, K. Liu, Y. Cai, S. Peng, D. Tu, Item Selection Methods in Multidimensional Computerized Adaptive Testing for Forced-Choice Items Using Thurstonian IRT Model. Behavior Research Methods, 2023, https://doi.org/10.1177/0146621618762748.

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

[83] J. Seybert, D. Becker, Examination of the test- retest reliability of a forced-choice personality measure, ETS Research Report Series 2019 (1) (2019) 1–17, https://doi.org/10.1002/ets2.12273.

[84] F.L. Oswald, A. Shaw, W.L. Farmer, Comparing simple scoring with IRT scoring of personality measures: the navy computer adaptive personality scales, Appl. Psychol. Meas. 39 (2) (2015) 144–154, https://doi.org/10.1177/0146621614559517.

[85] W.-C. Wang, X.-L. Qiu, C.-W. Chen, S. Ro, K.-Y. Jin, Item response theory models for ipsative tests with multidimensional pairwise comparison items, Appl. Psychol. Meas. 41 (8) (2017) 600–613, https://doi.org/10.1177/0146621617703183.

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

[87] L. Watrin, M. Geiger, M. Spengler, O. Wilhelm, Forced-choice versus likert responses on an occupational Big Five questionnaire, J. Indiv. Differ. 40 (3) (2019) 134–148, https://doi.org/10.1027/1614-0001/a000285.

[88] B. Zhang, T. Sun, F. Drasgow, O.S. Chernyshenko, C.D. Nye, S. Stark, L.A. White, Though forced, still valid: psychometric equivalence of forced-choice and single-statement measures, Organ. Res. Methods 23 (3) (2020) 569–590, https://doi.org/10.1177/1094428119836486.

[89] E. Wetzel, S. Frick, A. Brown, Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking, Psychol. Assess. 33 (2) (2020) 156–170, https://doi.org/10.1037/pas0000971.

[90] D.M. Dueber, A.M.A. Love, M.D. Toland, T.A. Turner, Comparison of single-response format and forced-choice format instruments using Thurstonian item response theory, Educ. Psychol. Meas. 79 (1) (2019) 108–128, https://doi.org/10.1177/0013164417752782.

[91] S.-P. Hung, H.-Y. Huang, Forced-choice ranking models for raters' ranking data, J. Educ. Behav. Stat. 47 (5) (2022) 603–634, https://doi.org/10.3102/10769986221104207.

[92] Y. Lin, Reliability estimates for IRT-based forced-choice assessment scores, Organ. Res. Methods 25 (3) (2022) 575–590, https://doi.org/10.1177/1094428121999086.

[93] T. Gnambs, Facets of measurement error for scores of the big five: three reliability generalizations, Pers. Indiv. Differ. 84 (2015) 84–89, https://doi.org/10.1016/j.paid.2014.08.019.

[94] A. Brown, A. Maydeu-Olivares, Ordinal factor analysis of graded-preference questionnaire data, Struct. Equ. Model.: A Multidiscip. J. 25 (4) (2018) 516–529, https://doi.org/10.1080/10705511.2017.1392247.

[95] X.L. Qiu, J. Torre, A dual process item response theory model for polytomous multidimensional forced-choice items, Br. J. Math. Stat. Psychol. (2023), https://doi.org/10.1111/bmsp.12303.

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

[97] K. Bunji, K. Okada, Linear ballistic accumulator item response theory model for multidimensional multiple-alternative forced-choice measurement of personality, Multivariate Behav. Res. 57 (4) (2021) 658–678, https://doi.org/10.1080/00273171.2021.1896351.

[98] Z. Guo, D. Wang, Y. Cai, D. Tu, An Item Response Theory Model for Incorporating Response Times in Forced-Choice Measures. Educational and Psychological Measurement, 2023, https://doi.org/10.1177/00131644231171193.