Forced-Choice Ranking Models for Raters’ Ranking Data

Su-Pin Hung
National Cheng Kung University

Hung-Yu Huang
University of Taipei

To address response style or bias in rating scales, forced-choice items are often used to request that respondents rank their attitudes or preferences among a limited set of options. The rating scales used by raters to render judgments on ratees’ performance also contribute to rater bias or errors; consequently, forced-choice items have recently been employed for raters to rate how a ratee performs in certain defined traits. This study develops forced-choice ranking models (FCRMs) for data analysis when performance is evaluated by external raters or experts in a forced-choice ranking format. The proposed FCRMs consider different degrees of raters’ leniency/severity when modeling the selection probability in the generalized unfolding item response theory framework. They include an additional topic facet when multiple tasks are evaluated and incorporate variations in leniency parameters to capture the interactions between ratees and raters. The simulation results indicate that the parameters of the new models can be satisfactorily recovered and that better parameter recovery is associated with more item blocks, larger sample sizes, and a complete ranking design. A technological creativity assessment is presented as an empirical example with which to demonstrate the applicability and implications of the new models.

Keywords: forced-choice items; unfolding models; item response theory; rater errors

Introduction

Rater-mediated assessments conventionally require raters to provide ratings by directly responding to a scoring rubric (e.g., a Likert-type rating scale) to evaluate the degrees of proficiency with respect to specified criteria. Performance assessments are produced by external raters in many fields of applied social science, and well-known examples are performance appraisals in organizational behavior fields, writing assessments in language testing situations,
patients’ recovery evaluations in medical contexts, and creativity assessments in educational settings. Unlike measured outcomes obtained from the selected-response item format used in educational assessments and the self-reported rating scale used in psychological testing, human rater scoring is subject to rater effects on rater-mediated assessments (Myford & Wolfe, 2003). Although psychometric models, such as item response theory (IRT) and latent trait models, have been developed in earlier studies to account for several commonly observed rater effects (e.g., rater leniency, rater inconsistency, consensual observer drift, and halo effects; Engelhard, 1994; Hung et al., 2012; Jin & Wang, 2018; Linacre, 1989; Myford & Wolfe, 2003, 2004; Wang et al., 2014; Wang & Wilson, 2005; Wilson & Hoskens, 2001), rater scoring is more vulnerable to various sources of bias and error due to the nature of scale-based rating assessments (Myford & Wolfe, 2003). Alternatively, a comparative judgment approach (i.e., forced-choice formats) to evaluating personal performance rather than direct scoring (i.e., single-stimulus formats) may provide a new direction to improve measurement properties in rater-mediated assessments (Laming, 2004; Pollitt, 2012).

Two types of comparative judgment approaches can be applied to rater-mediated assessment contexts based on Thurstone’s (1927) theorem of cognitive processes in comparative judgment. When objects are compared with respect to a prespecified attribute—for example, students’ assignments are evaluated with respect to creative thinking—a rank order of objects (e.g., students) can be created to indicate different proficiency levels in the measured attribute. This comparative approach, known as the between-object comparison design, is widely used in a variety of contexts, such as art appraisals (Newhouse, 2014), educational achievement measures (Crompvoets et al., 2020), essay evaluations (Steedle & Ferrara, 2016), and creativity ranking (Florida et al., 2015). Because the between-object comparison design requires multiple pairwise comparisons among a large number of objects, the high rater burden and time-consuming process may compromise the efficiency and validity of measurement. In addition, existing measurement models mostly require raters to perform comparisons with respect to a single attribute under a unidimensional testing structure (e.g., Crompvoets et al., 2020), which limits the applicability of this approach to comparisons of multiple attributes. Rather than producing comparisons among persons, an alternative approach to distinguishing between attributes requires raters to compare items representing different attributes and to measure distinct latent traits in an item block by ranking these items within each block. The between-trait comparison design uses forced-choice formats rather than single-stimulus formats, so that the response distortions that have been observed in Likert-type rating scales, such as those due to response styles, can be eliminated across items under comparison (Cheung & Chan, 2002). In this study, we focus on the between-trait comparison approach and develop a variety of measurement models for raters’ ranking data.
There is growing interest in the application of rank-ordered choice formats to rater-mediated assessments, and the corresponding measurement models are developed progressively. For example, 360-degree appraisals of employees in organizations are widely used for purposes of promotion, salary increases, and training to evaluate employee behavior and performance in job roles from distinct perspectives (e.g., those of superiors, peers, and subordinates) with respect to different competencies and can be thought of as a type of high-stakes assessment in organizations. Because the Likert-type rating rubrics used in 360-degree appraisals are more likely to be subject to rater biases, Brown et al. (2017) designed an analysis of the rater-mediated ranking task, in which four items measuring distinct latent traits in a block are partially ranked by external raters according to their judgments of ratees’ performance. Additionally, the Thurstonean IRT (TIRT) model (Brown & Maydeu-Olivares, 2011, 2012, 2013) is used to calibrate employees’ propensities in job performance. With the fit of the TIRT model to forced-choice ranking data, Brown et al. found that several potent rater errors, such as acquiescence, extreme tendency, and halo effects, could be effectively eliminated by direct item comparison. Although their results were significant and the analysis considered multiple latent traits simultaneously, several limitations should be noted.

First, for the TIRT model, the model parameters are estimated using confirmatory factor analysis based on traditional bivariate information rather than full information, and the two-step estimation procedure (i.e., estimating traits after structural parameters have been calibrated) results in imprecise parameter estimation because estimation errors are ignored. Second, for the purpose of model identification, the TIRT model requires inverse items to be designed as item blocks of opposite polarity. This design has been found to be less robust against response biases because respondents prefer direct items to inverse items (Morillo et al., 2016). Third, personnel assessments, such as 360-degree appraisals, often instruct raters to produce partial rankings (i.e., to select the items that best and worst describe the ratee) rather than complete rankings. When partial ranking data are collected, as in Brown et al.’s study, each block should then be converted into multiple binary dummy items by means of pairwise comparisons, and missing responses to binary dummy items during the partial ranking process should be statistically imputed to avoid bias in parameter estimation (Brown & Maydeu-Olivares, 2012). Although the TIRT model can be applied to complete ranking analysis without statistical imputation, the current application is limited to partial rankings and has been found to be less efficient than the full ranking method (Hontangas et al., 2015).

It is a common practice in econometric fields and marketing studies to ask respondents to provide a complete preference ordering among a set of option alternatives, such as different modes of transportation and different brands of a certain product (e.g., Ahn et al., 2006; Calfee et al., 2001). Therefore, several probabilistic choice models for rank-ordered choice data have been developed to
relate the desirability of a choice pattern for respondents with the combination of the person’s characteristics (e.g., demographic variables) and attributes of each choice option. Based on the ranking choice theorem (Luce, 2005), the probability of a ranking event can be decomposed into the product of multiple-choice probabilities in the rank-ordered choice set, in which the choice events for each option are assumed to be statistically independent and follow a multinomial logit model (Gensch & Recker, 1979). By computing the product of the probabilities represented by the multinomial logit model for each choice event, the rank-ordered logit model (also known as the exploded logit model) and several variants can be formulated to express the choice probability of rank ordering, as has been demonstrated in the choice behavior model literature (e.g., Beggs et al., 1981; Chapman & Staelin, 1982; Fok et al., 2012). Although the traditional probabilistic choice models for rank-ordered choice data are statistically efficient and have been widely applied to marketing surveys for understanding customers’ preference judgments, several limitations should be outlined and discussed for applications to psychological measures and rater-mediated assessment.

