Received: 4 August 2021 | Accepted: 7 March 2023

DOI: 10.1111/bmsp.12303

ARTICLE

A dual process item response theory model for polytomous multidimensional forced-choice items

Xuelan Qiu[1] | Jimmy de la Torre[2]

[1]Institute for Learning Sciences & Teacher Education, Australian Catholic University, Brisbane, Queensland, Australia

[2]Faculty of Education, The University of Hong Kong, Hong Kong, China

Correspondence
Xuelan Qiu, Institute for Learning Sciences & Teacher Education, Australian Catholic University, Brisbane, Qld 4000, Australia.
Email: shqiu@acu.edu.au

Funding information
Faculty Research Fund, Faculty of Education, The University of Hong Kong

Abstract

The use of multidimensional forced-choice (MFC) items to assess non-cognitive traits such as personality, interests and values in psychological tests has a long history, because MFC items show strengths in preventing response bias. Recently, there has been a surge of interest in developing item response theory (IRT) models for MFC items. However, nearly all of the existing IRT models have been developed for MFC items with binary scores. Real tests use MFC items with more than two categories; such items are more informative than their binary counterparts. This study developed a new IRT model for polytomous MFC items based on the cognitive model of choice, which describes the cognitive processes underlying humans' preferential choice behaviours. The new model is unique in its ability to account for the ipsative nature of polytomous MFC items, to assess individual psychological differentiation in interests, values and emotions, and to compare the differentiation levels of latent traits between individuals. Simulation studies were conducted to examine the parameter recovery of the new model with existing computer programs. The results showed that both statement parameters and person parameters were well recovered when the sample size was sufficient. The more complete the linking of the statements was, the more accurate the parameter estimation was. This paper provides an empirical example of a career interest test using four-category MFC items. Although some aspects of the model (e.g., the nature of the person parameters) require additional validation, our approach appears promising.

KEYWORDS
cognitive model of choice, item response theory, multidimensional forced choice, psychological differentiation

1 | INTRODUCTION

The use of multidimensional forced-choice (MFC) items in non-cognitive tests to measure traits such as career interests, values and personality has a long history (Johnson et al., 1988). Take career interest tests as an example. A typical MFC item is 'Which activity do you prefer: visiting museums or attending parties?', which pairs activities measuring artistic and social interests and presents them in a single item. Respondents are required to select their preferred activity from the two statements. There are other types of MFC items with more than two statements. For example, in multidimensional ranking items, respondents are asked to rank several (e.g., four) statements according to their levels of preference for the statements. MFC items have been found to effectively reduce response bias (e.g., acquiescence or social desirability) and detect faking (Murphy et al., 1993; Salgado et al., 2015). Popular tests using MFC items include the Jackson Vocational Interest Survey (JVIS; Jackson, 1977), the Edwards Personal Preference Schedule (EPPS; Ashman & Telfer, 1983), the Career Interest Test (CIT; Bartlett et al., 2016), the Allport–Vernon–Lindzey Study of Values (SOV; Kopelman et al., 2003), and the ipsative Occupational Personality Questionnaire (OPQ; SHL, 2006).

Multidimensional forced-choice items are usually scored dichotomously, with 1 being assigned to the statement preferred by the respondent and 0 to the non-preferred statement (analogous to 1 for correct response and 0 for incorrect responses in the context of ability testing). Such items are henceforth referred to as *dichotomous* MFC items. For the example item above, the respondent gets a score of 1 if s/he selects 'visiting museums', and a score of 0 otherwise. This scoring method results in scores with a unique feature referred to as ipsative (from the Latin *ipse*: he, himself), where each individual has identical summed scores (Cattell, 1944; Meade, 2004). Ipsative scores allow comparisons of different latent traits within a person, but not between persons (Meade, 2004). For example, these scores can be used to compare a student's distinct career interests (e.g., 'Mary has a higher level of social interest than artistic interest'). However, ipsative scores cannot be used to compare one student's career interest with another's (i.e., it would be incorrect to say 'Mary has a higher level of social interest than John does'). Thus, researchers are sometimes reluctant to use MFC items in their tests (Matthews & Oddy, 1997).

Recently, there has been a surge of research interest in developing item response theory (IRT) models to analyse MFC items. Such models can be broadly categorized into two frameworks: dominance and ideal point. Models developed within the dominance framework assume that the higher a person's trait level is and the more attractive a statement is, the greater the probability of the person selecting the statement. Models in the dominance framework include Thurstonian IRT models (Brown & Maydeu-Olivares, 2011, 2013) and the Rasch ipsative models (RIMs; Wang et al., 2016; Wang et al., 2017). The ideal point approach assumes that the closer locations between a person's trait and a statement's utility (attractiveness) are, the greater the probability of the person selecting the statement. Examples of models developed within the ideal point framework are the multi-unidimensional pairwise-preference model for multidimensional pairwise comparison items and ranking items (Hontangas et al., 2015; Stark et al., 2005).

Despite these attempts, nearly all existing IRT models are for dichotomous MFC items. However, MFC items with more than two categories are used in real tests. For instance, Brown and Maydeu-Olivares (2018) described MFC items that were used in their testing, which require respondents to indicate their degree of preference for one statement over another by selecting one of four categories. Figure 1 shows two items that are similar in form to those discussed by Brown and Maydeu-Olivares (2018). Another example appears in Part I of the SOV (Kopelman et al., 2003), which requires respondents to distribute three points over a pair according to the degree to which they prefer one statement over another. This yields four response patterns, namely (3,0), (2,1), (1,2), and (0,3). The MFC items in these applications yield more than two categories of responses and are henceforth referred to as *polytomous* MFC items. This item format is conceived to inherit the advantages of both MFC items and polytomous items. Specifically, the format consists of statements that usually have similar levels of social desirability, making it more powerful to resist the response bias than the Likert-type format. At the same time, it provides information about preference and intensity of the preference, and thus is more informative than the dichotomous MFC format which provides preference information only. Thus, the item format is particularly useful in

	I prefer A *much more* (j = 3)	I prefer A *a little more* (j = 2)	I prefer B *a little more* (j = 1)	I prefer B *much more* (j = 0)
1. (A) Visiting museums (B) Attending parties	○	○	○	○
2. (A) Playing sports activities (B) Doing scientific experiments	○	○	○	○

FIGURE 1 Examples of multidimensional forced-choice items that produce polytomous responses.

those testing contexts in which researchers are more interested in the question 'how strong are the preferences?' than in 'which do participants prefer?'

Polytomous MFC items are less well understood. To the best of our knowledge, only one model (Brown & Maydeu-Olivares, 2018), which is an extension of the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011) for ordinal preference data, treats the categories as ordinal responses. Thus, the present study aims to develop a new IRT model for polytomous MFC items based on a cognitive modelling approach. The new model and Brown and Maydeu-Olivares's model are similar in the sense that both models are developed within the dominance framework, but are different in many other important aspects. Analogous to their Thurstonian IRT model for dichotomous MFC items (Brown & Maydeu-Olivares, 2011), the new model can recover the normative or absolute levels of latent traits, albeit relying on the items consisting of both positively and negatively keyed statements. By contrast, the dual process IRT model (DPM) aims to estimate the relative levels of latent traits that represent construct differentiation, without requiring special design for paired statements in items. Moreover, as its name implies, the DPM assumes dual cognitive processes underlying preferential choice behaviours, whereas Brown and Maydeu-Olivares's model does not.

The DPM is a type of tree-based IRT model (Böckenholt, 2012; De Boeck & Partchev, 2012) which has been applied to study the judgement or choice behaviours associated with Likert-type scales. However, the DPM focuses on preferences when respondents are presented with multiple statements, whereas previous tree-based models are mainly concerned with preferences for a single item. To the best of our knowledge, the tree-based models have never been applied to study choice behaviours in MFC items.

The rest of this paper is organized as follows. In Section 2 the new model is introduced and formulated. In Section 3 parameter estimation for the new model using the freeware Just Another Gibbs Sampler (JAGS; Plummer, 2003) is presented. Section 4 reports on a series of simulations conducted to assess parameter recovery under various combinations of conditions, and summarizes the results. In Section 5 an empirical example is provided to demonstrate the implications and applications of the new model. Section 6 concludes with a summary and discussion.

2 | THE DUAL PROCESS IRT MODEL FOR POLYTOMOUS MFC ITEMS

The cognitive model of choice for decision-making describes the cognitive processes underlying human preferential choice behaviours. This study specifically uses the multialternative decision field theory (MDFT; Roe et al., 2001), because it can uniquely explain the dynamic and sequential process of choosing from more than two options. In the MDFT, a person's preference for each alternative evolves by focusing on the most important attribute of the options and evaluating the specific aspects of this attribute. The evaluation depends on whether some options have a similar or higher utility for the attended attribute. Then the evaluation switches to another less important attribute and compares the aspects relevant to this second attribute based on the previous preference.

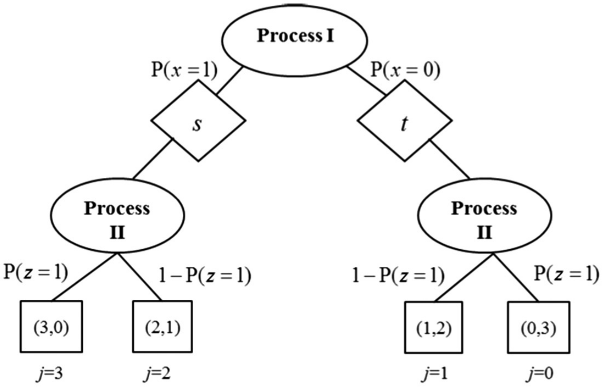

FIGURE 2 Diagram of dual process model for a four-category multidimensional forced-choice item.

Based on the MDFT, when they encounter a polytomous MFC item, respondents are assumed to first focus on the most important attribute of the options (e.g., do the statements describe artistic or social activities?) and choose the statement they prefer from the pair based on their evaluation of the attribute. They then decide how much they prefer the chosen statement (e.g., the specific activity). These two processes are referred to as the choice process and the judgement process, respectively. Figure 2 illustrates the dual process for polytomous MFC items with four category responses. Since this study focuses on MFC items, for which ties are not allowed, the items always have even numbers of response categories. The initial preference responses (denoted by *s* or *t*) are indicated in rhombuses, suggesting which statement is preferred in Process I (i.e., the choice process). At this stage, only dichotomous responses are observed. In Process II (i.e., the judgement process), for the preferred statement, a further decision is made in terms of the intensity of the preference; the outcomes are indicated in squares. For example, when the initial preference response is *s*, and a respondent decides that he or she prefers statement *s much more* than statement *t*, the observed response will be (3,0). Collectively, the category $j = 3$ will be selected.

