Item Response Theory Models for Ipsative Tests With Multidimensional Pairwise Comparison Items

Applied Psychological Measurement
2017, Vol. 41(8) 600–613
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617703183
journals.sagepub.com/home/apm

Wen-Chung Wang¹, Xue-Lan Qiu¹, Chia-Wen Chen¹,
Sage Ro², and Kuan-Yu Jin¹

Abstract

There is re-emerging interest in adopting forced-choice items to address the issue of response bias in Likert-type items for noncognitive latent traits. Multidimensional pairwise comparison (MPC) items are commonly used forced-choice items. However, few studies have been aimed at developing item response theory models for MPC items owing to the challenges associated with ipsativity. Acknowledging that the absolute scales of latent traits are not identifiable in ipsative tests, this study developed a Rasch ipsative model for MPC items that has desirable measurement properties, yields a single utility value for each statement, and allows for comparing psychological differentiation between and within individuals. The simulation results showed a good parameter recovery for the new model with existing computer programs. This article provides an empirical example of an ipsative test on work style and behaviors.

Keywords

item response theory, Rasch model, ipsative data, forced-choice items, pairwise comparison

Tests (including examinations, inventories, questionnaires, and observations) are popular data collection methods in the human sciences. In general, tests can be classified into two categories: normative and ipsative (from the Latin *ipse*: he, himself; Cattell, 1944). Normative tests allow for comparing different persons along some latent traits (e.g., who is more proficient in problem solving, who is more motivated to learn). Most tests for personnel selection and college admission are normative, because decision making relies on interindividual comparisons. As the name ipsative implies, the major purpose of ipsative tests is to compare different latent traits within persons. For example, in career guidance, it is important for each client to identify an area worth pursuing among many careers, rather than comparing with others.

When using normative tests, concerns arise regarding their assessment of noncognitive traits (e.g., personality, values, and interests). Take career interest tests as an example. A typical

¹The Education University of Hong Kong, Hong Kong

²IBM, Minneapolis, MN, USA

Corresponding Author:

Wen-Chung Wang, Department of Psychology, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, New Territories, Hong Kong.

Email: wcwang@eduhk.hk

normative item would be the following: “How much do you like the following activities: attending parties, visiting museums, etc., on a 5-point rating scale (e.g., *not at all*, *very little*, *a little*, *quite a lot*, *a great deal*)?” Two prominent issues arise with such a normative test design (Matthews & Oddy, 1997). The first issue is social desirability. A job applicant’s intention to make a good impression so as to increase his or her chance of getting hired can distort the responses on normative tests. The second issue is a problem of differentiation. If a person responds consistently in a single category (e.g., *a great deal*) throughout all items, it is impossible to differentiate his or her career interests. Low differentiation makes the tests practically useless in such a setting, where the purpose of tests is to provide advice on career choices.

To resolve these two issues, researchers have proposed the use of forced-choice items in areas such as career guidance assessments and workforce assessments. Among the various forced-choice items, multidimensional pairwise comparison (MPC) items are popular; they consist of a pair of statements that are similar in social desirability, but that measure different dimensions. Respondents are required to choose the statement that they prefer over another statement. For example, a typical MPC item is “Which activity do you prefer: attending parties or visiting museums?” The first statement of attending parties is designed to measure a latent trait of social interest, whereas the second one of visiting museums is designed to measure a latent trait of artistic interest. If two statements in a pair have very different degrees of social desirability, it will be easy for respondents to “fake good” or “fake bad.” Popular examples of tests using the MPC format include the Jackson Vocational Interest Survey (Jackson, 1977), Edwards Personal Preference Schedule (Ashman & Telfer, 1983), Chinese Personality at Work (Hui, Pak, & Cheng, 2009), and the Allport–Vernon–Lindzey Study of Values (SOV; Kopelman, Rovenpor, & Guan, 2003). In addition, the Programme for International Student Assessment 2012 adopted the MPC format to measure students’ mathematics intentions (Organisation for Economic Co-Operation and Development, 2014).

Responses to MPC items are typically scored as a summed frequency (or simply a summed total score)—that is, how many times those statements measuring the same latent trait are preferred across MPC items. The outcome is person-centered scores (or ipsative scores), because the number-correct total scores are the same for all people (Meade, 2004). The following example explicates the nature of ipsative tests with MPC items.

Let there be 10 MPC items measuring two latent traits: social interest and artistic interest. Of the 10 MPC items, Person A selects statements measuring social interest 6 times and statements measuring artistic interest 4 times, resulting in number-correct total scores of 6 for social interest and 4 for artistic interest. Person B receives number-correct scores of 9 for social interest and 1 for artistic interest. They (in fact, all persons) scored 10 points, which is the total number of MPC items. Given the result, it is legitimate to draw the following two conclusions:

1. Both Persons A and B prefer social activities to artistic activities, because the number-correct scores for social activities are higher than those for artistic activities.
2. Relatively speaking, the preference for social activities over artistic activities is weaker in Person A than in Person B, because the odds of preferring social activities to artistic activities are 6:4 for Person A and 9:1 for Person B.

