An IRT Scoring Model for Multidimensional Forced-Choice Tests

Authors: Liu Juan, Zheng Chanjin, Li Yunchuan, Lian Xu, Zheng Chanjin

Date: 2021-12-14T00:00:00+00:00

Abstract

Forced-choice (FC) testing is widely employed in non-cognitive assessments due to its ability to control response biases associated with traditional Likert methods. However, conventional scoring approaches for forced-choice tests generate ipsative data, which has been subject to criticism for its unsuitability for inter-individual comparisons. In recent years, the development of multiple forced-choice IRT models has enabled researchers to derive near-normative data from forced-choice tests, thereby renewing interest among researchers and practitioners in forced-choice IRT models. First, six relatively mainstream forced-choice IRT models are classified and introduced based on the decision models and item response models they employ. Subsequently, these models are compared and summarized from two perspectives: model construction rationale and parameter estimation methods. Furthermore, a review is conducted across three applied research domains: parameter invariance testing, computerized adaptive testing (CAT), and validity research. Finally, future research directions are proposed, suggesting that investigations could be advanced in four areas: model extension, parameter invariance testing, forced-choice CAT, and validity research.

Full Text

IRT-Based Scoring Methods for Multidimensional Forced-Choice Tests

Li Yunchuan[1], Zheng Chanjin[2], Li Juan[1], Lian Xu[2] [1]Beijing Insight Online Management Consulting Co., Ltd., Beijing 100102, China [2]East China Normal University, Shanghai 200062, China

Abstract: Forced-choice (FC) tests are widely used in non-cognitive assessments because they can control response biases inherent in traditional Likert methods. However, conventional scoring of forced-choice tests produces ipsative data, which has been criticized as unsuitable for inter-individual comparisons. In

recent years, the development of multiple forced-choice IRT models has enabled researchers to obtain normative data from forced-choice tests, reigniting interest among researchers and practitioners. This paper first classifies and introduces six mainstream forced-choice IRT models based on their adopted decision models and item response models. It then compares and summarizes these models from two perspectives: model construction rationale and parameter estimation methods. Next, it reviews applied research across three domains: parameter invariance testing, computerized adaptive testing (CAT), and validity studies. Finally, future research directions are proposed in four areas: model expansion, parameter invariance testing, forced-choice CAT, and validity research.

Keywords: forced-choice test, ipsative data, TIRT, MUPP, GGUM-RANK

1 Introduction

Psychological assessments can be categorized into cognitive and non-cognitive tests based on their content. Cognitive tests measure individual abilities such as numerical computation, typically have standard answers, and yield higher scores for greater ability. Non-cognitive tests represent one of the most important methods for understanding personality traits, values, and attitudinal tendencies, widely applied in clinical psychological diagnosis, career planning, and personnel decision-making. Numerous validity studies have demonstrated that personality effectively predicts job performance (SHL, 2018; Sitser et al., 2013; Hurtz & Donovan, 2000). Unlike cognitive tests, most non-cognitive assessments use Likert-type rating scales that require individuals to independently evaluate each item (e.g., "I am an organized person") on a scale from 1 (least like me) to 5 (most like me), with no correct answers. In high-stakes assessment contexts such as hiring and selection, individuals can easily manipulate responses to certain items (e.g., those reflecting high conscientiousness or optimism) to appear more desirable to organizations, even when such self-descriptions are inaccurate. This tendency is known as faking or impression management, which compromises the test' s ability to differentiate among candidates and seriously undermines fairness.

To eliminate or reduce faking, researchers typically employ pre-emptive or post-hoc control techniques (Luo & Zhang, 2007). Post-hoc methods include embedding faking detection scales, using bifactor models to control for faking factors (Brown et al., 2017; Hendy et al., 2021), applying mixed Rasch models to identify faking response patterns (Luo & Zhang, 2007), and building decision tree models based on historical data to detect fakers (Ziegler et al., 2012). These approaches aim to identify faked data to avoid making decisions based on contaminated responses. A critical challenge is ensuring high detection accuracy, as falsely identifying honest respondents as fakers is highly undesirable. Pre-emptive techniques, in contrast, aim to prevent faking before or during test administration to obtain uncontaminated data. These include warnings, bogus pipeline techniques, and forced-choice tests. Warnings are the easiest to implement, comprising faking detection warnings and consequence warnings.

Meta-analytic results show that only consequence warnings effectively suppress faking, with optimal results achieved by combining both types (Dwight & Donovan, 2003). However, warnings may elicit more sophisticated faking strategies or increase test anxiety. Therefore, warning only those showing faking tendencies is considered a better solution, with decision tree models determining when to issue warnings (Ziegler et al., 2012). Bogus pipeline techniques deceive respondents into believing they are taking a lie detection test when they are actually completing a genuine assessment, forcing authentic responses. However, this approach is ethically problematic and has been condemned (Aguinis & Handelsman, 1997).

Forced-choice tests require individuals to select the most and least characteristic statements from a set of similarly desirable items or to rank items by preference, making it impossible to endorse all items positively. Because items within a block share similar desirability levels, none is clearly preferable, reducing the likelihood of socially desirable responding. Compared to Likert scales, forced-choice tests are more resistant to faking (Saville & Willson, 1991; Jackson et al., 2000; Wetzel et al., 2020). This format also eliminates other response biases inherent in Likert scales, such as halo effects, central tendency, extreme responding, and acquiescence. Additionally, forced-choice tests effectively reduce score inflation in the socially desirable direction (Cao & Drasgow, 2019) without decreasing respondent motivation (Sass et al., 2020) or causing adverse emotional or cognitive effects (Zhang et al., 2020). A meta-analysis by Bartram (2007) found that forced-choice assessments improved predictive validity for job performance by 50% compared to Likert rating scales. However, traditional scoring of forced-choice tests produces ipsative data, where scores represent intra-individual rankings rather than absolute levels. This data format has limited forced-choice applications in inter-individual comparisons (e.g., personnel selection). Over the past decade, several forced-choice measurement models have been developed to obtain normative latent trait estimates from forced-choice tests, overcoming ipsativity and making forced-choice tests a more promising anti-faking technology.