Mathematically, responding to a polytomous MFC item can be formulated as a hierarchy of statistical models for the two processes. An IRT model with dichotomous MFC items can be used for Process I (choice), and a conventional IRT model with dichotomous or Likert-type items can be used for Process II (judgement). This study uses the RIM (Wang et al., 2017) for Process I because the RIM was developed for multidimensional pairwise comparison (MPC) items and possesses the measurement property of specific objectivity (Whitely & Dawis, 1974). Specifically, let $x_{ni} = 1$ denote that person *n* prefers statement *s* over statement *t* for item *i* with $\{s, t\}$; then the probability of $x_{ni} = 1$ under the RIM is

$$p(x_{ni} = 1|\theta_n, \delta_{si}, \delta_{ti}) = \frac{\exp((\theta_{an} + \delta_{si}) - (\theta_{bn} + \delta_{ti}))}{1 + \exp((\theta_{an} + \delta_{si}) - (\theta_{bn} + \delta_{ti}))}, \tag{1}$$

where δ_{si} and δ_{ti} represent the overall utilities (attractiveness) of statement *s* and statement *t*, respectively, in item *i*; θ_{an} and θ_{bn} are the relative levels of latent trait *a* measured by statement *s* and latent trait *b* measured by statement *t*, respectively, for person *n*; and $\boldsymbol{\theta}_n^{\mathrm{T}} = (\theta_{an}, \theta_{bn})$. The θ variables represent the deviations of each latent trait from the mean of D latent traits, with an extreme value representing a greater differentiation level than a value close to zero. Thus, these variables reflect individuals' degree of construct differentiation (Witkin et al., 1979) such as personality differentiation (Harris et al., 2005), interest differentiation (Hirschi, 2009) and emotion differentiation (Barrett, 2004).

A unique and important property pertaining to the θ variables is worth noting. According to the definition of ipsative measure, the sum of the trait scores in MFC items is a constant C for every person (Meade, 2004), where C can be zero or a non-zero constant. Therefore, for *d*th dimensional trait score, it equals C minus the sum of the remaining $(D-1)$ trait scores. Without loss of generality, it is assumed that the *d*th dimension is the last (i.e., the *D*th) dimension. As shown in Appendix S1 in the Supporting

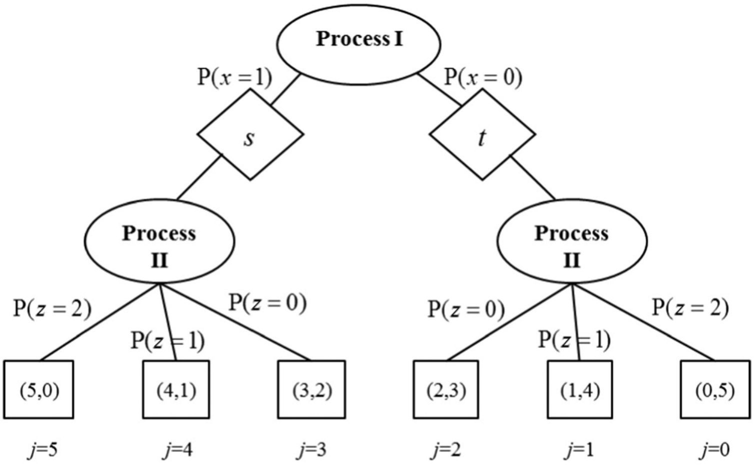

FIGURE 3 Dual process model for a six-category multidimensional forced-choice item.

Information, the value of C will affect the expected value (i.e., the scale origin) of the Dth trait only, but will not affect its variance and covariance with other traits. As in the IRT literature where zero is usually arbitrarily chosen as the scaling origin, C is set to zero for the θ variables in this study. Hence, $\theta_{Dn} = -\sum_{d=1}^{D-1}\theta_{dn}$ for person n.

In Process II, let z_{nij} be the outcome j of judging the preference intensity of person n for item i. For example, for a four-category item, as shown in Figure 2, there are only two outcomes for z_{nij} in this process. These outcomes are defined as follows:

$$z_{nij} = \begin{cases} 0 & \text{if } j \in \{1,2\} \\ 1 & \text{otherwise} \end{cases}. \tag{2}$$

In this definition, $z_{nij} = 1$ indicates that person n judges the intensity of preference for the chosen statement for item i to be *much more*, whereas $z_{nij} = 0$ denotes that the intensity of preference is *a little more* only. Any one of the IRT models for dichotomous items can be used for this probability. For example, the following Rasch model (Rasch, 1960) can be used:

$$p\left(z_{nij} = 1 | \gamma_n, \xi_{ij}\right) = \frac{\exp\left(\gamma_n + \xi_{ij}\right)}{1 + \exp\left(\gamma_n + \xi_{ij}\right)}, \tag{3}$$

where γ_n represents the overall level of preference intensity for person n, and ξ_{ij} represents the specific utility of category j for item i. A respondent tends to choose the *much more* category for the preferred statement when $\xi_{ij} > 0$.

The Rasch model, which is a dichotomous IRT model, is used in Equation (3) because there are only two categories as outcomes in Process II when four-category MFC items are used. In practice, more response categories may be found. Figure 3 shows an MFC item with six categories. Any polytomous IRT model, such as the graded response model (Samejima, 1969), can be used in Process II.

Finally, the two processes are combined multiplicatively to produce the observed responses. The probability of selecting category j in item i for person n can be expressed as

$$p\left(y_{ni} = j | \theta_n, \gamma_n, \delta_{si}, \delta_{ti}, \xi_{ij}\right) = p\left(y^{(\mathrm{I})} = x_{ni}\right) \times p\left(y^{(\mathrm{II})} = z_{nj}\right), \tag{4}$$

where $p\left(y^{(\mathrm{I})} = x_{ni}\right)$ is the probability of observing x_{ni} in Process I, which follows Equation (1); and $p\left(y^{(\mathrm{II})} = z_{nj}\right)$ is the probability of observing z_{nj} in Process II, which can follow any dichotomous or polytomous IRT model (e.g., Equation 3). Equation (4) is henceforth referred to as the dual process IRT model (DPM) for polytomous MFC items. For illustration, when J = 4, the probabilities are as follows:

$$\begin{aligned} p(y_{ni}=0|\boldsymbol{\theta}_n,\gamma_n,\delta_{si},\delta_{ti},\xi_{ij}) &= p(x_{ni}=0)\times p(z_{nij}=1), \quad j=0, \\ p(y_{ni}=1|\boldsymbol{\theta}_n,\gamma_n,\delta_{si},\delta_{ti},\xi_{ij}) &= p(x_{ni}=0)\times p(z_{nij}=0), \quad j=1, \\ p(y_{ni}=2|\boldsymbol{\theta}_n,\gamma_n,\delta_{si},\delta_{ti},\xi_{ij}) &= p(x_{ni}=1)\times p(z_{nij}=0), \quad j=2, \text{ and} \\ p(y_{ni}=3|\boldsymbol{\theta}_n,\gamma_n,\delta_{si},\delta_{ti},\xi_{ij}) &= p(x_{ni}=1)\times p(z_{nij}=1), \quad j=3. \end{aligned} \tag{5}$$

When applying the DPM, polytomous responses in the MFC items are treated as a set of pseudo-items. For example, for MFC items with four categories, as shown in Figure 2, two pseudo-items were introduced that corresponded to Process I and Process II, respectively. Specifically, category 4 (j = 3) was treated as two pseudo-items with a value of 1 in both; category 3 (j = 2) was treated as two pseudo-items with the values 1 and 0, respectively; category 2 (j = 1) was treated as two pseudo-items with a value of 0 in both; and category 1 (j = 0) was treated as two pseudo-items with the values 0 and 1, respectively. The RIM and the Rasch model can be used for Process I and Process II, respectively.

3 | PARAMETER ESTIMATION

As mentioned earlier, due to the ipsative nature of the data, $\theta_D = -\sum_{d=1}^{D-1}\theta_d$. In other words, only $D-1$ θ traits will be freely estimated. As such, there are a total of D ($D-1$ θ traits plus one γ trait) random-effect parameters in the DPM. Apart from this constraint, the mean statement utility of each dimension should be constrained to be zero and the mean of category utilities to be zero to identify the new model.

Ipsative tests often have many dimensions (e.g., the JVIS measures 34 dimensions and the EPPS measures 15 dimensions). Following previous studies on IRT models for ipsative data (Wang et al., 2016, 2017), the freeware JAGS (Plummer, 2003) which implements the Bayesian Markov chain Monte Carlo (MCMC) method, was used in this study. For the DPM, the full posterior distribution is given by

$$p(\boldsymbol{\theta},\delta,\xi|\mathbf{Y}) \propto \prod_{n=1}^{N}\prod_{i=1}^{I} p(y_{nij}|\boldsymbol{\theta}_n)p(\boldsymbol{\theta}_n|\boldsymbol{\mu},\boldsymbol{\Sigma})p(\boldsymbol{\mu}|\boldsymbol{\Sigma})p(\boldsymbol{\Sigma})p(\delta)p(\xi), \tag{6}$$

where $\boldsymbol{\theta}_n^T = (\theta_{n1}, \ \ldots, \ \theta_{nD},\gamma_n)$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $\mathbf{Y}$ denotes the item responses; $p(y_{nij}|\boldsymbol{\theta}_n)$ denotes the probability of choosing category j of item i for person n with latent trait $\boldsymbol{\theta}_n$, which is calculated as in Equation (4); $p(\boldsymbol{\theta}_n|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the conditional probability of latent trait $\boldsymbol{\theta}_n$ for person n; and $p(\boldsymbol{\mu}|\boldsymbol{\Sigma})$, $p(\boldsymbol{\Sigma})$, $p(\delta)$, and $p(\xi)$ are the priors for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, δ and ξ, respectively. Like previous studies (Wang et al., 2016, 2017), this study used non-informative priors, unless specified otherwise. Specifically, for the D random-effect parameters, the priors were specified as $\mu_d \sim N(0, \ 1)$, where μ_d is the mean level for the dth dimension, and $\boldsymbol{\Sigma}$ follows the inverse Wishart distribution $W^{-1}(\mathbf{I}, df)$, where $\mathbf{I}$ denotes the identity matrix, and $df = D$. For the fixed-effect parameters, they were specified as $\delta \sim N(0, \ 1)$, and $\xi \sim N(0, \ 1)$. It was found in the pilot study that non-informative priors did not lead to improper posterior distributions.