One may wonder whether the following conclusion could be drawn:

3. The social interest level of Person A is lower (6 points) than that of Person B (9 points).

Conclusion 3, drawn from a normative comparison, is problematic in this situation because the scores are person-centered or ipsative. Person A, who scored 6 on social activities and 4 on

artistic activities, may be interested in both social activities and artistic activities, but shows a preference for social activities when comparing these two types of activities. By contrast, Person B, who scored 9 and 1, respectively, may not be interested at all in either social activities or artistic activities, but shows a very strong preference for social activities when comparing these activities. Thus, Conclusion 3 is invalidated in ipsative tests.

Böckenholt (2004) suggested three approaches to identify the latent traits on normative scale metric, including administering additional normative items with the ipsative tests. Following the suggestion, researchers (McCloy, Heggstad, & Reeve, 2005; Stark, Chernyshenko, & Drasgow, 2005) proposed embedding a small number of unidimensional Likert-type items in the ipsative tests and employing standard item response theory (IRT) models to calibrate the utility of each statement. Then, these utilities are treated as known in MPC data to obtain person parameter estimates. As the statement utilities are estimated from normative tests, the person estimates are on a normative scale as well. In this case, the latent traits on normative scores can be identified and Conclusion 3 can be justified.

Existing IRT Models for Ipsative Tests With MPC Items

Many IRT models exist for normative tests. However, few have emerged in the development of IRT models for ipsative tests, because their ipsative nature precludes their applicability to common IRT models. Brown (2016a) provided a comprehensive and excellent summary of some IRT models. In the literature, there are two major approaches for ipsative tests: the ideal-point approach and the dominance approach. In the ideal-point approach, the distance between a person's location and a statement's location determines the response function, where the shorter the distance, the greater the probability of endorsement. The models include the hyperbolic cosine model for pairwise preferences (Andrich, 1995), McCloy–Heggstad–Reeve's unfolding model (McCloy et al., 2005), and the multi-unidimensional pairwise preference model (MUPP; Stark et al., 2005). For example, in the MUPP, the probability of preferring statement s to statement t in item i for person n , $P_{in}(s|(s, t))$, is defined as follows:

$$P_{in}(s|(s, t)) = \frac{P(1, 0)}{P(1, 0) + P(0, 1)} \approx \frac{P_s(1)P_t(0)}{P_s(1)P_t(0) + P_s(0)P_t(1)}, \quad (1)$$

where $P(1, 0)$ is the joint probability of preferring statement s and not preferring statement t ; $P(0, 1)$ indicates the joint probability of not preferring statement s and preferring statement t ; $P_s(1)$ represents the probability of preferring statement s ; $P_s(0) = 1 - P_s(1)$; $P_t(1)$ is the probability of preferring statement t ; and $P_t(0) = 1 - P_t(1)$. In the MUPP, $P_s(1)$ and $P_t(1)$ follow the generalized graded unfolding model (Roberts, Donoghue, & Laughlin, 2000). Recently, the MUPP has been applied to adaptive tests (Stark, Chernyshenko, Drasgow, & White, 2012) and extended to handle (partial) ranking items, which have more than two statements in an item (Hontangas et al., 2015).

In the dominance approach, persons are assumed to select the stimulus with the highest utility where the term "utility" represents the property of an object that produces benefits or advantages (Thurstone, 1927). Based on their previous work on IRT models for unidimensional paired comparison and ranking data (Maydeu-Olivares & Brown, 2010), Brown and Maydeu-Olivares (2011) developed a Thurstonian IRT model for MPC items which can be expressed as follows:

$$P_{in}(s|(s, t)) = \Phi(\alpha_i + \beta_{si}\eta_{an} - \beta_{ti}\eta_{bn}), \quad (2)$$

where α_i is the location parameter for item i (it can be viewed, in some sense, as the difference in utility between the two statements); β_{si} and β_{ti} are the slope parameters of statements s and t

in item i , respectively; η_{an} and η_{bn} are the levels of latent trait a (measured by statement s) and trait b (measured by statement t); Φ is the cumulative standard normal distribution function; and others are defined as above. It is possible to recover the parameters using *Mplus* computer software when the items consist of positively and negatively keyed statements. The model has been applied to real tests (Guenole, Brown, & Cooper, 2016) and extended to compositional questionnaire data (Brown, 2016b).

In the following sections, the new Rasch ipsative model (RIM) for MPC items in the dominance approach is introduced, and the fundamental differences between the RIM and the Thurstonian model are shown because both models are proposed in the dominance approach. A series of simulations are conducted to assess parameter recovery using the computer program ACER ConQuest (Adams, Wu, & Wilson, 2015) or JAGS (Plummer, 2003), and the results are summarized. An empirical example to demonstrate the implications and applications of the new model is provided. Finally, conclusions are drawn, and suggestions for future studies are provided.