This paper systematically introduces forced-choice item types, characteristics, traditional scoring methods, and the drawbacks of ipsative data. It then reviews and evaluates six forced-choice IRT models from the perspectives of item response models and decision models. Next, it compares these models in terms of construction rationale, parameter estimation methods, and application research status. Finally, it proposes four future research directions from a practical standpoint: model expansion, parameter invariance testing, forced-choice CAT research, and validity studies.

2 Forced-Choice Test Design and Traditional Scoring

Forced-choice tests typically consist of several item blocks measuring different dimensions. Each block contains a fixed number of statements/items from different or identical dimensions with similar social desirability levels; these state-

ments serve as observable indicators of dimensions (i.e., latent traits). Because items within a block usually measure different dimensions, they are called multidimensional forced-choice (MFC) items.

2.1 Forced-Choice Test Design

According to Hontangas et al. (2015), forced-choice blocks have three common formats: PICK, RANK, and MOLE, distinguished primarily by their instructions. PICK (Table 1) requires selecting the most characteristic statement from a block. RANK (Table 2) requires complete ranking from most to least characteristic. MOLE (Table 3) requires selecting both the most and least characteristic statements. MOLE formats with more than three items are also called partial rankings.

Table 1. PICK Format
Instructions: Choose the statement that best describes you from the following two options
A. Identify shortcomings in things
B. Explore unfamiliar territories

Table 2. RANK Format
Instructions: Rank the following statements
A. Identify shortcomings in things
B. Explore unfamiliar territories
C. Make decisions based on data analysis

Table 3. MOLE Format
Instructions: Select the statement that is most like you and least like you from the following options
A. Identify shortcomings in things
B. Explore unfamiliar territories
C. Make decisions based on data analysis
D. Perform work requiring precision

Block size refers to the number of statements per block, with 2-4 items being most common. PICK-2 requires only one comparison, while larger blocks increase cognitive complexity, potentially disadvantaging less educated or lower-literacy respondents (Brown, 2016). Common block types include PICK-2 (Oswald et al., 2015), RANK-3 (Lian et al., 2014; SHL, 2018), and MOLE-4 (SHL, 1997). RANK-3 offers a balance—less cognitively demanding than MOLE-4 yet more efficient than PICK-2, providing maximum information (Hontangas et al., 2015; Joo et al., 2018).

Q-Sort (Block, 1963) represents a special forced-choice format where all questionnaire items (e.g., over 30) form one large block. Respondents gradually assign items to a few preference categories, first selecting the most characteristic items, then the least characteristic from the remainder, until all items are classified. This method requires processing many statements simultaneously and

is suitable for vocabulary-type items (Brown, 2016).

When assembling forced-choice tests, matching item desirability is the primary principle and critical for anti-faking effectiveness, followed by block size and instructions. Researchers typically calculate the mean absolute difference in item desirability as a matching index, where larger differences indicate poorer matching. However, this mean-based approach ignores variability across raters. Pavlov et al. (2021) proposed an alternative index, IIA (Inter-item agreement), which incorporates BP and AC indices (Gwet, 2014) into desirability matching, enabling better matching of items without mean differences. Practitioners can use the R package autoFC (Li et al., 2021) to calculate IIA indices and automate test assembly.

2.2 Traditional Scoring and Ipsative Data Issues

2.2.1 Traditional Scoring Traditional forced-choice scoring typically assigns +1 to the most characteristic or highest-ranked item, -1 to the least characteristic or lowest-ranked item, and 0 to unselected or middle-ranked items. Dimension scores are obtained by summing item scores within each dimension.

Item directionality affects scoring: negatively worded items require score reversal (multiplying by -1). For example, a negatively worded item like "I often expect negative outcomes" would receive -1 when selected as most characteristic. Because negatively worded items typically have low desirability, matching desirability across opposite-direction items is difficult. Placing items with large desirability differences in the same block makes it easy for respondents to select the more positive option, especially in high-stakes situations where almost all respondents choose the seemingly positive option (Bürkner et al., 2019). This creates measurement precision issues and undermines anti-faking effectiveness, making mixed-direction blocks rare in practice.

2.2.2 Ipsative Data and Its Problems Using the MOLE-4 format in Table 3 as an example, regardless of how an individual responds, the selected most and least characteristic items receive +1 and -1 respectively, making the sum of item scores within each block zero. Consequently, the total test score also sums to zero. This creates interdependence among dimension scores—high scores on some dimensions necessarily mean low scores on others, preventing all dimensions from being simultaneously high or low. This is ipsative data. In contrast, normative data (e.g., Likert scales) have independent item ratings, allowing variable total scores.

The internal dependency of ipsative data violates a fundamental assumption of classical test theory: independence of error variance. This affects statistical analysis and interpretation of forced-choice scores (Baron, 1996), including reliability analysis, ANOVA, and regression, increasing Type I error rates while reducing statistical power (Wang et al., 2014). Ipsative data also distorts dimension relationships, contaminating structural and criterion-related validity

(Brown & Maydeu-Olivares, 2013), and precludes factor analysis (Closs, 1996). Finally, interpreting ipsative data normatively for inter-individual comparisons is inappropriate, as it may distort true characteristics. In interest inventories, for instance, ipsative results represent only intra-individual preference rankings; direct group comparisons may seriously over- or under-estimate true interest profiles (Closs, 1996).