Because respondents’ choice behavior always involves a deterministic process, a stochastic utility model that includes deterministic and random components is used to represent the underlying process of making a probabilistic decision (Manski, 1977). The deterministic component of the utility function operationalizes the choice probabilities of the multinomial logit model and is determined by the respondent’s observed characteristics rather than latent traits or attributes. In psychometric fields, measuring respondents’ latent traits is of interest and importance for obtaining their performance levels on a prespecified latent continuum; therefore, the rank-ordered logit model is not an ideal approach to provide respondents’ specific latent trait measures. Second, the random utility framework assumes a dominance process in a preference ranking task; that is, the larger the utility value an option has, the higher the probability of that option being selected. It is not always the case, however, that personal preference and choice follow a cumulative response process because a person may agree with or prefer an option that has an ideal description for them (Andrich & Luo, 2019; Coombs, 1964). In contrast, an ideal-point stochastic model may be more appropriate than a dominance model to describe the underlying process of comparative judgment (see below for more details). Third, self-reported rankings have regularly been used in preference rank-ordered surveys, and the “true” preference judgments of respondents can be directly reflected by their actual ranking patterns. In some cases, however, individual performances may be ranked by external raters with respect to several criteria (e.g., employees’ performance is ranked by superiors with respect to four items represented by distinct latent traits; see Brown et al., 2017). The results of the sorting on these criteria do not necessarily reflect the real performance of the respondents because the raters’ characteristics (e.g., rater leniency; Engelhard, 1994) would influence the deterministic choice
Following the consecutive choice tradition and the explosion rule (Luce, 2005), the forced-choice ranking IRT model has been proposed. This model replaces the stochastic utility model and the multinomial logit model for modeling the choice probability of each ranking option with a specific IRT model (de la Torre et al., 2012; Joo et al., 2018; Lee et al., 2019). Similar to traditional probabilistic ranking models, self-reported rankings rather than rater-mediated rankings have been used in the forced-choice ranking IRT model, and rater effects have not been considered in previous studies. Our study builds on the contributions of the forced-choice ranking IRT model and extends it by quantifying raters’ leniency levels and introducing multiple measurement facets into the probabilistic choice function. Furthermore, we evaluate intrarater consistency by considering the possible randomness of raters’ leniency when their ranking data are collected. Note that we use “item” to describe the ranking option, and an item block is composed of a limited number of items measuring distinct latent traits to be compared and ranked. In typical raters’ ranking data, multiple item blocks are designed, and raters must rank the items according to their judgment of ratees’ performance for each item block. The newly developed forced-choice ranking models (FCRMs) have several advantages: Multiple latent traits can be measured and compared simultaneously, item response functions can be flexibly formulated by the dominance or ideal-point approach, and different measurement facets (e.g., respondents, items, raters, and materials) can be combined in relation to each item choice probability. This approach not only satisfies the practical demands of rater-mediated assessments but also contributes to theoretical progress in psychometric fields.

In recent decades, creativity has become an indispensable and essential skill for students’ future adaptability, and both schools and educational policymakers have sought approaches to improve and develop students’ creative potential in educational contexts by providing a supportive environment (e.g., Hernández-Torrano & Ibrayeva, 2020; Plucker et al., 2018). Abundant accumulated empirical evidence has shown the positive contribution of creativity to scholastic performance, problem-solving skills, and overall life success (Freund & Holling, 2008; Gajda et al., 2017; Sternberg, 2002). Among creativity assessments, the consensual assessment technique (Amabile, 1996) has frequently been used to evaluate a skill that is theorized to be relative to creativity. In such assessments, individuals are asked to create something, for example, by writing poems, telling stories, drawing pictures, or engaging in crafts, and experts are then asked to evaluate the products that the participants have created. The outcome measures in the consensual assessment technique are often obtained through human judgment on rating scales and have been found to be inconsistent with respect to raters’ rating procedures and use of rating scales due to differing rater cognition (for an intensive discussion, see Long & Pang, 2015). Corresponding to the purpose of
improving student creativity and innovation that is central to science, technology, engineering, arts, and mathematics education (Perignat & Katz-Buonincontro, 2019), we chose a technology creativity assessment based on the consensual assessment technique and forced-choice ranking formats as a demonstrative example for the application of the newly developed FCRMs. Additionally, it has been suggested that schools must provide opportunities for students to develop creativity capacity in correspondence with the increasing need for technology skills in future jobs (Keane & Keane, 2016). In this empirical analysis, our proposed raters’ ranking models are used to measure diverse aspects of creativity on the basis of the technology products that students create.

The remainder of this study is organized as follows. First, the Model Specification section describes the rationales and justifications for developing the new class of FCRMs in the context of rater-mediated assessments, accounting for multiple facets, such as rater leniency and the ranking of multiple tasks or materials, and controlling for the impacts of rater inconsistency. Then, three simulation studies are presented to assess model parameter recovery for each of the proposed models using Bayesian estimation. Next, an empirical example demonstrates the applications and implications of the models. Finally, we close this article by presenting an overall discussion of the results and suggestions for future research.

**Model Specification**

To prevent raters from endorsing (or disaffirming) all items due to the halo effect or response styles in Likert-type rating scales, it is justifiable to alternatively use forced-choice formats and ask raters to provide a complete rank ordering of a ratee’s performance (Brown et al., 2017). We begin this section with the rank-ordered probabilistic formulation, followed by issues associated with the choice of the item endorsement probabilistic function. At the end of this section, we provide the FCRM with raters’ leniency and further extend the new model toward a general formulation to correspond to theoretical and practical considerations.

**The Rank-Ordered Probabilistic Formulation**

When an individual is required to rank items within a block from most preferred to least preferred, the probability of a particular ranking pattern can be partitioned into a sequence consisting of the independent probability of picking the “most preferred” statement from the diminishing set of alternative statements (henceforth called the PICK probability) at each step in the decision process. Then, the overall probability of this ranking pattern is the product of these PICK probabilities (Beggs et al., 1981; Chapman & Staelin, 1982; de la Torre et al., 2012; Fok et al., 2012; Joo et al., 2018; Lee et al., 2019). When this concept is applied to the analysis of raters’ ranking data, the proposed modeling approach
Hung and Huang

can be illustrated for a hypothetical assessment scenario, in which individual students’ tasks (e.g., creative thinking, writing, or painting) are evaluated by raters using a forced-choice ranking format based on three criteria or dimensions (e.g., fluency, flexibility, and originality). Each dimension is measured by multiple items, from which a composite item (i.e., a forced-choice block) comprising three items that measure distinct dimensions can be formulated. The raters are required to rank such a composite item from most to least representative of the target’s performance. In the case of three items \((i_A, i_B, \text{ and } i_C)\) to be ranked in a forced-choice block, one possible ranking pattern among the \(3! = 6\) possible ranking order responses is that item \(i_A\) is the most representative, followed by \(i_B\), and \(i_C\) is the least representative. The probability of this ranking pattern \((i_A > i_B > i_C)\) in item block \(l\) for person \(n\) can be expressed as

\[
P_{nl,(i_A > i_B > i_C)}(\theta_{nl,A}, \theta_{nl,B}, \theta_{nl,C}) = P_{nl,(i_A > i_B, i_C)} \times P_{nl,(i_B > i_A, i_C)},
\]

where \(P_{nl,(i_A > i_B, i_C)}\) and \(P_{nl,(i_B > i_A, i_C)}\) are the PICK probabilities of selecting items \(i_A\) and \(i_B\), respectively, from among the remaining items in item block \(l\) for person \(n\) and \(\theta_{nl,A}, \theta_{nl,B}\), and \(\theta_{nl,C}\) denote the three domains measured by items \(i_A\), \(i_B\), and \(i_C\), respectively.

The PICK probabilities \(P_{nl,(i_A > i_B, i_C)}\) and \(P_{nl,(i_B > i_A, i_C)}\) can be further expressed as

\[
P_{nl,(i_A > i_B, i_C)} = \frac{P_{nl,i_A}(1)P_{nl,i_B}(0)P_{nl,i_C}(0)}{P_{nl,i_A}(1)P_{nl,i_B}(0)P_{nl,i_C}(0) + P_{nl,i_A}(0)P_{nl,i_B}(1)P_{nl,i_C}(0) + P_{nl,i_A}(0)P_{nl,i_B}(0)P_{nl,i_C}(1)},
\]

and

\[
P_{nl,(i_B > i_A, i_C)} = \frac{P_{nl,i_B}(1)P_{nl,i_C}(0)}{P_{nl,i_A}(1)P_{nl,i_C}(0) + P_{nl,i_B}(0)P_{nl,i_C}(1)},
\]

respectively, where \(P_{nl,i_A}(1), P_{nl,i_B}(1), \text{ and } P_{nl,i_C}(1)\) are the probabilities of selecting items \(i_A, i_B, \text{ and } i_C\), respectively, in item block \(l\) for person \(n\), and correspondingly, \(P_{nl,i_A}(0) = 1 - P_{nl,i_A}(1), P_{nl,i_B}(0) = 1 - P_{nl,i_B}(1), \text{ and } P_{nl,i_C}(0) = 1 - P_{nl,i_C}(1)\). The calculation of the joint probabilities of the other possible ranking patterns \((i_A > i_C > i_B, i_B > i_A > i_C, i_B > i_C > i_A, i_C > i_A > i_B, \text{ and } i_C > i_B > i_A)\) follows the same logic.