To obtain appropriate estimates of the parameters, it is important that the MCMC algorithms sample from the target posterior distribution after it has converged to a stationary distribution. Thus, the number of iterations to discard (burn-in) and the number of subsequent iterations for further analysis should be specified, and the convergence of the MCMC chain(s) should be checked. In this study, two chains with 10,000 (in the simulation study) or 15,000 (in the empirical study) iterations as burn-in, followed by an additional 5000 iterations, were run. The burn-in period was determined based on the pilot analysis. In the simulation study, only the δ parameters were provided initial values which were their generating values, while in the empirical example, no initial value was provided. Graphical methods such as history plots and the Gelman–Rubin statistical index $\hat{R}$ (Cowles & Carlin, 1996; Lunn et al., 2012) have been used in many previous studies (Wang et al., 2017), as well as in this study. As a rule of thumb, a value of $\hat{R}$ near 1 indicates the means of the samplers in first and second half are the same and thus convergence has been achieved.

4 | SIMULATION STUDY

4.1 | Design and analysis

In the simulation study, the following four variables were manipulated: the number of latent traits or dimensions, D (2, 6 or 12); the number of statements in each dimension, L (6 or 10); the sample size, N (300 or 1000, representing a median or a large sample size, respectively); and the linking design (complete linking or spiral linking). In a complete linking design, as shown in Figure 4 (left), any two statements from different latent traits are paired to form an MFC item. Consequently, there are $L^2 \times [D \times (D-1)/2]$ MFC items, where D is the number of dimensions and L is the balanced number of statements in each dimension. In a spiral linking design, the statements are linked as shown in Figure 4 (right), which yields $D \times L$ MFC items. Obviously, a complete linking design will produce a large number of items when D is high. For example, for $D = 6$ and $L = 6$, 540 items are constructed. If L increases to 10, there will be a total of 1500 MFC items. For this reason, the complete linking design was not considered under $D = 6$ and $D = 12$ in this study, because the calibration of the data was beyond the computer's memory capacity. The simulation design is shown in Table 1, and a total of 16 conditions were examined in this simulation study. The condition $D = 6$, $L = 6$, $N = 300$, and spiral linking mimicked the design of the empirical example in this study.

A Matlab program was written to generate item responses using the DPM. The data-generation procedure contained the following steps. First, the utilities of the statements (δ) were generated from $N(0, 1)$ because theoretically the parameter can be in the range between $-\infty$ and $+\infty$. The number of

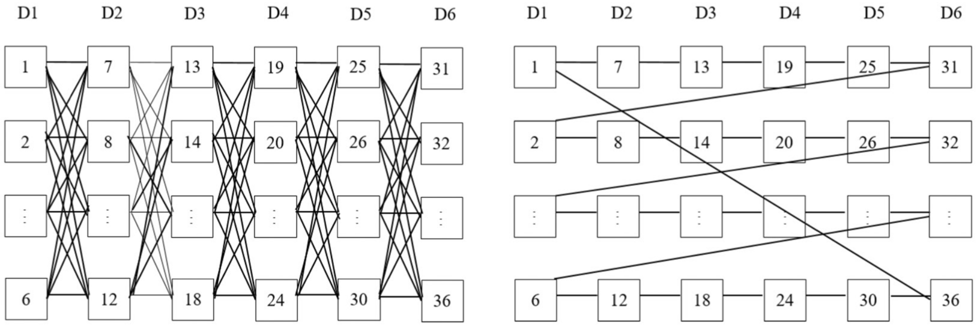

FIGURE 4 Complete (left) and spiral (right) linking design under the condition $D = 6$ and $L = 6$.

TABLE 1 Simulation design in the simulation study.

	Number of statements in dimension, L		Sample size, N		Linking design		Number of conditions
	6	10	300	1000	Complete	Spiral	
$D = 2$	✓	✓	✓	✓	✓	✓	8
$D = 6$	✓	✓	✓	✓	–	✓	4
$D = 12$	✓	✓	✓	✓	–	✓	4
Total							16

categories was fixed at four. The utilities of the categories (ξ) were generated from $N(0, 1)$. For conditions that mimicked the empirical example, the estimates from the real data were used as the generating values. Second, the latent traits were randomly generated from a D-dimensional ($D-1$ θ traits plus one γ trait) multivariate normal distribution with mean zero vector and a particular correlation matrix. Interested readers can find the Matlab code and the correlation matrix that was used to generate data for the $D = 6$ conditions in Appendix S2 of the Supporting Information. For $D = 2$ and 12 conditions, the correlation matrices used were obtained from previous studies (Brown & Maydeu-Olivares, 2011; Wang et al., 2017). The ability level for the Dth dimension was not generated. Rather, it was computed as $-\sum_{d=1}^{D-1}\theta_{dn}$ in line with the nature of ipsative data. For example, for conditions of six dimensions, we have $\theta_6 = -(\theta_1 + \cdots + \theta_5)$. Third, these generated random-effects parameters and the fixed-effects statement utility parameters were used to compute the corresponding category probabilities and their cumulative probabilities using the DPM. Fourth, these cumulative probability values were compared with a randomly generated number from a uniform distribution $U(0, 1)$, and the simulated item response was defined as the highest score category in which the random number was less than or equal to the associated cumulative probability. After the data had been generated, the data-generating models were used to fit the data sets. Thirty replications were conducted in each condition using JAGS. The number of replications was decided by referring to previous studies (Wang et al., 2017), and, more importantly, the accurate and stable estimation of parameters across replications, as shown in the simulation study results.

The dependent variables were the bias and root mean square error (RMSE) in the estimates across replications. For the simulation study, it was expected that the larger the sample size was, the better the parameter recovery would be. The complete linking design was expected to yield a more accurate estimation than the spiral linking design because there are many more data points (and thus smaller sampling variances and RMSE values) under the former than the latter. Conversely, for the spiral design, the larger the number of dimensions (D) and the more statements in each dimension there were (L), the poorer the parameter estimation was expected to be. To interpret, an index that represents the degree of linkage (dl) is computed as the proportion of the *linked* items relative to *all possible* items. It appears that that the larger the value of dl, the less missingness of data, the more accurate the parameter estimates are. For the complete design conditions, $dl = 1$. For the spiral design conditions,

$$dl = \frac{D \times L}{L^2 \times [D \times (D-1)/2]} = \frac{2}{(D-1) \times L}, \tag{7}$$

where dl is inversely proportional to D and L.

4.2 | Results

All analysis were performed on a computer with two Xeon 2.6 GHz cores and 64 GB memory. The computation time for each replication under various conditions of $D = 2$ was between 1.5 and 18 h. Due to space constraints, Tables 2 and 3 show the summarized results of $L = 6$ and $L = 10$, respectively. More information is provided in Appendix S3 in the Supporting Information, where Figures S1–S4 present the bias and RMSE values of the parameters for the condition $L = 6$, and Figures S5–S8 plot the results for $L = 10$. The convergence checks of MCMC for simulation study are provided in Appendix S4 in the Supporting Information.

Specifically, for $L = 6$ (Table 2), under the conditions of the complete linking design, the bias values were between −.056 and .056, and the RMSE values were between .024 and .158. Under the conditions of the spiral linking design, the bias values were between −.060 and .076, and the RMSE values were between .033 and .177. As in previous studies (Wang et al., 2017), most bias and RMSE values were less than .1, indicating that the parameters were well recovered. In general, the recovery of the parameters under the complete linking design was better than that for the spiral linking design, and the accuracy of estimation increased as the sample size increased.

TABLE 2 Summary of bias values and root mean square errors for the parameter estimates of the DPM in the simulation study with two dimensions and six statements in each dimension.

		N300_complete		N1000_complete		N300_spiral		N1000_spiral	
Par.	Est.	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
δ	Mean	.000	.057	.000	.031	.000	.153	.000	.083
	SD	.012	.007	.007	.005	.047	.015	.013	.020
	Min.	−.017	.048	−.014	.024	−.051	.128	−.023	.065
	Max.	.024	.075	.006	.039	.076	.177	.021	.127
ξ	Mean	.000	.138	.000	.076	.000	.138	.000	.074
	SD	.036	.024	.013	.012	.044	.017	.020	.014
	Min.	−.056	.097	−.024	.047	−.060	.103	−.032	.049
	Max.	.056	.158	.035	.103	.058	.159	.024	.093
$\boldsymbol{\mu}$	Mean	.000	.041	.000	.027	.000	.037	.000	.022
	SD	.020	.036	.008	.024	.005	.032	.001	.019
	Min.	−.020	.062	−.008	.041	−.005	.056	−.001	.033
	Max.	.020	.062	.008	.041	.005	.056	.001	.033
$\boldsymbol{\sigma}$	Mean	.003	.064	.001	.034	−.004	.097	−.003	.040
	SD	.008	.032	.008	.018	.015	.055	.006	.020
	Min.	−.006	.073	−.004	.034	−.032	.082	−.014	.042
	Max.	.017	.085	.017	.051	.009	.141	.000	.052

Note: δ is the statement utility parameter; ξ is the category utility parameter; $\boldsymbol{\mu}$ is the mean level of the $\boldsymbol{\theta}$ variables; $\boldsymbol{\sigma}$ represents the variance–covariance elements between the latent variables; N300_complete indicates a sample size of 300 and the complete linking design; N300_spiral indicates sample size 300 and the spiral linking design; and so on.

Abbreviations: Est., estimates; Max., maximum value; Min., minimum value; Par., parameters; RMSE, root mean square error; *SD*, standard deviation.

For L = 10 (Table 3), under the conditions of the complete linking design, the bias values were between −.053 and .046, and the RMSE values were between .021 and .168. Under the conditions of the spiral linking design, the bias values were between −.131 and .085, and the RMSE values were between .036 and .239.

A comparison of Tables 2 and 3 reveals that when the complete design is used, the recovery of parameters with $L = 10$ is better than that with $L = 6$. As expected, when the spiral linking design is used, the recovery of parameters is poorer under L = 10 than that under L = 6. The reason is that there were 100 possible items for the conditions of L = 10, of which 20 were used for linking, whereas there were 36 possible items for $L = 6$, of which 12 items were used. Following Equation (7), *dl* values were .200 for $L = 10$ and .333 for $L = 6$, respectively. The former conditions have smaller *dl*, suggesting more missingness of the data, and hence yielded poorer estimations.