RIM for MPC Items

Let there be D latent traits (e.g., social, artistic, and investigative interests) in the test. Let item i consists of statements s and t , which measure latent traits a and b , respectively. In the RIM, the logit of selecting statement s over statement t in item i for person n is defined as follows:

$$\text{logit}_{in} \equiv \log \left(\frac{P_{in}(s|(s,t))}{1 - P_{in}(s|(s,t))} \right) = (\theta_{an} + \delta_{si}) - (\theta_{bn} + \delta_{ti}), \quad (3)$$

where θ_{an} and θ_{bn} represent the levels of latent traits a and b for person n , respectively; δ_{si} and δ_{ti} are the utilities of statements s and t , respectively, in item i ; and others are defined as above. To be in line with the ipsative nature, where the sum of the number-correct score is constant for every person (Meade, 2004), the θ variables sum to zero across D latent traits (i.e., $\sum_{d=1}^D \theta_d = 0$) for every person. Thus, only $D - 1$ θ latent traits can be freely estimated. Taking one trait (e.g., the D th θ variable) arbitrarily as fixed, it can be calculated as $\theta_D = -\sum_{d=1}^{D-1} \theta_d$ for every person. The θ variables represent the degrees of psychological differentiation (Witkin, Goodenough, & Oltman, 1979), with an extreme value denoting a greater differentiation than a value close to zero. These variables are useful for assessing personality differentiation (Harris, Vernon, & Jang, 2005), interest differentiation (Holland, 1973), and emotion differentiation (Barrett, 2004).

As Equation 3 illustrates, the logit is determined by two competing parts. The first part is the sum of the utility of statement s and the corresponding latent trait a , and the second part is the sum of the utility of statement t and the corresponding latent trait b . Each part reflects how attractive the statement is to the person: the higher the value, the more attractive the statement is. When the first part is greater than the second part (i.e., the first part wins), the logit will be positive and the probability of selecting statement s will be greater than 0.5; when both parts are equal, the logit will be 0 and the probability will be equal to 0.5; when the difference between the two parts is negative, the logit will be negative and the probability will be smaller than 0.5. In essence, the RIM inherits Thurstone's (1927) Law of Comparative Judgment and extends to MPC items measuring multiple latent traits. It is assumed that when a person is required to indicate a preference on a pair of statements, the person will evaluate each statement separately to determine the total amount of attractiveness (e.g., $\theta_a + \delta_s$) and then select the statement that has a higher total amount of attractiveness.

Figure 1. Logit of selecting statement s over statement t across different values of $\theta_a - \theta_b$ and $\delta_s - \delta_t$ in the Rasch ipsative model.

The RIM has four important features. First, it yields a single utility value for each statement, regardless of what is in the other statement. This property is in line with most models, such as Thurstone's (1927) models, and is critical for measuring a statement's utility.

Second, the response function is monotonically increasing, as demonstrated in Figure 1. A larger value of $\theta_a - \theta_b$ leads to a higher probability of preferring statement s to statement t . Thus, between-person comparison is enabled on the θ variables. Furthermore, by definition, a positive value of $\theta_a - \theta_b$ indicates that a person has a higher level of latent trait a than latent trait b ; a positive value of $\theta_b - \theta_c$ indicates that a person has a higher level of latent trait b than latent trait c . Thus, within-person comparison is also enabled on the θ variables.

Third, the RIM has a good measurement property of specific objectivity (i.e., item and person parameters can be separated). When persons n and n' respond to item i with statements (s , t), it follows from Equation 3 that

$$\text{logit}_{i,n} = (\theta_{an} + \delta_s) - (\theta_{bn} + \delta_t), \quad (4)$$

$$\text{logit}_{i,n'} = (\theta_{an'} + \delta_s) - (\theta_{bn'} + \delta_t). \quad (5)$$

Then,

$$\text{logit}_{i,n} - \text{logit}_{i,n'} = (\theta_{an} - \theta_{bn}) - (\theta_{an'} - \theta_{bn'}), \quad (6)$$

which is independent of item parameters. Likewise, when a person responds to two items i and i' with statements (s , t) and (s' , t'), respectively, it follows that

$$\text{logit}_{i,n} = (\theta_{an} + \delta_s) - (\theta_{bn} + \delta_t), \quad (7)$$

$$\text{logit}_{i',n} = (\theta_{an} + \delta_{s'}) - (\theta_{bn} + \delta_{t'}),$$

Then,

$$\text{logit}_{i,n} - \text{logit}_{i',n} = (\delta_s - \delta_t) - (\delta_{s'} - \delta_{t'}), \quad (9)$$

which is independent of person parameters.

Fourth, when making a pairwise comparison within the same latent trait (e.g., both statements s and t measure latent trait a), the RIM becomes

$$\text{logit}_{i,n} = (\theta_{an} + \delta_s) - (\theta_{an} + \delta_r) = \delta_s - \delta_r, \quad (10)$$

which is equivalent to the Bradley–Terry–Luce model (Bradley & Terry, 1952; Luce, 1959).