Issues Associated With the Choice of Item Endorsement Probabilistic Models

In traditional probabilistic ranking models, the probability of selecting an item is determined by a multinomial logit model under the random utility framework (e.g., Fok et al., 2012). In this study, we instead formulize the item endorsement probability using a binary IRT model because of the advantages mentioned above. For IRT modeling, the use of dominance (as in the two-parameter logistic model or 2PLM; Birnbaum, 1968) versus unfolding (as in the generalized graded unfolding model or GGUM; Roberts et al., 2000) to model the underlying process
of forced-choice item evaluation remains controversial (Andrich & Luo, 2019; Drasgow et al., 2010). Within the framework of factor analysis, the TIRT model assumes a dominance process, whereas the IRT framework uses the almost equivalent multiunidimensional pairwise preference model developed by Morillo et al. (2016). For several reasons, we adopt an unfolding (or ideal-point) model for forced-choice responses. In pairwise comparisons, impersonal judgment implies that the rater’s judgment is objective. This eliminates personal parameters, meaning that probability is determined only by the stimuli (i.e., items) in the cumulative (dominance) IRT model (see Andrich & Luo, 2019, pp. 183–185). In practice, however, human raters’ evaluations do involve some subjective elements and rarely provide completely objective information about target behavior (Brown et al., 2017; Van der Heijden & Nijhof, 2004). For this reason, a single-peaked discrimination response function (i.e., unfolding) is more appropriate for assessing comparative judgments based on rater preference and choice (Andrich & Luo, 2019, pp. 186–188).

Formulation of the FCRM With Raters’ Leniency

Several probability functions have been proposed for modeling the selection process represented in Equations 2 and 3; for the purposes of the present study, unfolding IRT models are adopted to capture the nature of the raters’ subjective judgments. When applied to the probability of preferring a given item, the GGUM-rank model can be used to analyze multidimensional forced-choice items and ranking data (Joo et al., 2018; Lee et al., 2019; Roberts et al., 2000). However, because the GGUM-rank model was not developed for rater data and does not account for rater leniency in the selection of each item for comparison, a new forced-choice ranking IRT model is needed. As in many-faceted IRT models, a leniency parameter for each rater can be included in the unfolding probability function to adjust the range within which a rater’s evaluation of an item is more likely to be positive (i.e., representative of the target’s performance) than negative (i.e., unrepresentative of the target’s performance). Because the GGUM threshold parameter cannot be interpreted as the intersection of two adjacent categories (Roberts et al., 2000), it is difficult to adjust the range corresponding to a positive response on the continuum due to rater leniency. As an alternative, the threshold parameter in the hyperbolic cosine model (HCM) represents the latitude of acceptance of a stimulus and can be used to describe the intersection between the probabilities of positive and negative responses (Andrich, 1995; Andrich & Liu, 1993). Thus, a generalized HCM can be formulated by adding a slope (discrimination) parameter to each item in the HCM, and the probability function can be expressed as

\[
P_{nl,i,i,a,b,c}(1) \equiv P(Y_{nl,i,a} = 1 | \theta_{nd}, \delta_{id}, \alpha_{ia}, \rho_{ia}) = \frac{\psi(\alpha_{ia} \rho_{ia})}{\psi(\alpha_{ia}(\theta_{nd} - \delta_{id})) + \psi(\alpha_{ia} \rho_{ia})}, \tag{4}
\]
where $\theta_{nd}$ represents the substantive latent trait $d$ for person $n$ (in this case, for three latent traits, $d = A$, $B$, and $C$), and the vectors $\Theta_n = (\theta_{nd}, \theta_{nB}, \theta_{nC})'$ are assumed to be normally distributed, with a zero mean vector and a variance–covariance matrix $\Sigma$; $\delta_{id}$ and $\alpha_{id}$ are the location and discrimination parameters, respectively, for item $i_d$ in item block $l$; $p_{id}$ is the latitude of acceptance (threshold) parameter for item $i_d$ in item block $l$; and $\psi(\cdot)$ denotes an operational function that produces a symmetrical probability function in the unfolding model. Luo (2001, pp. 242–244) demonstrated that the graded unfolding model (Roberts et al., 2000) can be viewed as a special case of the HCM in a general form for polytomously scored responses when a specific operational function is chosen. In our study, because the dichotomously scored item format is used, Luo’s operational function can be simplified as follows:

$$
\psi(t) = \cosh\left(\frac{2C + 1}{2} t\right) / \cosh\left(\left[\left(\frac{2C + 1}{2} - 1\right) t\right]\right),
$$

where $C$ is the number of response categories minus one, and thus $C = 1$, and $\cosh$ represents the hyperbolic cosine function, which is defined as $\cosh(x) = \frac{\exp(x) + \exp(-x)}{2}$. Under this operational function formulation, in addition, the generalized HCM threshold parameter has a nonlinear relationship with the GGUM threshold parameter (see also Wang et al., 2013, pp. 181–182).

The generalized HCM described in Equation 4 is appropriate for application to self-reported ranking data. When preference responses are produced by external human raters, a leniency parameter $\tau_k$ (in an exponential form) can be added to Equation 4 to account for the raters’ differing leniency levels; the probability that rater $k$ will endorse ratee $n$ on item $i_d$ of item block $l$ can then be formulated as

$$
P_{nlk,i_d,A,B,C}(1) \equiv P(Y_{nlk,i_d} = 1|\theta_{nd}, \delta_{id}, \alpha_{id}, p_{id}, \tau_k) = \frac{\psi[\alpha_{id} \rho_{ik} \exp(\tau_k)]}{\psi[\alpha_{id} (\theta_{nd} - \delta_{id})] + \psi[\alpha_{id} \rho_{ik} \exp(\tau_k)]}.
$$

When the subscript $k$ is added to Equations 2 and 3 (i.e., $P_{nlk,(i_d,i_e,i_b,i_c)}$ and $P_{nlk,(i_d|i_b,i_c)}$), distinct PICK probabilities can be derived for all possible ranking patterns. Because $\rho_{ik}$ is a unit parameter (i.e., $\rho_{ik} \geq 0$) indicating the distance between two adjacent unfolded thresholds and describing the range within which the performance of a ratee is more likely to be judged positively (i.e., representative of the target’s performance) than negatively (i.e., unrepresentative of the target’s performance) with respect to an item, the latitude of positive judgment for rater $k$ can be adjusted by a nonnegative value to compute $\exp(\tau_k)$. A similar approach is adopted for attitude and personality measurements to adjust the latitude of acceptance in unfolding models when Likert-based ratings involve.
subjective judgment (Wang et al., 2013). The larger the $\tau$ parameter is, the more lenient the rater and the higher the probability of a positive response to the target’s performance.

Take the following scenario as an example. When three raters with $\tau = -1, 0, 1$ evaluate ratees’ performance with respect to an item with $\delta = 0, \alpha = 1, \rho = 1$, ratees with trait levels ranging between $-0.37$ and $0.37$ are more likely to be rated positively than negatively by the severe rater ($\tau = -1$), ratees with trait levels ranging between $-1$ and $1$ are more likely to be rated positively than negatively by the objective rater ($\tau = 0$), and ratees with trait levels ranging between $-2.72$ and $2.72$ are more likely to be rated positively than negatively by the lenient rater ($\tau = 1$). Figure 1 shows the probability of a positive response versus a negative response on the given item for the three respective raters along the ratees’ trait continuum.

When the generalized HCM (i.e., with discrimination parameters) is used to compute the item selection probability in the rater-medicated ranking model, as shown in Equation 6, we classify it as the two-parameter FCRM with raters’ leniency (2P-FCRM-L), and when the HCM (i.e., without discrimination parameters) is used, we classify it as the one-parameter FCRM with raters’ leniency (1P-FCRM-L). Corresponding to the self-reported ranking contexts, the 2P-FCRM-L and 1P-FCRM-L are reduced to the 2P-FCRM and 1P-FCRM, respectively, without raters’ leniency. To simplify the model and reduce the computational burden, the threshold parameter $\rho_{id}$ can be constrained to be identical across items corresponding to the same dimension; that is, $\rho_{id} = \rho_d$. For the purposes of model identification, the diagonal elements of the variance–covariance matrix $\Sigma_0$ are all set to one for both the 2P-FCRM-L and 2P-FCRM, and the mean leniency parameter is constrained to zero for both the 2P-FCRM-L and 1P-FCRM-L.

Notably, the proposed model relates to a single task assessment conducted by raters. In some cases, more than one task may be accomplished by individuals, and these tasks are evaluated by experts based on the same rubric. For example, college students in Hong Kong taking an English test wrote two essays (i.e., two tasks) that were evaluated by external raters along three criteria: organization, vocabulary, and grammar (Jin & Wang, 2018). When performance on multiple tasks is ranked by external raters with respect to various criteria, as in many-faceted IRT models of dominance responses, a topic difficulty parameter $\eta_j$ can be incorporated to indicate the effect of task $j$ on the probability function given in Equation 6, as follows:

$$P_{nlkj;id;i,b,c}(1) \equiv P(Y_{nlkj;id} = 1|\theta_{nd}, \delta_{id}, \alpha_{id}, \rho_{id}, \tau_k, \eta_j)$$

$$= \frac{\psi[x_{id}\rho_{ij}\exp(\tau_k)]}{\psi[x_{id}(\theta_d - \delta_{id} - \eta_j)] + \psi[x_{id}\rho_{ij}\exp(\tau_k)]},$$

(7)
FIGURE 1. Probability function for three raters’ leniency parameters of $-1$, $0$, and $1$ on an illustrative item.
where the other parameters and the operational function are as previously defined, and the subscript $j$ is added to the PICK probabilities of all ranking patterns to represent the effects of different tasks. For example, the probability of the ranking pattern described in Equation 1 becomes

$$P_{nlkj,(i_d > i_g > i_c)}(0_{nd}, 0_{ng}, 0_{nc}) = P_{nlkj,(i_d,i_g,i_c)} \times P_{nlkj,(i_g,i_b,i_c)}.$$  (8)

As in many-faceted IRT models, the mean topic difficulty parameter is constrained to zero for model identification.