The number of measured dimensions and their interrelationships substantially affect ipsativity. Bartram (1996) found that with fewer than 10 dimensions or inter-dimension correlations $ $0.3, ipsative scores become unreliable, with reliability decreasing sharply as dimension count drops or correlations increase. Clemans (1966) also noted that low-dimensional forced-choice tests suffer more severe ipsativity. Baron (1996) pointed out that if true scores are uniformly distributed around the mean, ipsative and normative scores will be similar. However, when most dimensions are above or below the mean, they diverge substantially—though this difference decreases with more dimensions, as the probability of multiple dimensions being simultaneously high or low declines. Similarly, when all inter-dimension correlations are highly positive or negative, the likelihood of uniformly high/low scores increases; when correlations are mixed, extreme score patterns become less likely. Saville and Willson (1991) demonstrated that with over 30 dimensions and low inter-dimension correlations, reliability reaches acceptable levels and trait recovery resembles normative data, making normative interpretation feasible. Thus, increasing dimension count is an effective traditional approach to mitigating ipsativity, though only a compromise.

In summary, ipsative data limitations have restricted forced-choice test applications. While increasing dimensions helps, traditional scoring treats ranking results as absolute ratings, failing to capture the psychological decision-making process underlying forced-choice responses. This is fundamentally inappropriate. Resolving ipsativity requires abandoning traditional scoring in favor of modern measurement models that reflect respondents' decision processes (Brown & Maydeu-Olivares, 2013), extracting latent trait scores from observed comparisons to restore normative properties.

3 IRT Scoring Models for Forced-Choice Tests

Over the past decade, numerous IRT scoring models for forced-choice tests have been developed to establish relationships between observed responses and latent traits, yielding normative latent trait scores for inter-individual comparisons. Among the most widely studied are Brown' s (2011) Thurstonian IRT model (TIRT) and Stark et al.' s (2005) MUPP (Multi-Unidimensional Pairwise Preferences) framework, which has spawned two new models (Morillo et al., 2016; P. Lee et al., 2019). Other models include Wang et al.' s (2017) Rasch ipsative model (RIM) and H. Lee and Smith' s (2020a) Bayesian random block IRT model (BRB-IRT) based on Bayesian testlet models (Bradlow, Wainer, & Wang, 1999). These models share three components: forced-choice format, item

response model, and decision model. Their essential differences lie in assumed item response patterns (Morillo et al., 2016) and decision model types. Item response patterns reflect the relationship between response intensity and measured dimensions, while decision models capture the choice process among items. Format and decision model jointly determine the framework, with decision models bridging observed choices to item response intensities, which are then linked to latent traits via item response models. This paper first clarifies different item response patterns and decision model types, then classifies and systematically introduces the models accordingly, and finally compares them across model construction, parameter estimation, and application research.

3.1 Item Response Patterns

Items are observable indicators of traits, with their relationship to latent traits requiring measurement models. In personality assessment, models can be categorized as dominance or unfolding models based on assumed response processes. Dominance models assume that higher trait levels increase the probability of positive responses, with Rasch and 2PL models following this pattern. Unfolding models assume that positive response probability relates directly to the proximity between item location and trait level. For example, for the item "I enjoy quietly chatting with friends in a café," extremely introverted individuals may disagree due to disliking public spaces, while extremely extraverted individuals may also disagree due to preferring more stimulating environments (Drasgow et al., 2010); those at intermediate levels are more likely to agree. The item response function is unimodal, with response probability peaking when trait level matches item location. The representative unfolding model is the Generalized Graded Unfolding Model (GGUM) (Roberts et al., 2000).

Which model better captures non-cognitive response characteristics remains unresolved (Wang et al., 2014; Morillo et al., 2016; Hontangas et al., 2016). Some simulation and empirical studies support unfolding models, particularly for attitude measurement, where unfolding items perform as well as or better than dominance items (Chernyshenko et al., 2001; Tay et al., 2011). Unfolding models are considered more flexible because they reduce to dominance models when item location parameters are extreme. However, this superiority is not universal; scales composed entirely of unfolding items show inferior psychometric properties compared to dominance-based scales, including lower reliability and criterion validity (Huang & Mead, 2014). Additionally, unfolding models cannot directly reverse-score negative items, potentially reducing estimation accuracy (Brown & Maydeu-Olivares, 2010). Dominance models are generally more parsimonious with fewer parameters; unless clear evidence favors complex models, simpler models should be preferred (Oswald & Schell, 2010). Unfolding items are also more difficult to write and interpret. Further discussion can be found in Drasgow et al. (2010).

Item response patterns are item-level characteristics, not trait characteristics. When combining single items into forced-choice blocks, any response pattern can

be used because they measure the same latent traits, whose distributions remain invariant across populations. In practice, researchers should select dominance or unfolding models based on item or data characteristics. No studies have mixed both models within the same test.

3.2 Decision Theory

Forced-choice tests require comparative judgments among items rather than independent evaluations, yet absolute evaluations of individual items form the basis for trait measurement. Following Brown (2016), individuals' comparative judgments are based on their absolute evaluation levels for each item. Modeling forced-choice data requires appropriate decision theory to explain the relationship between decision outcomes (observed responses) and absolute evaluations, thereby assessing latent traits. Two main decision theories have been used: Thurstone' s Law of Comparative Judgment (Thurstone, 1927) and Luce' s Choice Axiom (Luce & Duncan, 1959) with the Bradley-Terry model (Bradley & Terry, 1952), the latter being a special case of the former (Brown, 2016).

3.2.1 Thurstone' s Law of Comparative Judgment Thurstone (1927) used utility to represent response tendency toward each item. Utility is a latent variable representing an item' s psychological value to an individual. Thurstone viewed item evaluation as utility measurement. Let y_{ij} represent the observed outcome of comparing items i and j, where $y_{ij} = 1$ indicates choosing item i as most characteristic, otherwise $y_{ij} = 0$. Let t_i denote the utility of item i, where $t_i > t_j$ means item i has higher utility than j, making i more likely to be chosen. The relationship between utility and observed response can be expressed as:

$$y_{ij}^* = t_i - t_j$$

where y_{ij}^* represents the utility difference. When applied to forced-choice modeling, utility differences can be decomposed into systematic and random components. The systematic component $f(\theta)$ relates to latent trait levels, while the random component is error ε_i, assumed independent across items and normally distributed. Thus:

$$t_i = f(\theta) + \varepsilon_i$$

where θ_a is the individual' s level on trait a measured by item i.