The RIM Versus the Thurstonian IRT Model

As Equation 2 demonstrates, the Thurstonian IRT model is a two-parameter model, whereas the RIM (Equation 3) is a one-parameter model. One may question whether the RIM is a special case of the Thurstonian IRT model when the β parameters in Equation 2 are set at 1. Actually, they are fundamentally different. The Thurstonian model aims to recover the η variables that represent the normative or absolute levels of latent traits (without the constraint of zero sum across dimensions for every person), whereas the RIM aims to recover the θ variables (with the constraint of zero sum across dimensions for every person to reflect the ipsative nature). Even when it is possible to recover the η variables, it is crucial to check whether they have good measurement properties.

Take the following case as an example. Let there be two items with equal-utility statements (i.e., $\alpha_i = 0$), and Item 1 has $\beta_{s1} = 1$ and $\beta_{r1} = 0.5$, and Item 2 has $\beta_{s2} = 1$ and $\beta_{r2} = 1.5$ (for ease of interpretation, the two items have both positive-keyed statements); let there be two persons, where Person 1 has $\eta_{a1} = \eta_{b1} = 0$, and Person 2 has $\eta_{a2} = \eta_{b2} = -1$. When these two persons respond to the two MPC items, according to Equation 2, the probabilities of preferring the first statement in both items are 0.5 for Person 1, but are approximately 0.3 in Item 1 and 0.7 in Item 2 for Person 2.

As the η variables are to represent the absolute levels of latent traits, it is logical to anticipate that Person 1 should have a probability of preferring the first statement (measuring latent trait a) of an item higher than that of Person 2, because $\eta_{a1} > \eta_{a2}$. This is analogous to the logical anticipation that a person with a higher ability should have a probability of success on an item higher than that of a person with a lower ability. If the anticipation does not hold (e.g., Person 1 has a probability equal to or lower than that of Person 2), then the η variables cannot be compared between persons. Actually, such an anticipation holds for Item 1 but not for Item 2.

Similar logic can be applied to the comparison within persons. For both persons, as they have an identical level on the two latent traits and the statements in an item have an identical utility, it is logical to anticipate that for both persons, the probability of preferring the first statement should be 0.5 for both Items 1 and 2. Actually, such an anticipation holds within Person 1 but not within Person 2. All in all, the scaled score in the Thurstonian model cannot be used for between- and within-individual comparisons.

The Thurstonian model relies substantially on the items that consisted of both positively and negatively keyed statements to yield satisfactory estimates of item and person parameters. The parameter recovery is very poor when all statements are positively keyed (Brown & Maydeu-Olivares, 2011). However, pairing statements of different directions in an item may be problematic in practice. First, when encountering two statements like “I keep my paperwork in order” and “I struggle to keep my temper” (this is a sample item in Brown and Maydeu-Olivares, 2011), it is very likely that participants (e.g., job applicants) tend to select the positive statement that is obviously more socially desirable regardless of whether it is true. Consequently, almost all persons will choose the same statement, yielding little information about persons. Second, changing the wording of the statements may essentially change the construct being measured by these statements (Greenberger, Chen, Dmitrieva, & Farruggia, 2003). In other words, reverse coding of negative statements can be problematic.

Finally, the Thurstonian model does not yield unique utility for an individual statement. Suppose there are two latent traits and each has 10 statements, there will be 100 possible MPC items. According to Equation 2, for item i , only the difference between the utilities of the two

statements in an item (i.e., α_i , $i = 1, \dots, 100$) is estimated, but the utility for each statement is not identifiable. Consequently, it is impossible to answer the simple question “What is the utility of the described statement?”

In the RIM, as Person 1 has $\eta_{a1} = \eta_{b1} = 0$, and Person 2 has $\eta_{a2} = \eta_{b2} = -1$, according to the constraint that the θ variables must sum to 0 for every person, both persons have $\theta_a = \theta_b = 0$. In other words, the two persons have no differentiation between the statements. As the statements in each item have an equal utility ($\delta_{si} = \delta_{ti}$), it is logical to anticipate that both persons should have a probability of preferring the first statement of 0.5 in each item. According to Equation 3, the logical anticipation is met. The scales of the latent traits in the RIM are invariant. Finally, the RIM yields a single estimate for each statement to represent its utility.

Parameter Estimation of the RIM

The RIM can be identified by setting either the mean of each latent trait across persons at zero or the mean utility of each latent trait across statements at zero. The parameters can be estimated using the marginal maximum likelihood estimation (MMLE) method or the Bayesian method with Markov Chain Monte Carlo (MCMC) sampling. For the former, one can use computer programs such as ConQuest (Adams et al., 2015) or TAM in R (Kiefer, Robitzsch, & Wu, 2016); for the latter, one can employ computer programs such as WinBUGS (Spiegelhalter, Thomas, & Best, 2003) or JAGS (Plummer, 2003). In the authors’ experience, the estimation of a small number of latent traits (e.g., less than four) using the MMLE method can converge within several minutes; when the number of latent traits is large, the Bayesian method is preferable.