For $D = 6$, the computation time under the spiral linking design ranged from about 2 to 17 h. Table 4 shows the summarized results for the parameter estimates. The parameter recovery was acceptable when N= 300 and satisfactory when N= 1000. For example, for the condition D = 6, L = 6 and N = 300, in which item responses were generated with the estimates of the DPM used in the empirical example shown in Table 6, the bias values were between −.070 and .098, and the RMSE values were between .060 and .267. The detailed results are shown in Figures S9 and S10 for L = 6, and in Figures S11 and S12 for L = 10. The respective *dl* values for L = 6 and L = 10 were .07 and .04. Again, a smaller *dl* under L = 10 led to relatively larger bias and RMSE values for the parameters. In particular, the RMSE values of the δ parameters are relatively large because less information is available for the estimation of the δ parameters, compared with the other parameters, due to the small number of items in the spiral design. For the condition D = 12, the computation was very time-consuming, ranging from 4 to 38 h. As shown in Table 5, the parameter recovery was poor for N = 300 due to the use of a complicated model, a large

TABLE 3 Summary of bias values and root mean square errors for the parameter estimates of the DPM in the simulation study with two dimensions and ten statements in each dimension.

		N300_complete		N1000_complete		N300_spiral		N1000_spiral	
Par.	Est.	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
δ	Mean	.000	.049	.000	.027	.000	.178	.000	.117
	SD	.008	.006	.005	.004	.064	.026	.035	.011
	Min.	−.023	.040	−.012	.021	−.131	.147	−.056	.095
	Max.	.016	.065	.008	.033	.085	.239	.067	.147
ξ	Mean	.001	.128	.001	.071	.001	.126	.001	.073
	SD	.023	.018	.013	.010	.026	.015	.016	.009
	Min.	−.053	.093	−.031	.049	−.083	.110	−.035	.059
	Max.	.046	.168	.036	.094	.071	.165	.044	.102
$\boldsymbol{\mu}$	Mean	.000	.038	.000	.020	.000	.039	.000	.024
	SD	.012	.033	.002	.017	.014	.034	.011	.021
	Min.	−.012	.058	−.002	.030	−.014	.058	−.011	.036
	Max.	.012	.058	.002	.030	.014	.058	.011	.036
$\boldsymbol{\sigma}$	Mean	.012	.070	−.001	.039	−.007	.080	.007	.045
	SD	.017	.038	.002	.023	.026	.051	.010	.025
	Min.	−.003	.066	−.006	.032	−.037	.059	−.010	.040
	Max.	.033	.100	.001	.064	.032	.143	.015	.070

Note: δ is the statement utility parameter; ξ is the category utility parameter; $\boldsymbol{\mu}$ is the mean level of the $\boldsymbol{\theta}$ variables; and $\boldsymbol{\sigma}$ represents the variance–covariance elements between the latent variables. N300_complete indicates a sample size of 300 and the complete linking design; N300_spiral indicates a sample size of 300 and the spiral linking design; and so on.

Abbreviations: Est., estimates; Max., maximum value; Min., minimum value; Par., parameters; RMSE, root mean square error; *SD*, standard deviation.

proportion of missingness in the spiral linking design, and small sample size. However, the recovery was acceptable when the sample size was increased to 1000. More results are provided in Figures S13–S16.

The simulations showed that the parameters in the DPM could be well recovered using JAGS. The complete linking design yielded more accurate estimates than the spiral linking design. In general, the larger the sample was, the more accurate the estimates were. Moreover, the effect of the number of statements in each dimension, L, varied between linking designs. When the complete design was used, the accuracy increased as L increased. In contrast, when the spiral linking design was used, the accuracy of parameter estimation decreased as L increased.

5 | AN EMPIRICAL EXAMPLE

A questionnaire was constructed to measure career interests using four-category polytomous MFC items, as shown in Figure 1. It was designed following the popular Holland code (Holland, 1997), which assesses six types of career interest: realistic (R), investigative (I), artistic (A), social (S), enterprising (E) and conventional (C). Statements were retrieved from open-access sample items, and all of the statements were positively keyed. Each career interest was measured with six statements, and the spiral linking design shown in Figure 4 (right) was implemented. Therefore, the questionnaire consisted of 36 polytomous MFC items, and each statement was presented to the participants twice. A counterbalancing method was used to control the sequence effects of the statements. Specifically, when a statement was presented as the first one for an item, it was later presented as the second one for another item. It is important to note that the utilities of a statement in two different items are constrained to be identical. The survey took about 20 min to complete.

TABLE 4 Summary of bias values and root mean square errors for the parameter estimates of the DPM in the simulation study with six dimensions and a spiral linking design.

		$L=6$				$L=10$			
		N300_spiral		N1000_spiral		N300_spiral		N1000_spiral	
Par.	Est.	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
δ	Mean	.000	.203	.000	.137	.000	.255	.000	.188
	SD	.048	.019	.044	.017	.121	.045	.041	.019
	Min.	−.070	.163	−.055	.097	−.292	.182	−.091	.155
	Max.	.098	.239	.084	.164	.223	.385	.061	.253
ξ	Mean	.000	.120	.000	.071	.000	.175	.000	.077
	SD	.027	.015	.011	.009	.043	.029	.017	.015
	Min.	−.063	.083	−.019	.056	−.109	.113	−.054	.044
	Max.	.047	.148	.028	.092	.099	.261	.041	.110
$\boldsymbol{\mu}$	Mean	−.003	.061	−.003	.034	.000	.075	.000	.031
	SD	.016	.028	.014	.018	.009	.036	.009	.014
	Min.	−.026	.060	−.031	.030	−.012	.066	−.013	.031
	Max.	.025	.078	.013	.057	.014	.108	.012	.044
$\boldsymbol{\sigma}$	Mean	.001	.109	−.001	.059	−.005	.116	.001	.054
	SD	.020	.039	.012	.024	.029	.046	.008	.023
	Min.	−.032	.067	−.023	.038	−.071	.071	−.015	.031
	Max.	.036	.267	.031	.154	.048	.321	.019	.152

Note: δ is the statement utility parameter; ξ is the category utility parameter; $\boldsymbol{\mu}$ is the mean level of the $\boldsymbol{\theta}$ variables; and $\boldsymbol{\sigma}$ represents the variance–covariance elements between the latent variables. N300_spiral indicates a sample size of 300 and the spiral linking design; N1000_spiral indicates a sample size of 1000 and the spiral linking design.

Abbreviations: Est., estimates; Max., maximum value; Min., minimum value; Par., parameters; RMSE, root mean square error; *SD*, standard deviation.

Three hundred and one students at a university in Hong Kong were invited to complete the questionnaire in return for a feedback report if requested. They were required to compare their preferences for two activities related to different careers and choose one of the four categories according to their preferences. Demographic data such as age range, gender, and educational level were also collected. Of the participants who provided responses to the questionnaire, 63.79% were female. The distribution across the age groups was as follows: 1.66% of participants were below 18 years of age, 85.38% were in the 18–25 age range, 8.97% were 26–35, and 3.99% were over 35. Furthermore, 81.06% of the participants were undergraduate students; the remaining participants were studying for master's or doctoral degrees. The study protocol was approved by the Institutional Ethical Review Board of the university at which the study was conducted.

The DPM was fitted to the data using JAGS (with the JAGS codes shown in the Appendix S5 in the Supporting Information). As mentioned earlier, the mean vector and covariance matrix for the five θ variables and one γ variable were freely estimated, and the sixth θ variable was computed as $\theta_6 = -(\theta_1 + \dots + \theta_5)$. Non-informative priors were used for the utility parameters, step parameters, and mean vector by setting as $N(0, 1)$. The priors for the covariance matrix were set using an inverse Wishart distribution $W^{-1}[\mathbf{R}, K]$, with $\mathbf{R} = \mathbf{I}$ and the hyperparameter $K = 6$. In the subsequent analysis, the posterior mean and standard deviation were treated as the point estimate and standard error, respectively.

A visual inspection of the sampling histories of the chain was used to examine their convergence (See Appendix S4 in the Supporting Information). For illustration, Figure S19 shows the history plots with two chains after burn-in for the overall utility of the first statement describing realistic interest ($\delta_{1,1}$), the category utility of the first statement describing realistic interest (ξ_1), and the mean level of realistic interest ($\mu_{\theta 1}$) under the DPM. The plots indicated convergence to a stationary distribution because the two

TABLE 5 Summary of bias values and root mean square errors for the parameter estimates of the DPM in the simulation study with 12 dimensions and spiral linking design.

		L **= 6**				***L*** **= 10**			
		N300_spiral		**N1000_spiral**		**N300_spiral**		**N1000_spiral**	
Par.	**Est.**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
δ	Mean	.000	.235	.000	.172	.000	.268	.000	.212
	SD	.115	.030	.087	.021	.178	.082	.100	.039
	Min.	−.234	.173	−.154	.125	−.482	.164	−.284	.145
	Max.	.207	.300	.159	.227	.332	.546	.173	.339
ξ	Mean	.000	.134	.000	.075	.000	.140	.000	.077
	SD	.035	.022	.016	.012	.036	.021	.015	.013
	Min.	−.084	.086	−.056	.052	−.086	.097	−.040	.044
	Max.	.096	.210	.033	.104	.084	.212	.038	.129
$\boldsymbol{\mu}$	Mean	.000	.073	.000	.046	.000	.079	.000	.045
	SD	.015	.025	.013	.016	.013	.026	.012	.014
	Min.	−.024	.067	−.026	.040	−.025	.073	−.017	.041
	Max.	.030	.117	.021	.072	.024	.108	.016	.058
$\boldsymbol{\sigma}$	Mean	−.015	.191	−.011	.120	−.010	.149	−.005	.088
	SD	.159	.088	.096	.048	.092	.066	.044	.032
	Min.	−.361	.069	−.203	.045	−.226	.056	−.110	.037
	Max.	.275	.539	.158	.270	.140	.533	.064	.249

Note: δ is the statement utility parameter; ξ is the category utility parameter; $\boldsymbol{\mu}$ is the mean level of the $\boldsymbol{\theta}$ variables; $\boldsymbol{\sigma}$ represents the variance–covariance elements between the latent variables; N300_spiral indicates a sample size of 300 and the spiral linking design; and N1000_spiral indicates a sample size of 1000 and the spiral linking design.