3.2.2 Luce' s Choice Axiom Luce (1959, 1977) extended the Bradley-Terry model for binary choices, using v_i to represent response intensity for item i. For choice set S, the probability of selecting i is proportional to v_i:

$$P(i|S) = \frac{v_i}{\sum_{k \in S} v_k}$$

Luce described ranking as a series of independent best-choice steps: select the most characteristic item from S, then the next most characteristic from the remaining $S-1$, continuing until all items are ranked (Hontangas et al., 2015). The ranking probability is the product of step probabilities. For three items $\{i, j, k\}$, the probability of ranking $i > j > k$ is:

$$P(i, j, k) = P(i|\{i, j, k\}) \times P(j|\{j, k\})$$

where $P(i|\{i, j, k\})$ is the probability of choosing i from $\{i, j, k\}$, and $P(j|\{j, k\})$ is the probability of choosing j from $\{j, k\}$.

When S contains only two items, Luce' s axiom reduces to the Bradley–Terry model:

$$P(i|\{i, j\}) = \frac{v_i}{v_i + v_j}$$

In forced-choice modeling, v_i is derived from item response functions related to latent traits. Other decision theories include Coombs' s Unfolding Preference Model (a special case of Thurstone' s law) and Andrich' s Forced Endorsement Model (equivalent to Bradley–Terry when simplified); see Brown (2016) for details.

3.3 TIRT Model

TIRT (Brown, 2011) is a dominance-item model based on Thurstone' s law, applicable to PICK-2, RANK, and MOLE formats with items from same or different dimensions. TIRT assumes respondents make independent pairwise comparisons among n items in a block, generating $n(n-1)/2$ comparisons. Data must be binary-coded to obtain pairwise results. For a RANK-3 block with items $\{i, j, k\}$ and ranking $i > j > k$, coding yields $y_{ij} = 1$, $y_{ik} = 1$, $y_{jk} = 0$, representing $i > j$, $i > k$, and $j < k$. TIRT is a probabilistic model built on this binary-coded data.

In TIRT, utility relates linearly to latent traits, with each item loading on only one trait (unidimensional). For item i measuring trait a:

$$t_i = \mu_i + \lambda_i \theta_a + \varepsilon_i$$

where μ_i is the mean latent utility, λ_i is the factor loading on θ_a, and $\varepsilon_i \sim N(0, \psi_i^2)$. θ_a follows a multivariate normal distribution. Substituting into the utility difference equation yields:

$$y_{ij}^* = (\mu_i - \mu_j) + (\lambda_i \theta_a - \lambda_j \theta_b) + (\varepsilon_i - \varepsilon_j)$$

Assuming $\varepsilon_i - \varepsilon_j \sim N(0, \psi_i^2 + \psi_j^2)$, the conditional probability of choosing i over j uses the normal ogive link:

$$P(y_{ij} = 1|\theta) = \Phi\left(\frac{\mu_i - \mu_j + \lambda_i\theta_a - \lambda_j\theta_b}{\sqrt{\psi_i^2 + \psi_j^2}}\right)$$

where Φ is the standard normal CDF. Binary-coded data share common items (e.g., y_{ij} and y_{ik} both involve item i), so their covariance is set to ψ_i^2 to account for dependency. This makes TIRT a special two-dimensional normal ogive IRT model following dominance response patterns.

Numerous simulation and empirical studies have examined TIRT' s applicability (Bürkner et al., 2019; Brown & Maydeu-Olivares, 2013; Schulte et al., 2021; Li et al., 2017; Lian et al., 2014). These studies demonstrate that TIRT partially overcomes ipsativity, improving measurement precision and yielding results closer to single-stimulus Likert scales (Joubert et al., 2015). However, TIRT requires restrictive test designs to outperform traditional scoring. For example, with few dimensions, good trait recovery requires mixed-direction blocks (Brown, 2011). Schulte et al. (2021) found that with fewer than 10 dimensions and all positively keyed items, reliability drops sharply even with high factor loadings. Conversely, with many dimensions (>30), TIRT accurately recovers trait scores and relationships without mixed-direction blocks (Schulte et al., 2021; Bürkner et al., 2019). However, mixed-direction blocks have drawbacks (Bürkner et al., 2019; Morillo et al., 2016): increased cognitive load, potential method factors from negative wording, and reduced anti-faking effectiveness.

3.4 MUPP Framework and Models

3.4.1 MUPP Framework and MUPP-GGUM Model Stark (2005) proposed the MUPP framework for pairwise preference (PICK-2) formats, significantly influencing later forced-choice model development (Brown & Maydeu-Olivares, 2013). For a block with items i and j measuring traits θ_a and θ_b, let $P(i)$ be the probability of endorsing item i and $Q(i)$ the probability of rejecting it, with $P(i) + Q(i) = 1$. The probability of choosing i as most characteristic is:

$$P(i \text{ over } j) = \frac{P(i)Q(j)}{P(i)Q(j) + P(j)Q(i)}$$

MUPP assumes independent evaluation of each item with unidimensional items that can measure same or different dimensions—hence "Multi-Unidimensional Pairwise Preference."This reflects the relationship between decision probabilities and single-item tendencies, using a Bradley–Terry decision model (Brown, 2016).

Stark (2005) assumed unfolding response patterns and used the binary version of GGUM to calculate $P(i)$ and $Q(j)$ in the above equation, creating the MUPP-GGUM model. Stark (2002, 2005) provided a recommended process for building

PICK-2 tests: (1) write many statements per dimension (3× target number); (2) administer them as 4- or 5-point Likert scales to ~1000 respondents per dimension; (3) estimate item parameters and test unidimensionality; (4) rate social desirability and use mean ratings; (5) pair items with similar desirability from different traits, including some same-dimension pairs to set the latent trait metric; (6) administer the forced-choice test; (7) estimate traits using MUPP-GGUM.