Even with rater training, scoring behavior is often influenced by a rater’s own characteristics and unique prior experiences (Wang & Engelhard, 2019, p. 775). Variations in raters’ leniency across ratees and local item dependence caused by interactions between raters and ratees may interfere with parameter estimation. Specifically, while interrater reliability can be captured by the estimation of different $\tau$ parameters, intrarater consistency can be monitored by considering variations in leniency. Similar to the random-effects facet model proposed to model the intrarater interaction effect between persons and items for dominance items (Wang & Wilson, 2005), the generalized HCM for raters’ rankings of multidimensional forced-choice items can be expressed in terms of the following probability function:

$$P_{nlkj,\delta_{id},\beta_{id},\gamma_{id}}(\tau_{nk}, \eta_j) = \frac{\psi[\delta_{id} + \rho_{id} \exp(\tau_{nk})]}{\psi[\delta_{id} - \delta_{id} - \eta_j] + \psi[\delta_{id} + \rho_{id} \exp(\tau_{nk})]},$$  (9)

where $\tau_{nk}$ is the leniency parameter of rater $k$ for ratee $n$ and is assumed to be normally distributed, with a mean of $\tau_k$ and a variance of $\sigma^2_{\tau_k}$. The magnitude of $\sigma^2_{\tau_k}$ represents the dispersion of rater $k$’s leniency across ratees, and the greater the variation in rater $k$’s leniency is, the lower the intrarater reliability.

Adopting Equation 9 as the item endorsement probability in the FCRM leads to the most general rater-mediated ranking model, which is classified as the two-parameter many-faceted FCRM with random leniency (2P-MF-FCRM-RL) or the one-parameter many-faceted FCRM with random leniency (1P-MF-FCRM-RL) when the discrimination parameters are all set to one in Equation 9. Note that when $\sigma^2_{\tau_k} = 0$, rater $k$ is fully consistent across ratees, simplifying Equation 9 to Equation 7, and the corresponding 2P-MF-FCRM-L and 1P-MF-FCRM-L arise. When both $\sigma^2_{\tau_k} = 0$ and only one task are evaluated, Equation 9 will be further reduced to Equation 6, and the corresponding ranking models will become the same as the 2P-FCRM-L and 1P-FCRM-L described above.

Simulation Studies

Three simulation studies were conducted to investigate parameter recovery for a range of FCRMs considering raters’ leniency. In each case, three
dimensions, or criteria, were used to generate items for comparison in each item block, and five raters were required to rank the three items measuring different dimensions in terms of the ratee’s performance on a given task. The simulation design, prior distribution settings of Bayesian estimation, and results for the three studies are described separately as follows.

Simulation 1

Method. In the first simulation, each ranking pattern corresponded to the joint probability of multiple PICK functions and was generated by following the 2P-FCRM-L, where the PICK probability could be computed using the generalized HCM of rater leniency for each item (Equation 6). Three independent variables were manipulated: (a) the number of ratees (500 or 1,000), (b) the number of raters assigning rankings to a ratee (two or five), and (c) the number of item blocks (five or 10). In other words, in these ranking designs, each of 500 or 1,000 ratees was evaluated by two or five raters (out of five raters in total) using five or 10 item blocks, and each item block, which comprised three items of different dimensions, was ranked by the raters. The ranking designs in which only two raters assigned rankings to a ratee were considered incomplete; in these designs, each rater should evaluate 400 ratees from the large sample or 200 ratees from the small sample of 500. Specifically, as shown in Table 1, the ratees were split into five groups of equal size. Adjacent raters evaluated the same ratee group, and each group was evaluated by two raters. This concept is similar to the balanced incomplete blocking design for linkages between different tests and persons (Lord, 1965) and is consistent with the incomplete rating design used in previous studies (e.g., Jin & Wang, 2018).

For the simulations, the \( \delta \) parameters were generated from a uniform distribution of \(-2\) to \(2\) for each dimension. The \( \alpha \) parameters were generated from a uniform distribution of \(0.5\)–\(1.5\) for each dimension. The \( \rho \) parameters were set to \(1.10\) for all items corresponding to the first dimension, \(0.79\) for all items...
corresponding to the second dimension, and 0.59 for all items corresponding to the third dimension. The \( \tau \) parameters were generated from a uniform distribution of \(-0.5\) to \(0.5\); thus, \( \exp(\tau) \) ranged from 0.61 to 1.65, representing reasonable values for controlling rater leniency when making judgments. The \( \theta \) vectors were sampled from \( N_3(0, \Sigma_0) \), where all variances in \( \Sigma_0 \) were set to 1 and all covariances were set to 0.5. The generated values for the response data and simulation designs were consistent with the findings of previous studies involving unfolding IRT models, multidimensional forced-choice IRT models, and many-faceted IRT models (Jin & Wang, 2018; Joo et al., 2018, Lee et al., 2019; Wang et al., 2013).

Parameter Estimation and Analysis. The WinBUGS program (Spiegelhalter et al., 2003) was used to calibrate the model parameters based on the Bayesian estimation and Markov Chain Monte Carlo (MCMC) methods. Before a joint posterior distribution could be produced using the MCMC method, a statistical model and a set of prior distributions for the model parameters were required to specify the full conditional parameter distributions through sequential sampling. In line with previous studies using Bayesian estimation in IRT models (e.g., Jin & Wang, 2018; Lee et al., 2019; Liu & Wang, 2016; Morillo et al., 2016; Wang et al., 2013), the prior distributions of the model parameters were specified as follows. In the first simulation study, a lognormal distribution with a mean of 0 and a variance of 1 was used for the \( \alpha \) and \( \rho \) parameters, a truncated normal distribution with a mean of 0 and a variance of 4 was used for the positive and negative \( \delta \) parameters, a normal distribution with a mean of 0 and a variance of 1 was used for the \( \tau \) parameters, and a multivariate normal distribution with a zero mean vector and a nonzero variance–covariance matrix was used for the \( \theta \) parameters, where the diagonals of the variance–covariance matrix were constrained to 1 and the off-diagonals (i.e., the correlations between latent traits) followed a uniform distribution between \(-1\) and 1.

Note that the uniform priors used for the correlation coefficients in the variance–covariance matrix were not restricted to specific bounds (e.g., positive values) because there may be a variety of correlations in practical situations (Jin & Wang, 2015). However, the sampling process in the MCMC method may yield numerical overflow due to the appearance of a nonpositive definite or singular matrix. According to our experience, setting appropriate initial values can improve the efficiency of parameter estimation and reduce the computation time needed to reach convergence. In real assessment situations, information on starting values can be obtained from experts or previous empirical analyses (e.g., Liu & Wang, 2016). In addition, if optimal initial values are not available, an alternative is to generate multiple sets of initial values to examine the parameter convergence over iterations within the Bayesian framework. The Online Appendix lists the WinBUGS codes for the 2P-FCRM-L, and the initial values were internally generated by the WinBUGS program.
Considering the large number of simulated conditions and the fact that each calibration required dozens of hours of computer time, we conducted 30 replications for each condition of the three simulation studies. Three parallel chains were implemented for five randomly selected simulated datasets under each condition to evaluate parameter convergence and determine the required number of iterations. Regarding the multivariate potential scale reduction factors (Brooks & Gelman, 1998), which were all less than 1.1, the results indicated that 15,000 iterations were sufficient to achieve satisfactory parameter convergence, with the first 5,000 iterations designated as the burn-in period.

Results. Because numerous parameters were estimated, in consideration of space constraints, the parameter recovery performance was examined by computing the means and standard deviations of the bias and root mean square error (RMSE) across items for the item parameters, across raters for the rater’s leniency parameters, and across dimensions for the covariance parameters of the latent traits. Tables 2 and 3 summarize the results of the proposed model tested in the first simulation study when using incomplete and complete ranking designs, respectively. The bias values were quite small in most cases except for some estimators in the five-item-block condition with the incomplete ranking design. When the complete ranking design was used, the bias values were closer to zero across all conditions. With respect to the RMSE, whether the incomplete or complete ranking design was used, the findings indicated that the RMSE magnitudes generally decreased when larger numbers of item blocks and ratees were used. As expected, the parameters could be recovered more satisfactorily when each ratee was evaluated by five raters (i.e., the complete ranking design) than when each ratee was evaluated by two raters (i.e., the incomplete ranking design), as shown in Tables 2 and 3.