For the Bayesian method, the full posterior distribution of the parameters given the data is as follows:

$$P(\boldsymbol{\theta}, \boldsymbol{\delta} | \mathbf{Y}) \propto \prod_{n=1}^N \prod_{i=1}^I P(y_{ni} | \boldsymbol{\theta}_n) P(\boldsymbol{\theta}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma}) P(\boldsymbol{\mu} | \boldsymbol{\Sigma}) P(\boldsymbol{\Sigma}) P(\boldsymbol{\delta}), \quad (11)$$

where $P(y_{ni} | \boldsymbol{\theta}_n)$ denotes the probability of choosing the first statement on item i for person n with latent trait $\boldsymbol{\theta}_n$, calculated as Equation 3 for the RIM; the $D - 1$ ability parameters $\boldsymbol{\theta}_n$ ($n = 1, \dots, N$) follow a multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. That is, $\boldsymbol{\theta}_n \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $\mathbf{Y}$ is the item responses; $P(\boldsymbol{\theta}_n | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the conditional probability of latent trait $\boldsymbol{\theta}_n$ for person n ; and $P(\boldsymbol{\mu} | \boldsymbol{\Sigma})$, $P(\boldsymbol{\Sigma})$, $P(\boldsymbol{\delta})$ are the priors for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\delta}$, respectively.

Simulation Studies

Simulation Design

A series of simulations were conducted to evaluate the parameter recovery of the RIM. Similar to previous studies (e.g., Brown & Maydeu-Olivares, 2011; Wang, Qiu, Chen, & Ro, 2016), four variables were manipulated: (a) number of dimensions: $D = 2, 3, 5$; (b) number of statements in each dimension: $L = 6$ and 12; (c) sample size: $N = 500$ and 1,000; and (d) linking design: complete linking and minimum linking. In the complete linking, each statement was paired with all statements in other dimensions, resulting in a maximum number of items of $\binom{T}{2} - \sum_{d=1}^D \binom{L}{2}$, where T is the total number of statements and L is defined as above. The minimum linking was adapted from the balanced incomplete block method (Johnson, 1992), where each statement is paired twice, as shown in Figure 1 of Wang et al. (2016).

This study considered linking design because it is often (if not always) desirable to put the utilities of statements on the same scale for comparison (e.g., which statement has the highest or lowest utility). To the best of the authors' knowledge, existing ipsative tests seldom have a linkage such that comparisons between statements are possible; only Stark et al. (2012) discussed and evaluated different linking designs for ipsative tests via simulations. This situation is analogous to the cases where different persons take different tests without proper linkages. With complete linking or minimum linking, all cross-dimension statements were connected so their utilities were placed on a common scale for direct comparison. As the minimum linking had much fewer data points than the complete linking, the parameter recovery would be less accurate.

A total of 20 conditions were simulated (for $D = 3$ and 5, there are 432 items and 1,440 items, respectively, under the complete linking design with $L = 12$; they were not simulated due to heavy computational burden). The RIM (Equation 3) was used to generate item responses. The statement utilities were sampled from $N(0, 1)$. The $D - 1$ latent traits were assumed to follow a multivariate standard normal distribution with a mean zero vector and a variance-covariance matrix with a variance of 1 and covariance of -0.2 . The level for the D th latent trait was computed as $\theta_D = -\sum_{d=1}^{D-1} \theta_d$ for every person, in line with the ipsative nature of the test. The simulated data were analyzed with the data-generating model using the computer program ConQuest (Adams et al., 2015). The generating values were used as the initial values of the parameters to facilitate the estimation. Alternatively, one may obtain approximate estimates with smaller nodes and use them as initial values for the second analysis to produce more accurate estimates with larger nodes. In the authors' experiences, the final parameter estimates were very similar. One hundred replications were conducted in each condition.

In addition, a condition that mimicked the design of the empirical example was simulated using the person and statement estimates as the generating values. JAGS was used to estimate the parameters because the MCMC method is more efficient for high dimensions (12 latent traits in the example). The priors and the settings for the MCMC procedure were the same as those in the empirical example. Thirty replications were conducted for this condition due to extensive computation time. The program codes of ConQuest and JAGS are available on request.

The bias and root mean square error (RMSE) for the statement utility and person distributional parameters were computed to evaluate parameter recovery. It was expected in general that the larger the sample size, the better the parameter recovery; the complete linking would yield a more accurate estimation than the minimum linking.

Results

Table 1 provides the mean and standard deviation of the bias and RMSE for the statement utilities parameters (δ), mean level of latent traits (μ), and variance-covariance of latent traits (σ) under the condition of six statements in each dimension ($L = 6$). According to Table 1, for the complete linking, the parameter recovery was fairly good across different numbers of dimensions and sample sizes. For example, under condition of $D = 5$ and $N = 500$, the bias values were close to 0 for δ s ($M = 0.000$, $SD = 0.006$), for μ s ($M = 0.000$, $SD = 0.005$), and for σ s ($M = -0.009$, $SD = 0.032$); and the RMSE values were rather small for δ s ($M = 0.022$, $SD = 0.003$), for μ s ($M = 0.000$, $SD = 0.003$), and for σ s ($M = 0.065$, $SD = 0.019$). For the minimum linking, the parameter recovery was satisfactory. For example, under condition of $D = 5$ and $N = 500$, the bias values were comparable with that of the complete linking. As expected, the RMSE values were relatively larger in the minimum linking than in the complete linking for δ s ($M = 0.196$, $SD = 0.012$), for μ s ($M = 0.003$, $SD = 0.002$), and for σ s ($M = 0.088$, $SD = 0.026$). The results for $L = 12$ had similar patterns and are provided as an online supplement. Compared with

Table 1. Summary of Parameter Recovery for the RIM With Six Statements in Each Dimension.