MUPP-GGUM is one of the oldest and most widely used forced-choice models, with standardized development procedures and pioneering applications in CAT for personality assessment, including U.S. military selection tests (Stark et al., 2012; Stark et al., 2014). Subsequent MUPP-based derivatives expanded to various formats and advanced forced-choice CAT research.

3.4.2 MUPP-2PL Model Morillo et al. (2016) argued that dominance items are also suitable for non-cognitive tests, offering advantages in item writing ease and model parsimony. They replaced the GGUM functions in MUPP with the classic 2PL model, creating MUPP-2PL. The probability of choosing item i over j becomes:

$$P(i \text{ over } j) = \frac{\exp(a_i\theta_a + d_{\text{block}})}{\exp(a_i\theta_a + d_{\text{block}}) + \exp(a_j\theta_b + d_{\text{block}})}$$

where L is the logistic function, a_i and a_j are discrimination parameters, θ_a and θ_b are latent traits, and d_{block} is an intercept combining 2PL parameters (though individual item b parameters are not identifiable).

Usami et al. (2016) also applied 2PL in MUPP but used pre-calibrated item parameters like Stark (2005). While convenient for algorithm and item bank management, personality tests lack correct answers and rarely require large item banks for parallel forms. Estimating parameters directly from forced-choice data is more realistic (P. Lee et al., 2019). Stark' s method also ignores parameter variability across contexts and estimation errors when estimating traits. Therefore, Morillo et al. (2016) used Bayesian MCMC for joint estimation of item and person parameters from forced-choice data. They found that test length affected parameter and trait recovery (longer tests yield better accuracy), and sample size importantly influenced item parameter estimation, with d_{block} estimated more accurately than a parameters. Empirically, MUPP-2PL showed some trait relationship differences from previous research, though whether this stems from sample or context changes remains unclear.

3.4.3 GGUM-RANK Model MUPP-GGUM and MUPP-2PL only accommodate PICK-2 formats. Hontangas et al. (2015) extended MUPP using Luce' s choice axiom to handle PICK, RANK, and MOLE formats. For three items $\{i, j, k\}$ in PICK-3 format:

$$P(i|\{i,j,k\}) = \frac{P(i)}{P(i)+P(j)+P(k)}$$

For RANK formats, ranking is conceptualized as a series of PICK decisions. For RANK-3 with result $i > j > k$:

$$P(i,j,k) = P(i|\{i,j,k\}) \times P(j|\{j,k\})$$

For MOLE-4 (adding item l), the two unselected items' order is unknown, so both possible rankings are combined. Let $P(i,k^{**})$ represent the probability of selecting i as most and k as least characteristic:

$$P(i,k^{**}) = P(i,j,l,k) + P(i,l,j,k)$$

where probabilities are calculated using the above logic.

This Luce-based extension integrates PICK, RANK, and MOLE formats into a nested framework, greatly expanding MUPP' s applicability. P. Lee et al. (2019) developed the GGUM-RANK model for RANK-3 formats (using GGUM for $P(i)$ in the equations) with MCMC joint estimation. Joo et al. (2018) developed two information indices: OII (Overall item information) for a block and OTI (overall test information) as the sum across blocks. These indices guide test assembly, with conditional OII plots enabling selection of blocks providing maximum information in target ability ranges, laying groundwork for GGUM-RANK CAT (Joo et al., 2020).

3.5 RIM Model

Wang et al. (2017) argued that identifying absolute latent trait levels from forced-choice tests is unrealistic, suggesting TIRT-derived trait scores cannot be used for intra- or inter-individual comparisons. They proposed RIM to obtain scores for intra-individual comparison rather than absolute trait scores, using the Rasch model as the item response function (suitable for dominance items). Like TIRT, RIM uses Thurstone's (1927) comparative judgment law, where item comparison reflects trait differences relative to item utilities. The probability of choosing i over j is:

$$P(i \text{ over } j) = \frac{\exp(\theta_a - \mu_i)}{\exp(\theta_a - \mu_i) + \exp(\theta_b - \mu_j)}$$

where θ_a and θ_b are latent traits for items i and j, and μ_i, μ_j are item utilities.

In trait estimation, the sum of all trait scores within an individual is fixed at zero, leaving only $D-1$ traits freely estimated (where D is the number of dimensions). Here, θ represents psychological trait differentiation—values

near zero indicate lower differentiation. Thus, θ values represent finer-grained internal rankings than traditional ipsative scoring. The absolute difference $|\theta_a - \theta_b|$ indicates relative differentiation between traits a and b within an individual. For parameter estimation, Wang et al. recommend MMLE for fewer than 4 dimensions and MCMC for higher dimensions.

Wang et al. (2016) extended RIM to RANK formats, creating ELIRT (exploded logit IRT) and GLIRT (generalized logit IRT). ELIRT' s extension logic matches Hontangas et al.'s (2015) RANK approach, while GLIRT enumerates all possible response patterns for each block, writes response functions for each, and constrains their sum to 1 (see Chen et al., 2020). For pairwise formats, ELIRT and GLIRT are equivalent to RIM. Simulation studies show similar results, allowing researchers to choose either.

3.6 BRB-IRT Model

H. Lee and Smith (2020a) used the Bayesian testlet model (Bradlow et al., 1999) as a foundation, adding a random block effect parameter γ_n to MUPP-2PL' s item response function (similar to testlet effects) to account for within-block item dependencies—creating BRB-IRT. Like TIRT, BRB-IRT supports multiple formats and requires binary coding for RANK-3, making its decision theory classifiable as Thurstonian.