Regarding the recovery of the person parameters, Table 4 shows the mean RMSEs of the person parameter estimates for the three dimensions across all replications. The results show that the use of both a larger number of item blocks and a complete ranking design increased the precision of person parameter estimation, as indicated by the substantial decrease in the mean RMSE values in both cases. In contrast, the sample size showed no systematic effect on person parameter recovery.

Simulation 2

Method. The second simulation study considered a practical testing situation involving multiple tasks (e.g., essays), requiring the incorporation of topic effects into the probability function. Thus, ranking patterns for the respondents were generated according to the 2P-MF-FCRM-L. In the considered scenario, three tasks created by each of the 500 ratees were each evaluated by five raters, who ranked multidimensional forced-choice items along three dimensions in...
five item blocks. For the simulations, the topic parameters \( \eta \) were set to \(-0.5, 0, \) and \(0.5\) for the first, second, and third tasks, respectively. For the other parameters, the generated values were the same as in the first simulation study. In order to evaluate the performance of the model, the simulations were conducted with different sample sizes and item blocks. The results were then compared to the true values to assess the accuracy of the parameter estimates. The table below summarizes the parameter recovery results for the two-parameter forced-choice ranking model with raters’ leniency in an incomplete ranking design.

| Parameter | Dimension 1 | Dimension 2 | Dimension 3 |
|-----------|-------------|-------------|-------------|
| \( \alpha \) Mean | \(-0.038\) | \(-0.039\) | \(0.013\) |
| \( \alpha \) SD | \(0.046\) | \(0.058\) | \(0.027\) |
| \( \beta \) Mean | \(-0.129\) | \(0.023\) | \(0.014\) |
| \( \beta \) SD | \(0.163\) | \(0.021\) | \(0.109\) |
| \( \gamma \) Mean | \(-0.052\) | \(0.042\) | \(0.000\) |
| \( \gamma \) SD | \(0.175\) | \(0.023\) | \(0.000\) |
| Covariance Mean | \(-0.17\) | \(-0.047\) | \(-0.017\) |
| Covariance SD | \(0.012\) | \(0.004\) | \(0.012\) |

Note: RMSE = root mean square error; \( \alpha \) = item slope parameter; \( \beta \) = item location parameter; \( \gamma \) = item threshold parameter; \( \tau \) = rater’s leniency parameter.
addition, the prior distribution for the η parameters was assumed to follow a normal prior distribution with a mean of 0 and a variance of 4. Prior distributions for other model parameters were the same as in the first simulation study,
Results. In the second simulation study, data were simulated in accordance with the performance of 500 ratees on three tasks each, evaluated by five raters in a complete ranking design, and the parameter recovery was assessed in terms of the bias and RMSE. As shown on the left-hand side of Table 5, the parameters were recovered satisfactorily, all bias values were close to zero, and the RMSE values were rather small. The additional topic parameters were estimated precisely, and it appears that the 2P-MF-FCRM-L produced better parameter estimates when multiple tasks were evaluated than when a single task was appraised under otherwise identical conditions (i.e., 500 ratees, five-item blocks, and the complete ranking design; see Tables 3 and 5). These findings are not surprising because when the raters ranked the item blocks multiple times for different tasks, a greater amount of information was obtained, facilitating accurate parameter estimation. Parameter recovery patterns similar to those for the item parameters were observed for the person parameters, as evidenced by the lower RMSE values of .229, .220, and .250 for the three latent trait estimates.

TABLE 4. Mean RMSEs of Person Parameter Estimates for the Two-Parameter Forced-Choice Ranking Model With Raters’ Leniency

| Sample Size | Item Blocks | Rater Ranking Design | Dimension |
|-------------|-------------|----------------------|-----------|
|             | 5           | 10                   | 5         | 10         |
| Incomplete  | First       | .538                 | .522      | .421       | .422       |
|             | Second      | .556                 | .562      | .393       | .388       |
|             | Third       | .598                 | .585      | .446       | .453       |
| Complete    | First       | .388                 | .375      | .279       | .275       |
|             | Second      | .386                 | .408      | .254       | .252       |
|             | Third       | .433                 | .429      | .299       | .296       |

Note. RMSE = root mean square error.

and the same criteria were used to assess the parameter recovery for the 2P-MF-FCRM-L.

Results. In the second simulation study, data were simulated in accordance with the performance of 500 ratees on three tasks each, evaluated by five raters in a complete ranking design, and the parameter recovery was assessed in terms of the bias and RMSE. As shown on the left-hand side of Table 5, the parameters were recovered satisfactorily, all bias values were close to zero, and the RMSE values were rather small. The additional topic parameters were estimated precisely, and it appears that the 2P-MF-FCRM-L produced better parameter estimates when multiple tasks were evaluated than when a single task was appraised under otherwise identical conditions (i.e., 500 ratees, five-item blocks, and the complete ranking design; see Tables 3 and 5). These findings are not surprising because when the raters ranked the item blocks multiple times for different tasks, a greater amount of information was obtained, facilitating accurate parameter estimation. Parameter recovery patterns similar to those for the item parameters were observed for the person parameters, as evidenced by the lower RMSE values of .229, .220, and .250 for the three latent trait estimates.

Simulation 3

Method. Because the third simulation study considered variations in rater leniency across ratees, the leniency parameters were treated as random-effect parameters rather than fixed-effect parameters. Five raters were considered, with the variances of $\tau_{n1}$, $\tau_{n2}$, $\tau_{n3}$, $\tau_{n4}$, and $\tau_{n5}$ set to 0.5, 0.5, 1, 1.5, and 1.5,
TABLE 5.
Summary of Parameter Recovery Results for the Second and Third Simulation Studies

| Simulation Study | Fitting Model | Data-Generating | Data-Generating | Ignoring Rater Effects |
|------------------|---------------|----------------|----------------|------------------------|
|                  | Criterion     | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| **Parameter**    | **α**         |      |      |      |      |      |      |
| Dimension 1      | Mean          | -.013 | .036 | -.025 | .084 | -.056 | .412 |
| SD               | .005         | .010 | .050 | .020 | .184 | .169 |
| **Dimension 2**  | Mean          | -.011 | .036 | -.037 | .090 | -.092 | .173 |
| SD               | .009         | .010 | .058 | .027 | .074 | .095 |
| **Dimension 3**  | Mean          | -.011 | .043 | -.041 | .096 | .024  | .283 |
| SD               | .006         | .009 | .060 | .033 | .065 | .194 |
| **δ**            | Dimension 1  | Mean | .007 | .074 | -.102 | .172 | .062  | .278 |
| SD               | .017         | .004 | .114 | .103 | .058 | .171 |
| **Dimension 2**  | Mean          | .015 | .065 | .034 | .112 | .058  | .128 |
| SD               | .009         | .012 | .044 | .033 | .042 | .032 |
| **Dimension 3**  | Mean          | .023 | .078 | .041 | .175 | .068  | .210 |
| SD               | .023         | .037 | .099 | .070 | .066 | .038 |
| **ρ**            | Mean          | -.010 | .055 | -.070 | .109 | .128  | .305 |
| SD               | .004         | .006 | .056 | .034 | .117 | .088 |
| **τ**            | Mean          | .000 | .035 | .000 | .082 | —     | —    |
| SD               | .018         | .009 | .064 | .021 | —     | —    |
| **Covariance**   | Mean          | -.010 | .035 | -.032 | .059 | .027  | .120 |
| SD               | .015         | .010 | .004 | .005 | .013 | .018 |
| **η**            | Mean          | .000 | .011 | —     | —     | —     | —    |
| SD               | .004         | .004 | —     | —     | —     | —    |
| **σ^2_t**        | Mean          | —    | —    | -.184 | .241 | —     | —    |
| SD               | —            | —    | .095 | .094 | —     | —    |

*Note.* For both simulation studies, 500 ratees were evaluated by five raters on five item blocks in a complete ranking design. RMSE = root mean square error; α = item slope parameter; δ = item location parameter; ρ = item threshold parameter; τ = rater’s leniency parameter; η = topic parameter; σ^2_t = variance of the rater’s leniency parameter; — = not applicable because of different model assumptions.
respectively, representing $\sigma^2_{vi}$ magnitudes from small to large. Specifically, $\tau_{vk}$ $(k = 1, \cdots, 5)$ parameters were generated from a normal distribution with a mean of $\tau_k$ and variance of $\sigma^2_{vi}$, where the values of $\tau_k$ were the same as in the first simulation study. The simulation conditions and generated values were the same as in the second simulation study except that the raters evaluated only a single task from each ratee. All simulated data were generated from the 2P-FCRM-RL and fit using the data-generating model to assess the effectiveness of model parameter estimation. In addition, the data generated in the third simulation study were fit using the corresponding FCRM without rater leniency (i.e., 2P-FCRM) to evaluate the consequences of ignoring rater effects on model parameter estimation. The priors for the means of the random-effect $\tau$ parameters were the same as in the first simulation study, and the priors for the inverse variances of the random-effect $\tau$ parameters were assumed to have a gamma distribution, with both hyperparameters being equal to .01. Priors for other parameters were the same as in the previous simulation studies.