Abbreviations: Est., estimates; Max., maximum value; Min., minimum value; Par., parameters; RMSE, root mean square error; *SD*, standard deviation.

chains were mixed and were located in the same parts of the target distribution. Moreover, the $\hat{R}$ values for all the parameters were near 1.0, indicating convergence in the MCMC estimation.

To demonstrate the advantages of fitting the DPM to polytomous MFC items, the RIM was also fitted to the data where the responses were dichotomized by transforming those of 'I prefer A much more' and 'I prefer A slightly more' in Figure 1 to score 0 and those of 'I prefer B much more' and 'I prefer B slightly more' to score 1. When fitting the RIM using JAGS, the specifications were similar to those of the DPM. The models were compared using two criteria: deviance information criterion (DIC; Lunn et al., 2012), which is a simple estimate of predictive error; and leave-one-out cross-validation (LOO; Vehtari et al., 2017), which compares the predictive performance of models on new data. The DIC was obtained directly from JAGS, whereas the LOO was computed using the R package *loo* (Vehtari et al., 2022). In particular, LOO estimates the expected log pointwise predictive density (elpd) as the measure to evaluate the out-of-sample prediction accuracy, and implements the efficient Pareto-smoothed importance sampling (PSIS) procedure to compute the LOO (PSIS-LOO). The results show that the DPM had a smaller DIC than the RIM (24,040 compared to 27,422), and the estimated elpd difference of 1964.75 with a standard error of 132.33 favours the DPM. Both the DIC and PSIS-LOO indicate that the DPM provided a better fit to the data. With respect to the six θ traits being measured in this example, the correlations between the estimates derived from the two models were .84, .83, .88, .92, .88, and .90, respectively, suggesting moderate to high correlations between the DPM and RIM estimates across the six traits. Nevertheless, it is worth noting that the DPM provides additional information about the preference intensity with γ estimates, while the RIM does not.

To check the model–data fit of the DPM, this study implemented the posterior predictive model checking (PPMC) method (Gelman et al., 1996). Specifically, when running the MCMC algorithm with

TABLE 6 Estimates and standard errors for statement parameters, mean level of latent traits, variance–covariance and correlation matrix between latent traits under the DPM in the empirical example.

Par.	Est.	*SE*	Par.	Est.	*SE*	Par.	Est.	*SE*	Par.	Est.	*SE*	Par.	Est.	*SE*
$\delta_{1,1}$	.27	.21	$\delta_{4,2}$	.08	.18	$\mu_{\theta 1}$	−.09	.04	$\sigma^2_{\theta 2,\theta 4}$	−.08	.03	$\varrho_{\theta 4,\theta 5}$	−.35	.15
$\delta_{1,2}$	−.19	.18	$\delta_{4,3}$	.00	.21	$\mu_{\theta 2}$	−.11	.04	$\sigma^2_{\theta 3,\theta 5}$	−.05	.03	$\varrho_{\theta 5,\theta 6}$	−.26	.11
$\delta_{1,3}$	.25	.19	$\delta_{4,4}$	−.32	.20	$\mu_{\theta 3}$	.01	.05	$\sigma^2_{\theta 4,\theta 6}$	−.03	.03	$\varrho_{\theta 1,\theta 3}$	.08	.17
$\delta_{1,4}$	.00	.21	$\delta_{4,5}$	−.20	.19	$\mu_{\theta 4}$	.31	.05	$\sigma^2_{\theta 1,\theta 4}$	−.08	.03	$\varrho_{\theta 2,\theta 4}$	−.40	.13
$\delta_{1,5}$	.02	.18	$\delta_{4,6}$	.00	.19	$\mu_{\theta 5}$	−.10	.05	$\sigma^2_{\theta 2,\theta 5}$	−.02	.03	$\varrho_{\theta 3,\theta 5}$	−.22	.14
$\delta_{1,6}$	−.31	.19	$\delta_{5,1}$	−.27	.19	$\mu_{\theta 6}$	.00	.05	$\sigma^2_{\theta 3,\theta 6}$	−.10	.03	$\varrho_{\theta 4,\theta 6}$	−.10	.11
$\delta_{2,1}$	.18	.21	$\delta_{5,2}$	−.02	.18	μ_{γ}	−1.22	.09	$\sigma^2_{\theta 1,\theta 5}$	.00	.02	$\varrho_{\theta 1,\theta 4}$	−.37	.12
$\delta_{2,2}$	.01	.18	$\delta_{5,3}$	.65	.21	$\sigma^2_{\theta 1}$	.14	.03	$\sigma^2_{\theta 2,\theta 6}$	.02	.02	$\varrho_{\theta 2,\theta 5}$	−.09	.16
$\delta_{2,3}$	.31	.20	$\delta_{5,4}$	−.09	.19	$\sigma^2_{\theta 2}$	.13	.03	$\sigma^2_{\theta 1,\theta 6}$	−.07	.02	$\varrho_{\theta 3,\theta 6}$	−.45	.09
$\delta_{2,4}$	−.41	.20	$\delta_{5,5}$	.14	.19	$\sigma^2_{\theta 3}$	.22	.05	σ^2_{γ}	2.13	.22	$\varrho_{\theta 1,\theta 5}$	.01	.14
$\delta_{2,5}$	.25	.18	$\delta_{5,6}$	−.41	.20	$\sigma^2_{\theta 4}$	.30	.06	$\sigma^2_{\theta 1,\gamma}$	.08	.06	$\varrho_{\theta 2,\theta 6}$	.13	.12
$\delta_{2,6}$	−.34	.20	$\delta_{6,1}$	−.18	.19	$\sigma^2_{\theta 5}$	.21	.04	$\sigma^2_{\theta 2,\gamma}$	−.05	.06	$\varrho_{\theta 1,\theta 6}$	−.39	.12
$\delta_{3,1}$	.17	.21	$\delta_{6,2}$	.23	.19	$\sigma^2_{\theta 6}$	.23	.04	$\sigma^2_{\theta 3,\gamma}$	−.15	.07	$\varrho_{\theta 1,\gamma}$	.13	.06
$\delta_{3,2}$	.37	.18	$\delta_{6,3}$	.11	.21	$\sigma^2_{\theta 1,\theta 2}$	.00	.02	$\sigma^2_{\theta 4,\gamma}$	−.01	.07	$\varrho_{\theta 2,\gamma}$	−.09	.06
$\delta_{3,3}$	.19	.21	$\delta_{6,4}$	.23	.19	$\sigma^2_{\theta 2,\theta 3}$	−.05	.03	$\sigma^2_{\theta 5,\gamma}$	.09	.06	$\varrho_{\theta 3,\gamma}$	−.19	.07
$\delta_{3,4}$	.02	.20	$\delta_{6,5}$	−.24	.19	$\sigma^2_{\theta 3,\theta 4}$	−.03	.04	$\sigma^2_{\theta 6,\gamma}$	.05	.07	$\varrho_{\theta 4,\gamma}$	−.02	.06
$\delta_{3,5}$	.01	.19	$\delta_{6,6}$	−.15	.21	$\sigma^2_{\theta 4,\theta 5}$	−.09	.04	$\varrho_{\theta 1,\theta 2}$	−.05	.16	$\varrho_{\theta 5,\gamma}$	.14	.06
$\delta_{3,6}$	−.76	.20	$\xi_{\min}$	−.72	.14	$\sigma^2_{\theta 5,\theta 6}$	−.06	.02	$\varrho_{\theta 2,\theta 3}$	−.30	.17	$\varrho_{\theta 6,\gamma}$	.07	.07
$\delta_{4,1}$	.43	.20	$\xi_{\max}$	.65	.17	$\sigma^2_{\theta 1,\theta 3}$	.01	.03	$\varrho_{\theta 3,\theta 4}$	−.13	.15			

Note. δs denote the statement utility parameters, with the first number in the subscript representing the latent trait θ (1, realistic; 2, investigative; 3, artistic; 4, social; 5, enterprising; 6, conventional) and the second number representing the statement; ξs denote the category utility parameters; μs denote the mean level of the latent variables; γ is the latent trait in Process II in the DPM; σs denote the variance–covariance elements between the latent variables; and ϱs denote the correlation between the latent variables.

Abbreviations: Est., estimates; Max., maximum value; Min., minimum value; Par., parameters; *SE*, standard error.

the DPM, the replicated data were obtained by drawing samples from the posterior distribution after JAGS converges. Then the differences in some selected discrepancy measures between the observed and replicated data were evaluated. If the posterior predictive *p*-value of the observed data is beyond a critical range (e.g., .025–.975) of the replicated data, it can be concluded that the DPM does not fit the observed data (Meng, 1994). Four discrepancy measures were used for PPMC in this study: the frequency of categories of each item; the frequency of categories in the test; the item score distribution, which is the number of examinees responding to each category for each item (Zhu & Stone, 2012); and Yen's Q3 statistic (Yen, 1993). The first statistic has previously been applied to examine model–data fit for ipsative tests with MPC items (Wang et al., 2017). The other three measures, which are powerful in detecting misfit in tests using Likert-type items (Zhu & Stone, 2012), have never been applied in ipsative tests. Due to space constraints, the detailed computation of the posterior predictive *p*-value for the discrepancy measures and the results are provided in the Appendix A. The findings indicate a good model–data fit of the DPM.

Table 6 shows the estimates and standard errors (*SE*) for the individual statement utilities and the mean level of traits under the DPM in the empirical example. For the 36 statements, the utilities ranged from −.76 to .65 (M = 0, SD = .28). The statement 'Like to meet important people', which measures enterprising interest, had the highest utility, whereas the statement 'Can play a musical instrument', which measures artistic interest, had the lowest utility. Table 7 shows the frequencies of the categories for items

TABLE 7 Frequencies of categories for the statements with highest and lowest utilities in the empirical example.