The probability of person m choosing item i over j in block n is:

$$P(i \text{ over } j) = \frac{\exp(a_i\theta_a + d_{\text{block}} + \gamma_{nm})}{\exp(a_i\theta_a + d_{\text{block}} + \gamma_{nm}) + \exp(a_j\theta_b + d_{\text{block}} + \gamma_{nm})}$$

where block n contains 2+ items measuring different dimensions. The random block effect γ_{nm} represents the block's dimensional influence on responding, analogous to testlet effects from shared stimuli (e.g., reading passages) in traditional testlet models. BRB-IRT uses MCMC estimation like Bayesian testlet models. Simulation studies yielded TIRT-like recommendations: mixed-direction blocks are needed for reliable parameter estimation, though only 3-dimension scenarios were tested—high-dimensional performance is unknown. Random block effect size did not affect parameter estimates.

Regarding concerns that mixed-direction blocks reduce anti-faking effectiveness, H. Lee and Smith suggest BRB-IRT suits low-stakes contexts where leveraging forced-choice benefits (avoiding Likert biases) without high anti-faking demands is valuable. For example, PISA 2012 used forced-choice formats for mathematics intention and learning strategy scales to control cross-cultural response biases when examining international differences. BRB-IRT can flexibly include covariates affecting items and trait scores to better analyze group differences. With covariates affecting all traits (e.g., gender):

$$P(i \text{ over } j) = \frac{\exp(a_i\theta_a + d_{\text{block}} + \gamma_{nm} + \beta_{\text{gender}})}{\exp(a_i\theta_a + d_{\text{block}} + \gamma_{nm} + \beta_{\text{gender}}) + \exp(a_j\theta_b + d_{\text{block}} + \gamma_{nm} + \beta_{\text{gender}})}$$

With trait-specific covariates:

$$P(i \text{ over } j) = \frac{\exp(a_i(\theta_a + \beta_{\text{gender},a}) + d_{\text{block}} + \gamma_{nm})}{\exp(a_i(\theta_a + \beta_{\text{gender},a}) + d_{\text{block}} + \gamma_{nm}) + \exp(a_j(\theta_b + \beta_{\text{gender},b}) + d_{\text{block}} + \gamma_{nm})}$$

These allow examining whether gender influences item choices or trait-specific responding.

4 Model Comparison

4.1 Model Construction Rationale

From a practical perspective, forced-choice model development aims to accommodate more block formats (PICK, RANK, MOLE) and different response patterns. Table 4 summarizes existing models by format and response model.

Table 4. Model Summary

Format	Unfolding Response Model	Dominance Response Model
PICK-2	MUPP-GGUM	TIRT/MUPP-2PL/BRB-IRT
RANK-3	GGUM-RANK	TIRT/BRB-IRT/ELIRT/GLIRT
MOLE-4	GGUM-RANK	TIRT/BRB-IRT/ELIRT/GLIRT

TIRT, MUPP-2PL, and BRB-IRT all suit dominance items using 2PL, though MUPP-2PL only handles PICK-2 while the others apply to multiple formats via binary coding. For PICK-2, TIRT and MUPP-2PL are theoretically equivalent (Morillo et al., 2016), differing only in link functions (probit vs. logit). Simulation studies confirm their equivalence, though MUPP-2PL shows slightly better trait and trait-relationship recovery. Empirically, MUPP-2PL and TIRT estimates correlate near 0.9, supporting the internal equivalence of Thurstonian and Bradley–Terry approaches for PICK-2, though TIRT' s lack of prior information yields more extreme estimates. BRB-IRT adds random block effects to MUPP-2PL to account for within-block dependencies, while TIRT uses a covariance matrix; empirical studies show high consistency between BRB-IRT and TIRT.

RIM also supports dominance items but uses the Rasch model and differs in trait score origin interpretation. Unlike TIRT, MUPP-2PL, and BRB-IRT, which interpret θ as normative trait scores with multivariate normal distributions, RIM

views θ as intra-individual differentiation degree, fixing within-person θ sums to zero without population distribution assumptions. Thus, RIM suits tests aiming to identify intra-individual trait rankings, while the others suit inter-individual comparisons. Whether RIM' s advantages over traditional ipsative scoring justify its complexity requires further research.

4.2 Parameter Estimation Methods

Parameter estimation involves item parameters and latent traits, using traditional algorithms or MCMC, with joint or two-step estimation strategies.

Among the six models, only MUPP-GGUM uses two-step estimation: item parameters for $P(i)$ and $Q(i)$ are pre-calibrated from Likert data (steps 2-3), then used with different forced-choice data for ability estimation (step 7). This implicitly assumes parameter invariance across test formats, facilitating item bank management and CAT development. For ability estimation, Stark et al. (2005, 2012) used BFGS (Broyden-Fletcher-Goldfarb-Shanno) for MAP estimation in high dimensions, avoiding complex Hessian matrix derivations. BFGS can be implemented via DFPMIN (Press et al., 1986) or R' s `optim` function with `method="L-BFGS-B"`. For item parameter calibration, recent GGUM estimation advances are supported by R packages GGUM (Tendeiro & Castro-Alvarez, 2018), mirt (Chalmers, 2012), and bmggum (Tu et al., 2021).

TIRT, developed within SEM, has mature software support (e.g., Mplus) and the R package thurstonianIRT (Bürkner, 2018). Brown and Maydeu-Olivares (2012) provided an Excel macro generating Mplus code from test specifications. The thurstonianIRT package interfaces with lavaan (Rosseel, 2012), Mplus, or Stan (Stan Development Team, 2020), automatically generating code for all three methods. Item parameters can be estimated via unweighted or diagonally weighted least squares. Stan uses MCMC for Bayesian models, providing a convenient TIRT interface. Trait estimation uses EAP (for 1-2 dimensions) or MAP (for higher dimensions) because EAP' s numerical integration nodes increase exponentially with dimensions (Brown, 2016).