Results. The right-hand side of Table 5 summarizes the parameter recovery results for the case in which the data were simulated in accordance with the 2P-FCRM-RL and were fit using both the data-generating model and the 2P-FCRM (i.e., the model ignoring rater effects) to assess the quality of parameter estimation in the more complicated extended model and to investigate the consequences of ignoring rater effects in the forced-choice ranking items for parameter estimation. As indicated by the larger bias and RMSE values obtained when fitting the data with the 2P-FCRM, model misspecification had the nontrivial effects of increasing bias and RMSE in parameter estimation. The same conclusion can be drawn for person parameter recovery, as the RMSE values for the three latent trait estimates were .420, .406, and .453 for the fit with the 2P-FCRM-RL and .539, .508, and .559 for the fit with the 2P-FCRM. When we further compare the results reported in Tables 3 and 5 that were obtained under the same simulation conditions, we find that adding additional random effects to the leniency parameters caused the quality of model parameter estimation to deteriorate, although the parameter recovery for the more complex model was still acceptable. In our modeling formulation, the raters’ leniency random effects were treated as random noise and assumed to be uncorrelated with each other and with the substantive latent trait. Because the orthogonal structure was imposed in the random-effect variable space, the estimation benefits obtained from multiple correlated latent variables were not accessible, and more uncorrelated random-effect variables would hinder the accuracy of parameter estimates (Wang & Wilson, 2005). As indicated by the consequences of ignoring rater effects on parameter estimation, different raters’ leniency levels and their variations among ratees should be considered when
raters are evaluating ratees’ performance with respect to several criteria using forced-choice ranking items.

Empirical Demonstration

A real data analysis is presented here to illustrate the application of the proposed models based on different underlying process assumptions. In Taiwan, creativity is currently stressed as an important form of literacy in schooling environments. To assess the technological creativity of junior high school students, in accordance with Amabile’s theory of creativity assessment (Amabile, 1996), an open-ended task was formulated and evaluated by experts. Specifically, 384 students were recruited to design a cell phone that met junior high school students’ needs and showed personal creativity or elegance. The students were presented with several popular cell phones and provided with examples that were viewed as demonstrating high and low creativity based on three specified dimensions (described below). After receiving the instructions, the students were asked to draw a cell phone schematic and provide a verbal description of the phone’s functions. Three experienced raters independently evaluated these products (i.e., cell phone designs) on a website by ranking the students’ performance based on three dimensions: presentation, usefulness, and originality. Each dimension was evaluated based on multiple indicators, and the raters were asked to rank three indicators measuring different dimensions in each item block with respect to the products that the students had designed. The presentation dimension was assessed by two indicators corresponding to the verbal and visual presentations; usefulness was evaluated in terms of cost, function, and practicability; and originality was assessed in terms of appeal, design, labeling, and coordination. Each item block was formed by combining three indicators representing the different dimensions, and a total of 24 item blocks were presented to be ranked by the raters in evaluating the students’ technological creativity.

A variety of forced-choice ranking IRT models based on different model assumptions were fit to the responses evaluated by the raters. Three major questions were addressed in this empirical analysis: (a) Did the raters exhibit varying degrees of leniency? (b) Did the items have different discrimination parameters? and (c) If the raters exhibited varying degrees of leniency, should the variations of the leniency parameters across students be incorporated into the fitting model? Accordingly, six models were fit to the data as described below. When all the discrimination parameters were set to one, the three models of the 1P-FCRM (i.e., without raters’ leniency), 1P-FCRM-L (i.e., with raters’ leniency), and 1P-FCRM-RL (i.e., with randomness of raters’ leniency) were fit and compared. When all the discrimination parameters were freely estimated, the corresponding analysis models were the 2P-FCRM, 2P-FCRM-L, and 2P-FCRM-RL. The Bayesian deviance information criterion (DIC;
Spiegelhalter et al., 2002) was used to determine which of the six models could provide the best model-data fit; the smaller the DIC value was, the better the model-data fit was considered.

Because a small sample size was used in this analysis, a common discrimination parameter across all items corresponding to the same dimension was adopted to reduce the computational burden, and the results should be interpreted with caution. Nevertheless, a comparison among the various fitting models with respect to person parameter estimation sheds light on how substantial the practical effects of the different approaches are in working with raters’ ranking data. The DIC values were 9,115.45, 10,683.10, 9,402.62, 9,753.43, 8,892.23, and 6,267.09 for the 1P-FCRM, 2P-FCRM, 1P-FCRM-L, 2P-FCRM-L, 1P-FCRM-RL, and 2P-FCRM-RL, respectively; therefore, the 2P-FCRM-RL was selected as the best fitting model because it had the smallest DIC value. To obtain evidence of the model-data fit of the 2P-FCRM-RL in an absolute sense, we applied the posterior predictive model checking method to the data within a Bayesian framework (Gelman et al., 1996) by assessing the plausibility of the replicated data against the observed data during numerous iterations based on the ranking pattern of each item block. The statistic of the Bayesian \( \chi^2 \) test (Jin & Wang, 2018; Sinharay et al., 2006) was used to evaluate the overall model-data fit, and the results indicated that the 2P-FCRM-RL provided a good fit to the data and was the final model of choice.

Furthermore, we chose the 2P-FCRM and 2P-FCRM-L for comparison with the 2P-FCRM-RL in terms of person parameter estimation because these three models estimate the discrimination parameters. Comparing the 2P-FCRM and 2P-FCRM-L to the best fitting 2P-FCRM-RL shows the consequences of estimating the person parameters using the two parsimonious models. Figure 2 shows a set of scatterplots of the \( \theta \) estimates from the 2P-FCRM and the 2P-FCRM-L versus the \( \theta \) estimates from the 2P-FCRM-RL for each of the three measured latent traits. The scatterplots show that the \( \theta \) estimates from each of the three fitting models are fundamentally different and that there is a nonlinear relationship between the two poorer fitting models (i.e., the 2P-FCRM and 2P-FCRM-L) and the best fitting model (i.e., the 2P-FCRM-RL). Furthermore, the 2P-FCRM \( \theta \) estimates appear to show greater discrepancies than the 2P-FCRM-RL \( \theta \) estimates and the 2P-FCRM-L \( \theta \) estimates, suggesting that neglecting both raters’ leniency levels and leniency variations results in serious consequences in terms of person performance inference. With the 2P-FCRM-RL, the leniency parameters of the three raters were estimated to be −0.13, 0.00, and 0.13, and the corresponding variances were estimated to be 0.15, 0.27, and 0.36. Although the values of these rater leniency estimates are not substantial, the effects of the different raters on the estimates of the students’ ability were nontrivial, especially considering the small sample size in this example.
FIGURE 2. Relationships between the latent trait estimates calibrated using the two-parameter many-faceted forced-choice ranking model with random leniency, two-parameter forced-choice ranking model with raters’ leniency, and two-parameter forced-choice ranking model for three dimensions. (A) Dimension 1: Presentation. (B) Dimension 2: Usefulness. (C) Dimension 3: Originality.
Conclusion

The evaluation of individuals’ performance as assessed by external raters is a commonly used approach for understanding the actual level at which a ratee performs in terms of various prespecified skills, attributes, or symptoms. Since individuals’ performance is assessed by human raters and subjective judgments are consequently inevitable, it is necessary to develop appropriate modeling methods for working with rater-mediated assessment data. Although many psychometric models and statistical post hoc corrections have been proposed in the literature to address rater errors, the methods proposed to control for the halo effect are controversial because the existing approaches have various apparent limitations. Acknowledging the possible interdependence among conceptually distinct traits when raters are issuing judgments across ratees on certain rating scales (Myford & Wolfe, 2003), we adopted forced-choice items rather than single-stimulus items to avoid the indistinguishability problem arising for judgments issued in the form of Likert-type questionnaires. An FCRM for raters’ ranking data and variants that incorporate a topic facet and the randomness of raters’ leniency were developed in this study to capture the nature of subjective human judgments and to control for the halo effect using a generalized HCM as the probability function to represent the relationship between item evaluation and the underlying cognitive process. The proposed models are highly flexible and general, allowing researchers to readily and easily develop customized models by modifying our models to fit the needs of various practical testing situations.