Dim	Par		N = 500				N = 1,000			
			Complete linking		Minimum linking		Complete linking		Minimum linking	
			Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
D = 2	δ	M	0.000	0.046	0.000	0.172	0.000	0.032	0.000	0.131
		SD	0.004	0.004	0.012	0.007	0.003	0.002	0.011	0.006
	μ	M	-0.008	0.048	-0.007	0.050	0.004	0.032	-0.003	0.034
		SD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	σ	M	-0.015	0.073	-0.033	0.084	-0.017	0.049	-0.031	0.061
		SD	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
D = 3	δ	M	0.000	0.031	0.000	0.154	0.000	0.022	0.000	0.104
		SD	0.005	0.003	0.013	0.010	0.002	0.002	0.010	0.005
	μ	M	0.000	0.000	-0.001	0.001	0.000	0.000	-0.001	0.001
		SD	0.005	0.000	0.031	0.001	0.003	0.000	0.023	0.000
	σ	M	-0.001	0.078	-0.002	0.094	-0.002	0.054	0.000	0.072
		SD	0.002	0.027	0.008	0.027	0.009	0.014	0.006	0.021
D = 5	δ	M	0.000	0.022	0.000	0.196	0.000	0.017	0.000	0.145
		SD	0.006	0.003	0.027	0.012	0.007	0.003	0.024	0.011
	μ	M	0.000	0.000	-0.005	0.003	0.000	0.000	0.000	0.001
		SD	0.005	0.003	0.031	0.002	0.002	0.000	0.016	0.001
	σ	M	-0.009	0.065	0.000	0.088	-0.009	0.050	-0.001	0.065
		SD	0.032	0.019	0.010	0.026	0.032	0.019	0.011	0.018

Note. RIM = Rasch ipsative model; Dim = dimension; Par = parameters; RMSE = root mean square error; δ = utility parameter; μ = mean vector of latent traits; σ = variance-covariance of latent traits.

the results of $L = 6$, the bias and RMSE of $L = 12$ were relatively smaller under the complete linking because of longer test length, but were relatively larger under the minimum linking because of fewer data points. For the brief simulation study mimicking the empirical example, it was found that both item and person parameters can be recovered satisfactory using JAGS, and the results are provided in online supplement.

To further examine the parameter recovery, a four-way factorial ANOVA of the bias and RMSE values were conducted on δ s, μ s, and σ s across number of dimensions, number of statements in each dimension, sample size, and linking design. For the bias, the results showed that the factors had no significant main effect or interaction effect. For the RMSE, statistically significant effects included the main effects of all four factors (partial η^2 are .022, .034, .066, and .300, respectively), and two-way interaction effect between the number of dimensions and linking design (partial η^2 is .010), between the number of statements in each dimension and linking design (partial η^2 is .013), and between sample size and linking design (partial η^2 is .012). In general, a larger number of dimensions, a larger number of statements in each dimension, a smaller sample size, and the minimum linking produced a larger RMSE.

An Empirical Study

Data and Analysis

To demonstrate the implications and applications of the RIM for MPC items, a dataset was decomposed from a real ipsative test with triad ranking items and analyzed. In the decomposed data, there were 88 MPC items with 88 statements, where the minimum linking was fulfilled such that all statements were linked and placed on the same scale. The test design is provided

Figure 2. Observed frequencies of selecting the first statement in each MPC item and the 95% credible intervals from replicated datasets under the Rasch ipsative model.

Note. The items are sorted according to their frequencies in selecting the first statement in an MPC item. MPC = multidimensional pairwise comparison.

as an online supplement. Each statement was designed to measure one of the 12 latent traits on work style and behaviors: Energy, Assertiveness, Sociability, Concern for Others, Dependability, Organization, Achievement Orientation, Initiative, Multitasking, Innovation, Self-Confidence, and Self-Control. Each latent trait was measured by six to nine statements ($M = 7.3$ statements). An example item is “Which statement is more like you: Trying things that stretch my abilities versus being seen as a winner?”

The RIM was fit to 1,594 responses using JAGS. The mean utility of each dimension across statements was set at zero for model identification. The mean vector and covariance matrix for the 11 θ variables were freely estimated and the 12th θ variable was the negative sum of the other 11 θ variables. The priors for the utility parameters and the mean vector were set as $N(0, 1)$. The priors for the covariance matrix were set as an inverse Wishart distribution $[\mathbf{R}, K]$ with $\mathbf{R} = \mathbf{I}$ and hyperparameter $K = 11$. The MCMC procedure ran 10,000 iterations with a burn-in length of 1,000.