While TIRT' s software support facilitates its widespread use, concerns exist. Bürkner et al. (2019) found severe convergence issues with Mplus and lavaan for large tests (e.g., 5 dimensions $\times$ 27 blocks yielded ~30% convergence). High memory demands are also problematic (e.g., 30 dimensions $\times$ 9 blocks required 32GB RAM), sometimes necessitating omission of fit indices to reduce computational load. Negative variance errors are common, requiring constraints on dimension relationships or factor loadings to promote convergence, though results then depend heavily on these fixed values. MCMC estimation avoids convergence and memory issues due to Bayesian advantages. However, Mplus' s unweighted least squares (a limited-information method) is often faster than Stan by several-fold, making it recommended for non-large-scale applications when convergence is achievable.

Unlike TIRT, newer models exclusively use MCMC—a full-information, prob-

abilistic method requiring no complex derivations, only reasonable posterior specification, achieving precision similar to frequentist methods. MUPP-2PL, GGUM-RANK, RIM, and BRB-IRT all use Metropolis-Hastings MCMC for joint parameter estimation, with similar prior specifications but different software: GGUM-RANK uses Ox (Doornik, 2009), BRB-IRT uses OpenBUGS 3.2.3 (Lunn et al., 2009), and RIM uses WinBUGS (Spiegelhalter et al., 2003) or JAGS (Plummer, 2003). For fewer than 4 dimensions, RIM recommends MMLE in ConQuest (Adams et al., 2015) or R package TAM (Kiefer et al., 2016). WinBUGS and OpenBUGS are relatively slow, while Bürkner et al.' s (2019) Stan implementation for TIRT uses advanced NUTS/HMC sampling for faster estimation. Convergence is assessed via Gelman and Rubin' s (1992) $\hat{R}$ statistic (<1.2 indicates convergence). While these models have fewer convergence problems, they require deeper MCMC knowledge and suffer from long estimation times (e.g., BRB-IRT took 6 days for one simulation condition with 1000 respondents, 3 dimensions, 8 RANK-3 blocks, and 25 replications).

Table 5 summarizes parameter estimation methods.

Table 5. Parameter Estimation Method Summary

Model	Estimation Content	Algorithm	Software	Advantages	Disadvantages
MUPP-GGUM	Two-step: pre-calibrate items from Likert data; estimate ability from forced-choice data	BFGS for ability; MMLE/MAP for items	DFPMIN/R stats::optim	Facilitates adaptive test management	Risk of parameter non-invariance across formats
TIRT	Joint estimation of items and traits	Weighted least squares/DWLS MCMC	Mplus; R:thurstonianIRT (Mplus/Lavaan/Stan)	Fast (Mplus); Easy to use	Poor convergence in high dimensions; High memory; May need to sacrifice fit indices
MUPP-2PL/GGUM-RANK/RIM/BRB-IRT	Joint estimation	MCMC	Ox/WinBUGS/JAGS/OpenBUGS; R:thurstonianIRT (Stan)	Good flexibility	Long estimation time; Steep learning curve

5 Application Research Status

Forced-choice IRT models are widely applied in industrial-organizational psychology. TIRT is used in commercial tests like OPQ32r and CCSQ personality tests (SHL, 2018; Brown & Maydeu-Olivares, 2011) and assessments of maladaptive personality traits (Guenole et al., 2016). Forced-choice formats with TIRT scoring show better structural and convergent validity than Likert scales in 360-degree feedback (Brown et al., 2017). MUPP-GGUM is used in the Adaptive Employee Personality Test (Adept-15) and U.S. military' s TAPAS (Stark et al., 2014), representing breakthroughs in forced-choice CAT. Parameter invariance testing methods are being developed, and substantial validity evidence has accumulated. This section reviews applied research in parameter invariance testing, CAT, and validity studies.

5.1 Parameter Invariance Testing

Test developers must examine parameter invariance (measurement consistency) to ensure all respondents interpret items similarly. For forced-choice tests, invariance has two aspects: cross-block consistency and cross-population consistency.

Cross-block consistency examines whether an item's parameters remain invariant when paired with different items (e.g., item A in blocks $\{A, B, C\}$ vs. $\{A, D, E\}$). Lin and Brown (2017) tested TIRT invariance across RANK-3 and MOLE-4 formats sharing 75% common items, finding only small bias for few items.

Cross-population consistency (differential item functioning, DIF) examines whether parameters vary across groups (e.g., gender, culture, stakes). Items showing DIF are influenced by background factors, reducing validity and fairness. Forced-choice test development requires ensuring single-statement items have acceptable psychometric properties and no DIF (Stark et al., 2005; SHL, 2018). However, combining items into blocks may create new DIF due to changed context. Thus, DIF testing with forced-choice data is essential.

H. Lee and Smith (2020b) proposed using multiple-group CFA fit index differences to test TIRT measurement invariance, suggesting critical values for metric and scalar non-invariance, though this doesn' t identify specific items. P. Lee et al. (2020) developed omnibus Wald tests for TIRT' s discrimination and intercept parameters, showing high detection rates and near-nominal Type I error in simulation studies. Qiu and Wang (2021) proposed three RIM DIF methods (EMD, AOS, CS), finding CS superior when DIF items are present.

5.2 Computerized Adaptive Testing

Given personality' s complexity, assessments often measure many dimensions (e.g., OPQ32r' s 32 traits), requiring many items and long tests. Lengthy tests increase fatigue and careless responding, particularly in hiring contexts, damaging corporate image. CAT addresses this by administering fewer items when

acceptable precision is reached, focusing remaining items on less certain dimensions, thereby improving efficiency and reducing costs.