Item	Statements	$j = 3$ (%)	$j = 2$ (%)	$j = 1$ (%)	$j = 0$ (%)	Missing (%)
2	(A) Like to meet important people (B) Like social activities	28.24	27.91	33.55	9.63	.66
30	(A) Like keeping records and files (B) Like to meet important people	10.30	31.23	39.21	19.27	0
4	(A) Can play a musical instrument (B) Can understand science and use information to figure things out	15.95	28.91	34.22	20.93	0
5	(A) Cooperate well with others (B) Can play a musical instrument	36.88	37.21	21.93	3.65	.33

Note: 3 = '*I prefer A much more*'; 2 = '*I prefer A a little more*'; 1 = '*I prefer B a little more*'; 0 = '*I prefer B much more*'. The statement 'Like to meet important people' had the highest utility; the statement 'Can play a musical instrument' had the lowest utility.

involving the statements with the highest and lowest utility estimates in the raw scores. For example, the statement 'Like to meet important people' was presented as the first statement in item 2 and the second statement in item 30. When presented with the first statement in item 2, 27.91% of the participants chose the category 'I prefer A a little more', and 28.24% of them chose the category 'I prefer A much more'. Therefore, 56.15% of the participants chose the categories 'I prefer A a little more' and 'I prefer A much more'. Similarly, when the statement was presented as the second statement in item 30, 58.48% of the participants chose the categories 'I prefer B much more' and 'I prefer B a little more'. The results showing that the participants preferred the statement 'Like to meet important people' over the paired statements are consistent with the DPM estimation that the aforesaid statement had the highest utility estimate. The frequencies of categories for the statement 'Can play a musical instrument' in Table 7 were consistent with the expectation that the statement had a lower utility than the paired statement.

In terms of the mean level of latent traits, the differential level of social interest displayed the highest mean ($\mu_{\theta 4} = .31$, $SE = .05$), whereas the differential levels of investigative ($\mu_{\theta 2} = -.11$, $SE = .09$) and conventional ($\mu_{\theta 5} = -.11$, $SE = .05$) interests showed the lowest means. The variances of the six latent traits were between .13 ($\sigma^2_{\theta 3}$) and .30 ($\sigma^2_{\theta 5}$), suggesting that artistic interest differentiation has the smallest variation among students while enterprising interest differentiation has the greatest variation. Most of the correlations between the differential levels of the six types of career interest were moderately negative. Artistic and conventional interests had the highest negative correlation ($r = -.45$), followed by that between investigative and social interests ($r = -.40$). The θ variables were negatively correlated, because the expected correlations between the ipsative measures were $-1/(D-1)$, where D is the number of latent traits (Dunlap & Cornwell, 1994), which is also proved in Appendix S1 in the Supporting Information. Whereas judgement ability (γ) was significantly and negatively correlated with artistic interest differentiation ($\varrho_{\theta 3,\gamma} = -.19$, $SE = .07$), it was significantly and positively correlated with realistic interest differentiation ($\varrho_{\theta 1,\gamma} = .13$, $SE = .06$) and with enterprising interest differentiation ($\varrho_{\theta 5,\gamma} = .14$, $SE = .06$).

For each student, the mean and standard deviation of the posterior distribution of each θ latent trait were computed as the point estimate of career interest differentiation and its standard error. As a demonstration, Table 8 provides the point estimates and their SE of career interest differentiation for two students (S124 and S179), and Figure 5 provides a radar chart based on the θ estimates. These two students were selected according to the variance of their θ estimates which reflects the level of differentiation. As shown in the last column of Table 8, the variance for S124 was .621 and that for S179 was .010. Thus, the two students represented persons who have a high and low level of differentiation, respectively. The results and the chart can be used to interpret the career interest differentiation of a student (intrapersonal comparison) and to compare the levels of differentiation between students (interpersonal comparison). Again, it is important to note that both intra- and interpersonal comparisons are based on the θ estimates in the DPM (Equation 4), which represent levels of career interest differentiation in this situation, rather than levels of career interest *per se*.

In terms of the intrapersonal comparison, as shown in Figure 5, the locations of the differentiation of the six types of career interest were very diverse for S124. According to Table 8, S124 had the highest

TABLE 8 Estimates of career interest differentiation for two selected students in the empirical example.

Student	Par.	Realistic	Investigative	Artistic	Social	Enterprising	Conventional	Variance
124	Est.	−.497	.163	−.756	1.074	−.735	.752	.621
	SE	.212	.205	.196	.251	.190	.261	
179	Est.	−.047	−.147	.126	.051	.067	−.049	.010
	SE	.279	.269	.233	.273	.217	0.247	

Abbreviations: Est., estimates; Par., parameters; *SE*, standard error.

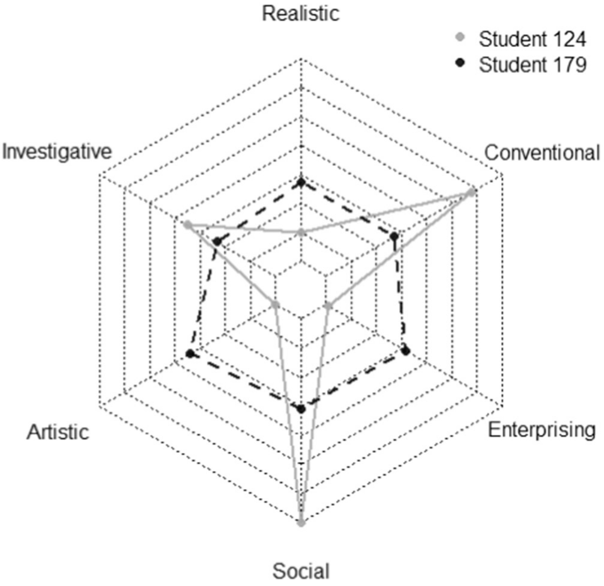

FIGURE 5 Radar chart of career interest differentiation for two students in the empirical example.

level of differentiation for social interest (1.074) and the lowest level of differentiation for artistic interest (−.756). Therefore, *S*124's level of differentiation for social interest was much higher than that for artistic interest. Conversely, the locations were very similar for *S*179, who had the highest level of differentiation for artistic interest (.126) and the lowest level of differentiation for investigative interest (−.147). It can be concluded that *S*179's artistic interest differentiation was slightly higher than his/her investigative interest differentiation.

In terms of the interpersonal comparison, first, the results could be used to compare students' differentiation levels within a specific interest. As shown in Figure 5, for artistic interest, *S*124's differential level was much higher than that of *S*179, whereas for artistic interest, *S*124's differential level was much lower than that of *S*179. Second, the results could be used to compare the students' holistic differentiation levels. As mentioned earlier, the variance of the six estimates for latent traits was computed for each student. As variance measured the spread of differentiation of the six career interests, it reflected the holistic differentiation for individual students. According to the variances that are shown in Table 8, *S*124 had a much higher holistic differentiation across career interests than *S*179 did.

The DPM also yields a γ estimate for the trait level of preference intensity in Process II for each student. To reveal the implications of the γ estimate, we compared three selected students as shown in Table 9. In the table, the students' responses to 12 items that associated with θ_1 are first given (due to space constraints, we focused on realistic interest, θ_1, only). For example, in the first item, statement A measured realistic interest while statement *B* measured other interest. For this specific item, *S*128 chose $j = 0$ ('*I prefer B much more*'), suggesting that the student much preferred the other interest over realistic

TABLE 9 Responses, frequencies of responses, and estimates for three selected students in the empirical example.

	Responses												Frequencies				DPM estimates		RIM estimates
	A	A	A	B	B	B	B	A	A	A	A	B	3	2	1	0	θ_1	γ	θ_1
*S*128	0	0	0	$\underline{0}$	3	$\underline{0}$	$\underline{0}$	0	0	0	0	3	2	0	0	10	−.273	3.127	−.153
*S*58	0	1	1	$\underline{1}$	$\underline{1}$	2	2	$\underline{2}$	1	1	1	$\underline{1}$	0	3	8	1	−.287	−2.070	−.096
*S*236	$\underline{3}$	$\underline{3}$	$\underline{3}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\underline{3}$	$\underline{3}$	$\underline{3}$	$\underline{3}$	3	8	0	0	4	.467	2.151	.265

Note: For responses, ‘A' means statement A in the item measured θ_1 and ‘B’ means statement B did so. 3 = ‘*I prefer A much more*’; 2 =’ *I prefer A a little more*’; 1 = ‘*I prefer B a little mor*e’; 0 = ‘*I prefer B much more*’. An underline indicates the statement that measured θ_1 was preferred by the student. θ_1 estimates are the differentiation of first (i.e., realistic) career interest obtained from the DPM or RIM; γ estimates are the trait level of judgement obtained from the DPM.

interest. Other responses can be interpreted similarly. Note that when statement that measured θ_1 was preferred by the student, the response is underlined. Of the 12 items, *S*128 preferred realistic interest 3 times, *S*58 4 times, and *S*236 11 times. Hence, it is expected that *S*128 and *S*58 should have very similar θ_1 estimates while *S*128 and *S*236 should have very different θ_1 estimates: to be precise, *S*236 should have a much higher θ_1 estimate than *S*128.

The next four columns in Table 9 display the frequencies of the four categories across 12 items, where the sum of each row is a constant value of 12 because of the ipsative nature of the raw scores. An examination of the frequencies of *S*128 and *S*58 shows that *S*128 selected categories indicating favouring statement B (i.e., $j = 1$ and $j = 0$) 10 times out of 12, while *S*58 selected them 9 times. However, *S*128 chose the high-intensity category ($j = 0$) 10 times, while *S*58 chose it only once. On the other hand, *S*128 selected categories indicating favouring statement A ($j = 3$ and $j = 2$) twice, while *S*58 selected them 3 times, with *S*128 consistently choosing the high-intensity category ($j = 3$) and *S*58 not at all. For *S*236, the frequencies indicate that the student's preference of θ_1 was substantially different from that of *S*128. *S*236 always selected high-intensity categories, favouring statement B ($j = 0$) 4 times and statement A ($j = 3$) 8 times.

In sum, from the observations it appears that *S*128 and *S*58 had similar preferences for a realistic career, but *S*128's preference was more intense than that of *S*58. Thus, these two students represent persons with similar θ, but different γ. By contrast, *S*128 and *S*236 had very different preferences for a realistic career, but their intensities of preferences are similar. Thus, they represent persons with different θ, but similar γ.

The next two columns in Table 9 provide θ_1 and γ estimates from the DPM. It can be seen that the estimates perfectly reflect the observed differences and similarities of the students discussed above. Specifically, *S*128 (−.273) and *S*58 (−.287) had close θ_1 estimates but very different γ estimates (3.127 for *S*128 and −2.070 for *S*58). Conversely, *S*236 had a very different θ_1 estimate (.467) from that of *S*128 but a similar γ estimate (2.151). As mentioned earlier, the data were also analysed with the RIM, and the estimates are provided in the last column of Table 9. The results can be utilized to understand different persons' preferences, but only to some extent.