A series of simulations were conducted to evaluate the success of parameter recovery with the proposed FCRMs using Bayesian estimation by manipulating the number of ratees, the number of item blocks, and the completeness of the ranking design. The results indicate that better parameter recovery for the model structural parameters is associated with a larger number of item blocks, a larger sample size, and a complete ranking design, while better parameter recovery for the latent trait parameters of individuals is associated with a larger number of item blocks and a complete ranking design. When multiple tasks are evaluated, the results indicate well-recovered parameters and better parameter estimates than when only a single task is ranked under otherwise identical conditions. In the final simulation study, the interactions between ratees and raters were considered by treating the leniency parameters as random-effect parameters rather than fixed-effect parameters, and the results show that although the parameter recovery performance deteriorated slightly compared to that in the previous simulation studies, the more complicated model nevertheless provided acceptable parameter estimates. In addition, severely biased estimates were obtained when rater effects were present but were ignored by fitting the simulated data to the traditional FCRM (without rater effects), suggesting that raters’ impact should not be neglected.
A technological creativity assessment was presented as an empirical example to show how the proposed models can be applied to fit data. The data were collected before the use of smartphones became prevalent, and the recruited students were expected to exhibit their innovation as best as they could. Limited by the available research resources, the dataset was composed of ranking results for only 384 students evaluated by three experts in the form of forced-choice ranking items presented as 24 item blocks measuring three dimensions. To address several important concerns, we used various FCRMs to fit the data. The latent trait estimates extracted with the 2P-FCRM, 2P-FCRM-L, and 2P-FCRM-RL were compared to show the impact on ability estimation of using misleading models for fitting, with the 2P-FCRM-RL being the best fitting model. The results indicated that neglecting raters’ leniency levels and leniency variations had nontrivial impacts on person parameter estimation. As mentioned above, however, the small sample size used in this analysis may influence the stability and precision of parameter estimation, and the results should be interpreted cautiously.

Recalling anonymous reviewers’ comments, further consideration that a rater’s ranking pattern may not be ideally decomposed into multiple successive subranking events is warranted, and therefore, the assumption of the explosion rule should be examined explicitly and justifiably for our empirical data analysis. Indeed, the literature has indicated that some nuisance factors may be introduced to the rank-ordered choice process if the exploded logit model was applied to fit the data and would result in biased estimation (Chapman & Staelin, 1982; Fok et al., 2012). For example, an unexperienced person is capable of selecting the most preferred items but may fail to indicate the less preferred items, is likely to be annoyed with a larger number of ranking items, or may rank order their choice set by successively deleting inferior items from consideration, all of which will compromise the validity of the exploded logit model. Although several probabilistic ranking models have been proposed to deal with those unexpected choice behaviors (e.g., Fok et al., 2012; Hausman & Ruud, 1987), we decided not to consider the likelihoods resulting from the nuisances in our raters’ ranking data for the following reasons. First, the appropriate fit of the proposed model (i.e., the 2P-FCRM-RL) to the ranking data was assessed statistically and theoretically within a Bayesian framework (Sinharay et al., 2006), and a good model-data fit implied that the explosion rule was verified to dominate raters’ rank order process. Second, the rank ordering of a limited number of comparative items produced by raters (e.g., three items used in our empirical data) can be expected to mitigate rater burden and increase measurement efficiency, as suggested by previous studies (Crompvoets et al., 2020; Steedle & Ferrara, 2016). Third, because rater training is necessarily arranged prior to any formal ranking, it is justifiable to expect that trained raters are more likely to exhibit rational choice behavior (i.e., items are ranked from most to least representative of the target’s
performance) than ordinary persons, and the explosion process can reasonably apply to the rank-ordered choice sets produced by human raters.

Although this study confirmed the efficiency and applicability of the proposed models in the simulation studies and empirical demonstration, an important question may arise regarding when to use the newly developed FCRMs to fit data in real rater-mediated assessments. From a methodological perspective, the nature of single-stimulus rating scales is definitely different from that of forced-choice ranking formats in terms of item construction, scoring, and interpretation. Performance assessments produced by external raters are dominated by Likert-type rating scales due to their popularity and accessibility (e.g., Wang & Engelhard, 2019). While raters’ rating data are collected and analyzed to inform decisions, researchers should be cautious in light of the detection of raters’ distortions in making judgments. Some rater errors can be addressed appropriately in the traditional IRT model framework; however, several rater effects, such as the halo effect and response style, have not been efficiently eliminated using single-stimulus formats (Murphy et al., 1993; Myford & Wolfe, 2003). If detection statistics flag raters as having halo bias (Myford & Wolfe, 2004), forced-choice ranking formats should be considered to replace single-stimulus formats and can be easily constructed by assembling multiple evaluation criteria (i.e., single-stimulus items) that measure distinct latent traits to form ranking blocks (as used in the forced-choice Big Five personality assessment; see Brown & Maydeu-Olivares, 2011). In addition, the WinBUGS codes are readily available in the Online Appendix, and readers can easily modify the codes to produce customized FCRMs for rater data analysis.

The proposed FCRMs assume that the probability of a particular ranking pattern can be partitioned into a sequence consisting of the independent probability of selecting the most preferred item from the diminishing set of alternative items. Within the context of the TIRT model, this assumption amounts to assuming a sequence of independent Thurstone’s Case V models. The actual TIRT model—employed, for instance, by Brown et al. (2017)—assumes that the probability of a particular ranking pattern depends instead on a set of interdependent response processes. The capture of ranking patterns by the TIRT model has been illustrated in previous studies (Maydeu-Olivares, 1999; Maydeu-Olivares & Böckenholt, 2005). Future studies are encouraged to compare the models introduced here in terms of model-data fit to alternative models, such as the TIRT model employed by Brown et al. (2017), to shed light on the nature of ranking pattern responses.

In this study, several commonly observed rater effects and rater biases have been considered to develop a new class of FCRMs. Other types of rater bias, however, such as logical errors (Linn & Gronlund, 2000), contrast errors (Guilford, 1954), raters’ attitudes (Murphy & DeShon, 2000), and order effects (Hopkins, 1998), have not been considered here and deserve further attention to develop more appropriate measurement models. In some cases, raters may
exhibit differential leniency, thus influencing their ratings or rankings for a particular group of ratees (e.g., in terms of gender, age, or ethnicity), and differential rater functioning may arise (Myford & Wolfe, 2003; Murphy & DeShon, 2000). To address this concern, the current FCRMs can be extended by allowing the leniency parameter of a rater to be separately estimated for different groups. Furthermore, differential rater functioning may arise across ranking criteria/dimensions rather than across ratee groups. For example, a rater might be more severe in selecting the “emotionally controlled” dimension than the “adaptable” dimension in the case of the forced-choice items on the Occupational Personality Questionnaire (SHL, 2013). If a rater exhibits differential leniency levels across different dimensions, then distinct leniency parameters for different dimensions should be included in the FCRM. Finally, an alternative model for handling forced-choice ranking items is the generalized logit IRT (GLIRT) model in the framework of the Rasch model (Wang et al., 2016). The questions of how to apply the GLIRT model to analyze raters’ ranking data and how effective this model is compared to our proposed models would be interesting topics for future study.

Acknowledgments
The authors would like to thank three anonymous reviewers and the Editor for their helpful and constructive comments on earlier versions of this 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 and/or authorship of this article: The first author was supported by the Ministry of Science and Technology, Taiwan (No. 108-2410-H-006-045), and the second author was supported by the Ministry of Science and Technology, Taiwan (No. 109-2410-H-845-015-MY3).

ORCID iD
Hung-Yu Huang https://orcid.org/0000-0001-6244-1950

References
Ahn, J., Lee, J., Lee, J. D., & Kim, T. Y. (2006). An analysis of consumer preferences among wireless LAN and mobile internet services. ETRI Journal, 28(2), 205–215. https://doi.org/10.4218/etrij.06.0105.0106
Amabile, T. M. (1996). Creativity in context. Westview Press.
Andrich, D. (1995). Hyperbolic cosine latent trait models for unfolding direct responses and pairwise preferences. Applied Psychological Measurement, 19(3), 269–290. https://doi.org/10.1177/014662169501900306
Andrich, D., & Luo, G. (1993). A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement, 17*(3), 253–276. https://doi.org/10.1177/014662169301700307

Andrich, D., & Luo, G. (2019). A law of comparative preference: Distinctions between models of personal preference and impersonal judgment in pair comparison designs. *Applied Psychological Measurement, 43*(3), 181–194. https://doi.org/10.1177/0146621617738014

Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics, 17*(1), 1–19. https://doi.org/10.1016/0304-4076(81)90056-7

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison Wesley.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

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. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135–1147. https://doi.org/10.3758/s13428-012-0217-x

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., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods, 20*(1), 121–148. https://doi.org/10.1177/1094428116668036

Calfee, J., Winston, C., & Stempski, R. (2001). Econometric issues in estimating consumer preferences from stated preference data: A case study of the value of automobile travel time. *Review of Economics and Statistics, 83*(4), 699–707. https://doi.org/10.1162/003465301753237777

Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research, 19*(3), 288–301. https://doi.org/10.1177/002224378201900302

Cheung, M. W.-L., & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9*(1), 55–77. https://doi.org/10.1207/S15328007SEM0901_4

Coombs, C. H. (1964). *A theory of data*. John Wiley.