The posterior predictive model checking method was used to examine the model–data fit. This method assumes that when the chosen model fits the observed data $\mathbf{y}^{\text{obs}}$, the replicated data $\mathbf{y}^{\text{rep}}$, which were simulated from the posterior predictive distribution with the model’s parameters, become very similar to $\mathbf{y}^{\text{obs}}$ (Gelman, Meng, & Stern, 1996). This study used the frequency of selecting the first statement in each MPC item as the discrepancy measure and evaluated the difference in this statistic between the observed and replicated data.

Results

Figure 2 shows the observed frequency of selecting the first statement in each item and the 95% credible intervals obtained from the replicated datasets. It appears that the frequency of the observed dataset was within the 95% credible intervals, suggesting that the model–data fit was acceptable.

Table 2. Means, Standard Errors, and Reliabilities of the Latent Traits in the Empirical Example.

	Energy	Asser	Soc	Conc	Depend	Organ	Achiev	Init	Mtask	Innov	Conf	Ctrl
L_d	8	6	6	8	8	9	6	7	9	8	6	7
M	0.162	−0.080	−0.025	−0.166	0.013	−0.083	−0.122	0.172	0.016	0.151	0.232	−0.270
SE	0.018	0.020	0.022	0.018	0.009	0.026	0.016	0.014	0.017	0.014	0.014	0.016
Reliability	.58	.56	.62	.28	.30	.59	.32	.44	.37	.41	.50	.57

Note. Energy = Energy; Asser = Assertiveness; Soc = Sociability; Conc = Concern for Others; Depend = dependability; Organ = Organization; Achiev = Achievement Orientation; Init = Initiative; Mtask = Multitasking; Innov = Innovation; Conf = Self-Confidence; Ctrl = Self-Control; L_d = number of statement in dimension d .

Individual statement utilities ranged from -1.41 to 1.01 ($M = 0.00$, $SD = 0.46$). Statement 5 measuring Dependability had the lowest utility (-1.410), while Statement 35 measuring Dependability had the highest utility (1.012). Table 2 shows the means, their standard errors, and the reliabilities of the 12 θ variables. Test reliability was calculated as $[\text{Var}(\theta_{\text{EAP}}) / \text{Var}(\theta)]$, where $\text{Var}(\theta_{\text{EAP}})$ was the estimated variance of expected a posteriori (EAP) estimates for person measures, and $\text{Var}(\theta)$ was the estimated variance of the person distribution (Mislevy, Beaton, Kaplan, & Sheehan, 1992). Self-Confidence displayed the highest mean (0.23), whereas Self-Control showed the lowest mean (-0.27). Test reliabilities were between $.28$ and $.62$. The reliabilities seem to be low due to the small number of statements in each dimension (6-9 per variable) and small variations in participants' levels on some latent traits (e.g., Concern for Others). The 12 θ variables were moderately correlated. Innovation and Self-Confidence showed the highest positive correlation ($.57$), while Sociability and Organization showed the largest negative correlation ($-.49$).

For each applicant, the mean and standard deviation of the posterior distribution on each θ variable were computed as the point estimate and its standard error, respectively. Figure 3(a) provides the profiles of point estimates on the 12 θ variables of Persons 8 and 10. Person 8 had a small variation on the 12 latent traits, whereas Person 10 had a large variation. It suggests that Person 8 was less differentiated among the latent traits than Person 10. Figure 3(b) presents the point estimates across these 12 θ variables for Persons 8 and 10, with variances of 0.07 for Person 8 and 0.27 for Person 10. The variance of the 12 θ variables for each person reflected the degree of psychological differentiation. For Person 8, the highest estimate was Energy (0.51) and the lowest estimate was Multitasking (-0.28); for Person 10, the highest estimate was Self-Confidence (0.78) and the lowest estimate was Sociability (-1.13). From the two cases, it seems reasonable to associate traits like Energy, Assertiveness, or Sociability with Person 8, and traits like Self-Confidence, Initiative, or Organization with Person 10.

Discussion and Conclusion

To meet the demand for forced-choice tests with MPC items in the noncognitive assessment and be consistent with the traditional scoring approach to ipsativity data, the RIM was proposed, which possesses desirable measurement properties, yields a single utility value for each statement, and allows for comparison in psychological differentiation between and within individuals. In the RIM, the relationship between item parameters and person parameters is specified so as to allow for parameter separation (i.e., specific objectivity).