6 | CONCLUSION AND DISCUSSION

In real tests, polytomous MFC items are used because of their own advantages: they can reduce response bias and are more informative by providing information about respondents' preferences and the intensity of these preferences. Specifically, the forced-choice form of the items is prevalent across a wide range of psychological assessments for non-cognitive abilities. In the field of vocational counselling, where, for the purpose of identifying a suitable career, discriminating between career interests is more relevant than comparing individuals. Thus, MFC items are employed in many tests in this field (e.g., JVIS, EPPS, CIT, OPQ, and situational judgement tests). Beyond vocational counselling, the MFC format is also often used in personality assessment (Cao & Drasgow, 2019), values assessment (Josef et al., 2017), and emotion assessment (Anguiano-Carrasco et al., 2015). Along with the forced-choice format, the polytomous form of the items makes them more informative than their dichotomous counterpart. From a psychometric point of view, the greater amount of information can yield more accurate ability estimates (de la Torre, 2008). Note, however, that items with more than two categories also have practical disadvantages: for example, they are likely to increase the cognitive load on respondents, and they may be unable to completely eliminate responding bias in the form of faking, central or extreme responding.

Nearly all existing IRT models for MFC items have been developed for dichotomous MFC items, and thus cannot be applied to polytomous MFC items. This study has developed a new IRT model, the DPM, for polytomous MFC items. A prominent feature of the DPM is its cognitive modelling approach. Using this approach, the DPM describes the underlying cognitive processes involved in choosing from multiple alternatives. Specifically, the participants decide which statement they prefer (choice), and then judge the intensity of their preference for the selected statement (intensity). Thus, the proposed model contributes to the understanding of the potential psychological mechanism of human choice-making. Another prominent feature of the DPM is its ability to allow researchers to assess and compare differentiation levels (with the $\boldsymbol{\theta}$ parameter) and preference intensity (with the γ parameter) within an individual and/or between individuals using polytomous MFC items. Such comparisons have important practical implications. Taking the applications in the fields of industrial and organizational settings as an example, studies have found that interests differentiation is an effective predictor of attributes and behaviours in career decisions (Hirschi, 2009), and individuals with higher differentiation interests tend to have more career-choice readiness and a greater vocational identity (Nauta & Kahn, 2007). Hence, in a counselling context, career consultants can use the results of the DPM to help the client better understand to what extent his/her career interests differ from each other (intrapersonal comparisons). Similarly, in a selection context, practitioners can use the results to identify candidates who have higher differentiation levels regarding specific career interests (interpersonal comparisons).

A series of simulations was conducted to evaluate parameter recovery under the new model. Implementing Bayesian methods with MCMC algorithms in JAGS was recommended to calibrate the new model, because of the high dimensionality of the data. The parameters were recovered fairly well with the complete linking design. When the spiral linking design was used, the accuracies of the statement utilities were lower due to the large proportion of missing data. These results can have useful practical implications. First, both linking designs used in the simulation study, as shown in Figure 4, connected the statements successfully. Note that successful connections are necessary to place the statements being calibrated on the same scale (Wang et al., 2016, 2017). Second, although statements can be connected regardless of the linking design, the complete linking design yields more accurate parameter estimates than the spiral linking design. However, when D is large, a complete linking design also yields a much larger number of items, which can be computationally challenging. For these reasons, researchers should strike a balance between successful connections and the number of items when constructing their MFC tests. One simple and practical solution is to add more items within the spiral linking design to ensure successful connections without using an inordinate number of items. Third, as mentioned before, the DPM belongs to the family of IRT tree models, for which one needs to pay special attention to the conditions of model identification when using the new model. The simulation results showed that the conditions for identifying the DPM provided in this paper are sufficient and can be applied to real data analysis. However, it would be helpful if future research could continue examining the identification issues associated with the DPM and other related models.

The empirical example of career interest was used to demonstrate the implications and applications of the new model. The model–data fit was examined using PPMC methods with various discrepancy indices. The results showed that the model–data fit was good. Of the six types of career interest, the differential level of social interest displayed the highest mean, whereas the differential level of investigative interest showed the lowest mean. This finding seems reasonable, because most of the participants in this study were undergraduate students, who are in a critical period in their development of social skills and social network building. Most of the latent traits had moderately negative correlations. Interestingly, artistic interest differentiation had a significant negative correlation with judgement ability, suggesting that in general, the lower the level of artistic differentiation is, the stronger the judgement ability is. Conversely, realistic interest differentiation and enterprising interest differentiation had a significant positive correlation with judgement ability. Hence, the higher the levels of realistic and enterprising interest differentiation are, the stronger the judgement ability is.

Although promising, the work in this paper is only an initial step in making tests with MFC items more useful and reliable in assessing psychological traits such as career interests, values and personality. It also serves as an impetus for future work in this area. First, contexts in which statements or activities are compared are likely to affect the consistency of choice for a person (Lin & Brown, 2017), and consequently harm the reliability of $\boldsymbol{\theta}$ estimates in the DPM. Future studies can explore the effects of contexts on MFC items. Second, although empirical data, albeit simulated, have shown that the θ and γ parameters can be well estimated, additional studies are needed to further establish the separability and interpretability of the person parameters. For example, it would be useful to collect experimental data from an intervention study specifically designed to change either the differentiation between the constructs or the intensity of differentiation. Finally, this study used the Bayesian Gibbs sampling method, which can be slow to converge when the number of dimensions is high (e.g., greater than six). Future studies could explore how to implement new estimation methods such as Metropolis–Hastings Robbins–Monro (Cai, 2010) to fit IRT models for MFC items.

It is important to note that although multidimensional ranking items are usually scored by assigning several numbers (ranks) to the statements, they are not the polytomous MFC items discussed in this paper. To clarify, let there be three statements, A, B and C, in a ranking item. These three statements produce three pairwise comparison items: (A, B), (A, C) and (B, C). The statements with higher ranks in these resulting pairwise comparison items will be scored 1, while the others 0 (Brown & Maydeu-Olivares, 2011; Wang et al., 2016). Essentially, the multidimensional ranking items scored in this way are binary MFC items. Nevertheless, they can be extended to be polytomous by asking the respondents to indicate how much they prefer the statements they are ranking. The format of polytomous ranking items is deemed more informative than that of dichotomous ranking items or polytomous pairwise comparison items. Future studies could investigate how the DPM can be adapted to fit the polytomous ranking items.

As a closing remark, it is important to note the issue of whether scores from tests that consist of purely MFC items can be used for interpersonal comparisons remains open. Some authors argue that scales of latent traits can be identified with MFC items (Brown & Maydeu-Olivares, 2011, 2013, 2018; Morillo et al., 2016), while other authors claim that mixing MFC items with some normative items, such as unidimensional Likert-type or unidimensional forced-choice items, is necessary to identify the origin of the scales (Böckenholt, 2004; Chernyshenko et al., 2009; Wang et al., 2017). This study tends to support the latter argument. Until a more unequivocal conclusion can be reached, people interested in using MFC items for interpersonal comparisons should be particularly cautious about how the MFC tests are designed and which specific IRT models are used.

AUTHOR CONTRIBUTIONS

Xuelan Qiu: Conceptualization; data curation; formal analysis; methodology; software; writing – original draft. **Jimmy de la Torre:** Conceptualization; supervision; writing – review and editing.

ACKNOWLEDGEMENTS

This study was supported by the Faculty Research Fund, Faculty of Education, The University of Hong Kong.

Open access publishing facilitated by Australian Catholic University, as part of the Wiley - Australian Catholic University agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

All authors declare that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The computer codes used in the simulation studies and the real data analysis are provided in the Supporting Information. Regarding the empirical data, due to the amount of work involved in developing the new assessment with the polytomous multidimensional forced-choice format and collecting the data, we would like to publish more papers based on more thorough analysis of these data before making them publicly available. However, we are open to sharing part of the data with interested readers, with the stipulation that the data cannot be disseminated or published without our formal permission.

ORCID

Xuelan Qiu https://orcid.org/0000-0002-5446-9758

REFERENCES

Anguiano-Carrasco, C., MacCann, C., Geiger, M., Seybert, J. M., & Roberts, R. D. (2015). Development of a forced-choice measure of typical-performance emotional intelligence. *Journal of Psychoeducational Assessment*, *33*(1), 83–97. https://doi.org/10.1177/0734282914550387

Ashman, A., & Telfer, R. (1983). Personality profiles of pilots. *Aviation, Space, and Environmental Medicine*, *54*(10), 940–943.

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of Personality and Social Psychology*, *87*(2), 266–281. https://doi.org/10.1037/0022-3514.87.2.266

Bartlett, C., Perera, H. N., & Mcllveen, P. (2016). A short form of the career interest test: 21-CIT. *Journal of Career Assessment*, *24*(2), 397–409. https://doi.org/10.1177/1069072715580579

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, *9*(4), 453–465. https://doi.org/10.1037/1082-989X.9.4.453

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

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

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

Brown, A., & Maydeu-Olivares, A. (2018). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 516–529. https://doi.org/10.1080/10705511.2017.1392247

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335. http://www.jstor.org/stable/40785074

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology*, *104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review*, *51*(5), 292–303. https://doi.org/10.1037/h0057299

Chernyshenko, O. S., Stark, S., Prewett, M. S., Gray, A. A., Stilson, F. R., & Tuttle, M. D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance*, *22*(2), 105–127. https://doi.org/10.1080/08959280902743303

Cowles, M., & Carlin, B. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*(434), 883–904. https://doi.org/10.2307/2291683

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. https://doi.org/10.18637/jss.v048.c01

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement*, *32*(5), 355–370. https://doi.org/10.1177/0146621607303784

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research*, *29*(1), 115–126. https://doi.org/10.1207/s15327906mbr2901_4

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistical Sinica*, *6*, 733–807.

Harris, J. A., Vernon, P. A., & Jang, K. L. (2005). Testing the differentiation of personality by intelligence hypothesis. *Personality and Individual Differences*, *38*(2), 277–286. https://doi.org/10.1016/j.paid.2004.04.007

Hirschi, A. (2009). Development and criterion validity of differentiated and elevated vocational interests in adolescence. *Journal of Career Assessment*, *17*(4), 384–401. https://doi.org/10.1177/1069072709334237

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Prentice Hall.

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Jackson, D. N. (1977). *Jackson vocational interest survey manual*. Research Psychologists Press.