Forced-choice CAT was applied in U.S. Navy selection 15 years ago in the Navy Computer Adaptive Personality Scales (NCAPS), measuring 19 traits via unidimensional PICK-2 blocks paired by desirability (Houston et al., 2006). Stark et al. (2012) proposed a 6-step CAT procedure for multidimensional PICK-2 under MUPP-GGUM, differing from traditional CAT by requiring unidimensional block proportions and pre-storing multidimensional block combinations. For a 3-dimension test, combinations include 1-1, 2-2, 3-3, 1-2, 1-3, 2-3, with content balance controlled via circular dimensional linking (e.g., 1-2, 2-3, 3-4, 4-5, 5-1 for 5 dimensions). Both studies showed CAT halved test length while maintaining accuracy. TAPAS, another U.S. military adaptive personality test, also uses MUPP-GGUM (Stark et al., 2014).

Recent work includes Joo et al.'s (2020) GGUM-RANK CAT for RANK formats, finding unidimensional blocks may not be necessary (contrasting with Stark et al.' s (2012) recommendation of 5% unidimensional blocks). Chen et al. (2020) proposed three subpool selection strategies to improve efficiency and control item exposure: Sequential Strategy (pre-building all dimension combinations), Multinomial Strategy (random subpool selection via multinomial distribution), and High-SE Strategy (selecting subpools for dimensions with highest standard errors). Compared to full-bank selection (6.72s), subpool strategies reduced time to <1s with minimal precision loss, though Sequential required more blocks and High-SE showed poorer content balance, making Multinomial optimal.

For item exposure control, Chen et al. (2020) proposed RSHO (revised Sympson-Hetter online) method. During block selection, calculate $P(S)$ (proportion in bank) and $P(A)$ (proportion already administered) for each item, compare with maximum exposure rate r, and compute p_{ks}. For a block, use the minimum p_{ks} as p_k, then draw a random number to decide selection. RSHO controls exposure with slight precision sacrifice.

Regarding CAT test-retest reliability, Seybert and Becker (2019) noted that item inconsistency introduces error, making CAT retest reliability more like alternate-forms reliability. Forced-choice CAT retest reliability is lower than Likert scales but comparable to Likert alternate-forms reliability.

5.3 Validity Research

Four research directions address whether forced-choice IRT latent trait scores reflect true characteristics. First, studies compare forced-choice IRT scoring to traditional scoring in recovering traits and relationships (Hontangas et al., 2015; Hontangas et al., 2016; Oswald et al., 2015). Almost all find significant precision improvements over traditional scoring, boosting confidence in forced-choice IRT development. However, Wang et al. (2017) offer alternative θ interpretations, and Schulte et al. (2021) note TIRT doesn' t always leverage IRT advantages, yielding partially ipsative scores even in high dimensions. Walton et al. (2020)

found TIRT showed poorer discriminant validity than traditional ipsative scoring on Big Five measures. The extent to which these scores can be treated as normative for selection or criterion analysis requires more research.

Second, studies examine relationships between forced-choice IRT and Likert latent trait scores (Zhang et al., 2020; Watrin et al., 2019; Joubert et al., 2015; Guenole et al., 2016), treating Likert scores as "true" values. High similarity in origin, scale, and dimension relationships would support equivalence and justify similar analyses. Third, research explores anti-faking effectiveness. When desirability is matched within blocks, forced-choice tests outperform Likert scales (Wetzel et al., 2020). Unlike TIRT, GRM analysis of Likert scales cannot effectively distinguish high-ability respondents because fakers tend to select extreme responses, reducing high-ability item discrimination (Dueber et al., 2018). Fourth, studies examine forced-choice IRT in non-self-report contexts. Rater biases in Likert-based 360-degree assessments reduce inter-rater reliability. Forced-choice IRT improves inter-rater reliability and structural validity across rater levels compared to Likert scales (Brown et al., 2017).

6 Research Outlook

As an effective format for resisting faking and response biases while improving efficiency, forced-choice tests and IRT models hold substantial potential, particularly for non-cognitive, high-stakes assessments. Based on unresolved issues, we propose four future directions: model expansion, parameter invariance research, forced-choice CAT research, and validity research.

6.1 Model Expansion

Existing models accommodate standard formats (PICK-2, RANK-3). Future research could explore adapting them to Q-Sort formats via recoding. Another variant is the multi-scored PICK-2 format used in Adept-15 (Aon Hewitt, 2015), where respondents select the most characteristic item and rate their confidence (Table 6), expanding from 2 to 4 score points but increasing cognitive load, limiting use to small blocks. No direct models exist for this format, warranting investigation.

Table 6. Multi-Scored PICK-2 Format

A. Identify shortcomings in things	
B. Explore unfamiliar territories	

6.2 Parameter Invariance Research

Building on Lin and Brown' s (2017) TIRT cross-block consistency study, research is needed on whether invariance holds with lower common-item proportions and for other models. Currently, only TIRT (H. Lee & Smith, 2020b; P. Lee et al., 2020) and RIM (Qiu & Wang, 2021) have DIF detection methods.

Future work should develop DIF methods for other models and improve existing methods' sensitivity to various DIF sources.

6.3 Forced-Choice CAT Research

While forced-choice CAT has empirical experience, current procedures use item parameters pre-calibrated from single-stimulus data and item banks (not block banks), forming blocks during selection. The impact of cross-block inconsistency on trait estimation in this context needs study. High dimensions increase block combinations and test length, challenging content balance and efficiency. Future research should explore maximizing CAT advantages in high dimensions. While Chen et al.' s (2020) subpool strategies and exposure control are format-agnostic, their performance with other models requires validation. Multinomial strategies don' t directly apply to variable-length tests, necessitating alternative approaches.

6.4 Validity Research

Many studies compare forced-choice and Likert results to demonstrate anti-faking effectiveness and normative score recovery. However, format differences and Likert biases introduce error. Future research should maximize control over these biases or develop better validity approaches. Larger blocks increase anti-faking effectiveness but also cognitive load (Wetzel et al., 2020); research should identify the optimal balance. Most validity research focuses on TIRT; new models like GGUM-RANK require similar investigation.

References

[The references section contains the full bibliography as provided in the original text, preserved exactly with all citations and formatting intact.]

Note: Figure translations are in progress. See original paper for figures.