Simulations studies were conducted to evaluate the parameter recovery of the new model by using ConQuest. The results suggested that parameters can be recovered fairly well with

Figure 3. (a) Profiles and (b) locations of the 12 latent traits for two persons in the empirical example. Note. 1 = Energy; 2 = Assertiveness; 3 = Sociability; 4 = Concern for Others; 5 = Dependability; 6 = Organization; 7 = Achievement Orientation; 8 = Initiative; 9 = Multitasking; 10 = Innovation; 11 = Self-Confidence; 12 = Self-Control.

complete linking and satisfactory with minimum linking. For the empirical data, JAGS was used for parameter estimation, because there were as many as 12 dimensions. The results indicated a good model–data fit and a slightly negative correlation of most of the 12 θ variables. Furthermore, the scaled scores for two persons to illustrate between- and within-individual comparisons in terms of psychological differentiation were provided.

The RIM opens up many theoretical research lines and practical issues. First, model generalizations are of great value. Ranking items are very common (e.g., the SOV), and they include pairwise comparison items as special cases. Recently, the RIM has been successfully generalized to accommodate multidimensional ranking items (Wang et al., 2016). Second, it is desirable to add covariates to the RIM to explain the item and person parameters directly, known as the explanatory IRT approach (De Boeck & Wilson, 2004). Third, computerized adaptive testing based on the RIM is much more challenging because of high dimensionality and huge item banks. For example, with 12 dimensions and 10 statements in each dimension, the item bank easily reaches 6,600 MPC items. It would be very challenging to select an MPC item to administer in a short time (e.g., less than 1 s) from such a huge bank and a high dimensionality when using Fisher’s item information criterion for item selection. Some quick and dirty item selection methods may be needed. Fourth, differential statement functioning (DSF) is another challenging topic. In the RIM, each statement has a single utility for all persons. In reality, a statement may have different utilities for different groups of persons. Take career interests as an example: “Like teaching and playing with young children” might have a higher utility for females, whereas “Like working with tools and machines” might have a higher utility for males. Further investigations are required to determine how to adapt current assessment methods of differential item functioning to the DSF assessment.

Acknowledgment

The authors thank two anonymous reviewers for their constructive comments on earlier drafts of the article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Grants Council of the Hong Kong SAR under GRF Project (845013).

Supplemental Material

The online supplement are available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621617703183>.

References

- Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). *ACER Conquest 4.0 [Computer program]*. Melbourne: Australian Council for Educational Research.
- Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. *Applied Psychological Measurement*, 19, 269-290.
- Ashman, A., & Telfer, R. (1983). Personality profiles of pilots. *Aviation, Space, and Environmental Medicine*, 54, 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, 266-281.
- Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods*, 9, 453-465.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block design I: The method of paired comparisons. *Biometrika*, 39, 324-345.
- Brown, A. (2016a). Item response models for forced-choice questionnaires: A common framework. *Psychometrika*, 81, 135-160.
- Brown, A. (2016b). Thurstonian scaling of compositional questionnaire data. *Multivariate Behavioral Research*, 51, 345-356.
- Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, 71, 460-502.
- Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review*, 51, 292-303.
- De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 3-41). New York, NY: Springer.
- Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6, 733-807.
- Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and Individual Differences*, 35, 1241-1254.
- Guenole, N., Brown, A., & Cooper, A. J. (2016). Forced choice assessment of work related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment*. Advance online publication. doi:10.1177/1073191116641181
- Harris, J. A., Vernon, P. A., & Jang, K. L. (2005). Testing the differentiation of personality by intelligence hypothesis. *Personality and Individual Differences*, 38, 277-286.
- Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice Hall.
- Hontangas, P. M., de la Torre, Ponsoda, J., V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, 39, 598-612.
- Hui, C., Pak, S., & Cheng, H. (2009). Validation studies on a measure of overall managerial readiness for the Chinese. *International Journal of Selection and Assessment*, 17, 127-141.
- Jackson, D. N. (1977). *Jackson vocational interest survey manual*. Port Huron, MI: Research Psychologists Press.

- Johnson, E. G. (1992). The design of the national assessment of educational progress. *Journal of Educational Measurement*, 29, 95-110.
- Kiefer, T., Robitzsch, A., & Wu, M. (2016). TAM: Test Analysis Modules (R package Version 1.995-0) [Computer program]. Retrieved from <https://cran.r-project.org/web/packages/TAM/index.html>
- Kopelman, R. E., Rovenpor, J. L., & Guan, M. W. (2003). The study of values: Construction of the fourth edition. *Journal of Vocational Behavior*, 62, 203-220.
- Luce, R. D. (1959). *Individual choice behavior*. New York, NY: John Wiley.
- 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, 169-182.
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, 45, 935-974.
- McCloy, R. A., Heggestad, E. D., & Reeve, C. L. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods*, 8, 222-248.
- Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, 77, 531-552.
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29, 133-161.
- Organisation for Economic Co-Operation and Development. (2014). *PISA 2012 Technical Report*. Author. Retrieved from <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>
- 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.
- Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, 24, 3-32.
- Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement*, 29, 184-203.
- Stark, S., Chernyshenko, O. S., Drasgow, F., & White, L. A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*, 15, 463-487.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 79, 281-299.
- 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). New York, NY: Springer.
- Witkin, H. A., Goodenough, D. R., & Oltman, P. K. (1979). Psychological differentiation: Current status. *Journal of Personality and Social Psychology*, 37, 1127-1145.