Crompvoets, E. A. V., Béguin, A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics, 45*(3), 316–338. https://doi.org/10.3102/1076998619890589

de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2012, April ). Examining the viability of recent models for forced-choice data [Paper presentation]. Meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.
Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! Industrial and Organizational Psychology, 3(4), 465–476. https://doi.org/10.1111/j.1754-9434.2010.01273.x

Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. Journal of Educational Measurement, 31(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Florida, R., Mellander, C., & King, K. (2015). The global creativity index 2015. Martin Prosperity Institute.

Freund, P. A., & Holling, H. (2008). Creativity in the classroom: A multilevel analysis investigating the impact of creativity and reasoning on GPA. Creativity Research Journal, 20(3), 309–318. https://doi.org/10.1080/10400410802278776

Fok, D., Paap, R., & van Dijk, A. (2012). A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. Journal of Applied Econometrics, 27(5), 831–846. https://doi.org/10.1002/jae.1223

Gajda, A., Karwowski, M., & Beghetto, R. A. (2017). Creativity and academic achievement: A meta-analysis. Journal of Educational Psychology, 109(2), 269–299. https://doi.org/10.1037/edu0000133

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4), 733–807. https://www.jstor.org/stable/2430636

Gensch, D. H., & Recker, W. W. (1979). The multinomial, multiattribute logit choice model. Journal of Marketing Research, 16(1), 124–132. https://doi.org/10.1177/002224377901600117

Guilford, J. P. (1954). Psychometric methods (2nd ed.). McGraw Hill.

Hausman, J., & Ruud, P. (1987). Specifying and testing econometric models for rank-ordered data. Journal of Econometrics, 34(1–2), 83–104. https://doi.org/10.1016/0304-4076(87)90068-6

Hernández-Torrano, D., & Ibrayeva, L. (2020). Creativity and education: A bibliometric mapping of the research literature (1975–2019). Thinking Skills and Creativity, 35, 100625. https://doi.org/10.1016/j.tsc.2019.100625

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

Hopkins, K. D. (1998). Educational and psychological measurement and evaluation (8th ed.). Allyn & Bacon.

Hung, S.-P., Chen, P.-H., & Chen, H.-C. (2012) Improving creativity performance assessment: A rater effect examination with many facet Rasch model. Creativity Research Journal, 24(4), 345–357. https://doi.org/10.1080/10400419.2012.730331

Jin, K.-Y., & Wang, W.-C. (2015). Item response theory models for carry-over effect across different scales. Applied Psychological Measurement, 39(5), 406–425. https://doi.org/10.1177/0146621615572250

Jin, K.-Y., & Wang, W.-C. (2018). A new facets model for rater’s centrality/extremity response style. Journal of Educational Measurement, 55(4), 543–563. https://doi.org/10.1111/jedm.12191
Joo, S. H., Lee, P., & Stark, S. (2018). Development of information functions and indices for the GGUM-RANK multidimensional forced choice IRT model. *Journal of Educational Measurement, 55*(3), 357–372. https://doi.org/10.1111/jedm.12183

Keane, L., & Keane, M. (2016). STEAM by design. *Design and Technology Education, 21*(1), 61–82. https://ojs.lboro.ac.uk/DATE/article/view/2085/2256

Laming, D. (2004). *Human judgment: The eye of the beholder*. Thomson Learning.

Lee, P., Joo, S. H., Stark, S., & Chernyshenko, O. S. (2019). GGUM-RANK statement and person parameter estimation with multidimensional forced choice triplets. *Applied Psychological Measurement, 43*(3), 226–240. https://doi.org/10.1177/0146621618768294

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Merrill.

Liu, C.-W., & Wang, W.-C. (2016). Unfolding IRT models for Likert-type items with a don’t know option. *Applied Psychological Measurement, 40*(7), 517–533. https://doi.org/10.1177/0146621616664047

Long, H., & Pang, W. (2015). Rater effects in creativity assessment: A mixed methods investigation. *Thinking Skills and Creativity, 15*, 13–25. https://doi.org/10.1016/j.tsc.2014.10.004

Lord, F. (1965). *Item sampling in test theory and in research design*. Educational Testing Service.

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Dover Publication.

Luo, G. (2001). A class of probabilistic unfolding models for polytomous responses. *Journal of Mathematical Psychology, 45*(2), 224–248. https://doi.org/10.1006/jmps.2000.1310

Manski, C. (1977). The structure of random utility models. *Theory and Decision, 8*, 229–254. https://doi.org/10.1007/BF00133443

Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika, 64*(3), 325–340. https://doi.org/10.1007/BF02294299

Maydeu-Olivares, A., & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods, 10*(3), 285–304. https://doi.org/10.1037/1082-989X.10.3.285

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P. M., 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/10.1177/014662161662226

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*(4), 873–900. https://doi.org/10.1111/j.1744-6570.2000.tb02421.x

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

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.
Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice*, 21(2), 205–220. https://doi.org/10.1080/0969594X.2013.868341

Perignat, E., & Katz-Buonincontro, J. (2019). STEAM in practice and research: An integrative literature review. *Thinking Skills and Creativity*, 31, 31–43. https://doi.org/10.1016/j.tsc.2018.10.002

Plucker, J. A., Guo, J., & Dilley, A. (2018). Research-guided programs and strategies for nurturing creativity. In S. I. Pfeiffer, E. Shaunessy-Dedrick, & M. Foley-Nicpon (Eds.), *APA handbook of giftedness and talent* (pp. 387–397). American Psychological Association.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354

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(1), 3–32. https://doi.org/10.1177/01466216000241001

SHL. (2013). *OPQ32r technical manual version 1.0*. SHL Group.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4), 298–321. https://doi.org/10.1177/0146621605285517

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Methodological*, 64(4), 583–616. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS Version 1.4* [Computer program]. MRC Biostatistics Unit, Institute of Public Health.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, 29(3), 211–223. https://doi.org/10.1080/08957347.2016.1171769

Sternberg, R. J. (2002). Raising the achievement of all students: Teaching for successful intelligence. *Educational Psychology Review*, 14(4), 383–393. https://doi.org/10.1023/A:1020601027773

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34(4), 273–86. https://doi.org/10.1037/h0070288

Van der Heijden, B., & Nijhof, A. (2004). The value of subjectivity: Problems and prospects for 360-degree appraisal systems. *The International Journal of Human Resource Management*, 15(3), 493–511. https://doi.org/10.1080/0958519042000181223

Wang, J., & Engelhard, G., Jr. (2019). Exploring the impersonal judgments and personal preferences of raters in rater-mediated assessments with unfolding models. *Educational and Psychological Measurement*, 79(4), 773–795. https://doi.org/10.1177/0013164419827345

Wang, W.-C., Liu, C.-W., & Wu, S.-L. (2013). The random-threshold generalized unfolding model and its application of computerized adaptive testing. *Applied Psychological Measurement*, 37(3), 179–200. https://doi.org/10.1177/0146621612469720
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., Su, C.-M., & Qiu, X.-L. (2014). Item response models for local dependence among multiple ratings. Journal of Educational Measurement, 51(3), 260–280. https://doi.org/10.1111/jedm.12045

Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random effects facet model. Applied Psychological Measurement, 29(4), 296–318. https://doi.org/10.1177/0146621605276281

Wilson, M., & Hoskens, M. (2001). The rater bundle model. Journal of Educational and Behavioral Statistics, 26(3), 283–306. https://doi.org/10.3102/10769986026003283

Authors

SU-PIN HUNG is an associate professor at the Center for Teacher Education, National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan; email: sphung@mail.ncku.edu.tw. Her research interests include creativity assessments, cognitive diagnosis modeling, and applied statistical analysis.

HUNG-YU HUANG is a distinguished professor at the Department of Psychology and Counseling, University of Taipei, No.1, Ai-Guo West Road, Taipei 10048, Taiwan; email: hyhuang@go.utaipei.edu.tw. His research interests include item response theory, cognitive diagnosis modeling, and computerized adaptive testing. He is the corresponding author of this article.

Manuscript received April 7, 2020
Revision received April 20, 2022
Accepted May 9, 2022