Johnson, C. E., Wood, R., & Blinkhorn, S. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology*, *61*(2), 153–162. https://doi.org/10.1111/j.2044-8325.1988.tb00279.x

Josef, M., Wolff, S., & Thomas, F. (2017). The motivational value systems questionnaire (MVSQ): Psychometric analysis using a forced choice Thurstonian IRT model. *Frontiers in Psychology*, *8*, 1626. https://doi.org/10.3389/fpsyg.2017.01626

Kopelman, R. E., Rovenpor, J. L., & Guan, M. W. (2003). The study of values: Construction of the fourth edition. *Journal of Vocational Behavior*, *62*(2), 203–220. https://doi.org/10.1016/S0001-8791(02)00047-7

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

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. CRC Press.

Matthews, G., & Oddy, K. (1997). Ipsative and normative scales in adjectival measurement of personality: Problems of bias and discrepancy. *International Journal of Selection and Assessment*, *5*(3), 169–182. https://doi.org/10.1111/1468-2389.00057

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*(4), 531–551. https://doi.org/10.1348/0963179042596504

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*(3), 1142–1160.

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

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology*, *78*(2), 218–225. https://doi.org/10.1037/0021-9010.78.2.218

Nauta, M. M., & Kahn, J. H. (2007). Identity status, consistency and differentiation of interests, and career decision self-efficacy. *Journal of Career Assessment*, *15*(1), 55–65. https://doi.org/10.1177/1069072705283786

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Paper presented at the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, *108*(2), 370–392. https://doi.org/10.1037//0033-295X.108.2.370

Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology*, *88*(4), 797–834. https://doi.org/10.1111/joop.12098

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *17*, 1–97. https://doi.org/10.1007/BF03372160

SHL. (2006). *OPQ32 technical manual*. SHL Group Ltd.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement*, *29*(3), 184–201. https://doi.org/10.1177/0146621604273988

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., Gelman, A., Goodrich, B., Piironen, J., & Nicenboim, B. (2022). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian Models*. R package version 2.5.1 https://github.com/stan-dev/loo

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing.*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Wang, W.-C., Qiu, X.-L., Chen, C.-W., & Ro, S. (2016). Item response theory models for multidimensional ranking items. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 49–65). Springer.

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*(8), 600–613. https://doi.org/10.1177/0146621617703183

Whitely, S., & Dawis, R. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement*, *11*(3), 163–178. https://doi.org/10.1111/j.1745-3984.1974.tb00988.x

Witkin, H. A., Goodenough, D. R., & Oltman, P. K. (1979). Psychological differentiation: Current status. *Journal of Personality and Social Psychology*, *37*(7), 1127–1145. https://doi.org/10.1037/0022-3514.37.7.1127

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zhu, X., & Stone, C. A. (2012). Bayesian comparison of alternative graded response models for performance assessment applications. *Educational and Psychological Measurement*, *72*(5), 774–799. https://doi.org/10.1177/0013164411434638

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Qiu, X., & de la Torre, J. (2023). A dual process item response theory model for polytomous multidimensional forced-choice items. *British Journal of Mathematical and Statistical Psychology*, *76*, 491–512. https://doi.org/10.1111/bmsp.12303

APPENDIX A: POSTERIOR PREDICTIVE MODEL CHECKING IN THE EMPIRICAL STUDY

As mentioned in the main text, the posterior predictive model checking (PPMC) method with the four discrepancy measures (the frequency of categories of each item, the frequency of categories in the test, the item score distribution, and Yen's Q3 statistic) was used to evaluate the model–data fit of the proposed dual process IRT model (DPM) in the empirical study. This appendix provides the detailed computation of the posterior predictive *p*-value (PPP) for the discrepancy measure $\mathbf{D}(\mathbf{y})$ of an item and the results.

Take the discrepancy measure of the frequency of categories in each item as an example. Let $\mathbf{y}^1, \dots, \mathbf{y}^r, \dots, \mathbf{y}^R$ (r = 1, …, R) be R (R = 100 in this study) sets of sampled responses after burn-in according to the models. The PPP for the discrepancy measure $\mathbf{D}(\mathbf{y})$ of item i was computed as follows.

1. Compute the frequency of category j (j = 0, …, 3) in item i for the rth simulated data set, denoted by $\mathbf{D}\left(\mathbf{y}_i^r\right) = \left[\upsilon_{0i}^r, \dots, \upsilon_{(J-1)i}^r\right]$, where υ is the frequency of category j across persons in replication r and J is the number of categories in item i. The subscript i is omitted for simplicity of notation in the following steps.
2. Repeat step 1 for the R samples. Collect $\mathbf{D}(\mathbf{y}^r)$ for a total of 100 samples into a matrix form, as $\mathbf{D}(\mathbf{y}^{\text{rep}})$.
3. Sort $\mathbf{D}(\mathbf{y}^{\text{rep}})$ for each category and obtain the value of $\mathbf{D}(\mathbf{y})$ at the 2.5% and 97.5% points, denoted by $\mathbf{D}_{.025}(\mathbf{y})$ and $\mathbf{D}_{.975}(\mathbf{y})$, respectively. The 95% credible intervals of $\mathbf{D}(\mathbf{y})$ for the items are thus defined using $\mathbf{D}_{.025}(\mathbf{y})$ and $\mathbf{D}_{.975}(\mathbf{y})$.
4. Compute the frequency of each category for the observed data, denoted by $\mathbf{D}\left(\mathbf{y}^{\text{obs}}\right)$.
5. Compute the *p*-value of $\mathbf{D}(\mathbf{y})$ ($p_{\mathbf{D}(\mathbf{y})}$) as the proportion of $\mathbf{D}\left(\mathbf{y}^{\text{obs}}\right)$ among the sorted $\mathbf{D}(\mathbf{y}^{\text{rep}})$ for each category. If the $p_{\mathbf{D}(\mathbf{y})}$ is beyond a critical range (e.g., .025–.975), it should be concluded that the chosen model does not fit the observed data (Meng, 1994).

The method was similarly employed for the other three discrepancy measures. For the frequency of categories across items, the PPP was the averaged $p_{\mathbf{D}(\mathbf{y})}$ across items. For the score distribution, it was the averaged $p_{\mathbf{D}(\mathbf{y})}$ across the four categories. For Yen's Q3 measure, it was the averaged $p_{\mathbf{D}(\mathbf{y})}$ across the 630 item pairs.

For the PPMC results with the discrepancy measure of the frequency of categories in each item, because as many as 36 items were used, a graphical display was used to evaluate the results. The observed frequency $\mathbf{D}\left(\mathbf{y}^{\text{obs}}\right)$ was plotted against the 95% credible intervals of $\mathbf{D}(\mathbf{y})$ for each category in each item. A poorly fitting model can be immediately identified by the extremeness of the observed frequency $\mathbf{D}\left(\mathbf{y}^{\text{obs}}\right)$ beyond the 95% credible intervals of $\mathbf{D}(\mathbf{y})$ for the item. Figure A1 shows the observed frequency $\mathbf{D}\left(\mathbf{y}^{\text{obs}}\right)$ for each of the 36 polytomous MFC items and the 95% credible intervals of $\mathbf{D}(\mathbf{y})$ for the items obtained from the replicated data sets under the DPM. It appears that for the majority of the items, the frequency from the observed data set was within the 95% credible intervals. The number of items beyond the 95% credible intervals for the categories was 4, 0, 0, and 2, respectively.

The results of PPMC with the other three discrepancy measures were as follows. For the frequency of categories in the test, the PPPs for j = 0, 1, 2 and 3 under the DPM were .272, .584, .488, and .695, respectively. For the item score distribution, the PPP was .322. For Yen's Q3 statistic, the PPP was .343. All of these PPPs were within the critical range of .025 to .975, indicating a good model–data fit.

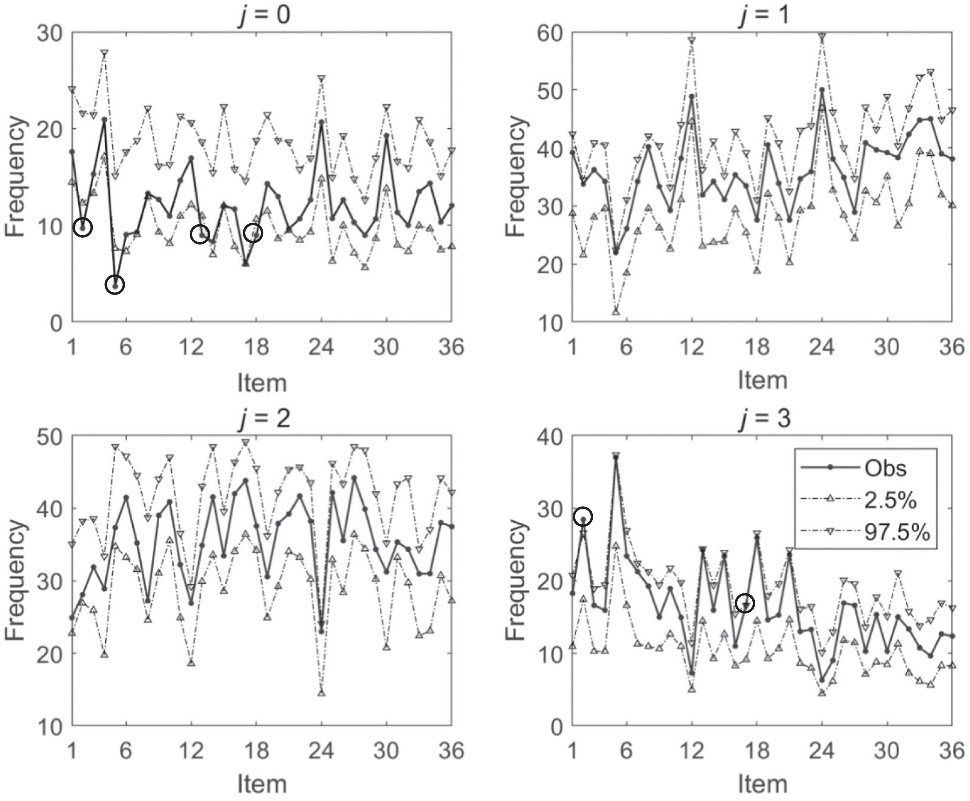

FIGURE A1 Results of posterior predictive model checking for the DPM with discrepancy measure of frequency of categories in each item. *Note*: Obs = observed data; Items beyond the 95% credible intervals are circled.