Diagnostic Classification Model for Forced-Choice Items and Noncognitive Tests

Hung-Yu Huang

Abstract
The forced-choice (FC) item formats used for noncognitive tests typically develop a set of response options that measure different traits and instruct respondents to make judgments among these options in terms of their preference to control the response biases that are commonly observed in normative tests. Diagnostic classification models (DCMs) can provide information regarding the mastery status of test takers on latent discrete variables and are more commonly used for cognitive tests employed in educational settings than for noncognitive tests. The purpose of this study is to develop a new class of DCM for FC items under the higher-order DCM framework to meet the practical demands of simultaneously controlling for response biases and providing diagnostic classification information. By conducting a series of simulations and calibrating the model parameters with a Bayesian estimation, the study shows that, in general, the model parameters can be recovered satisfactorily with the use of long tests and large samples. More attributes improve the precision of the second-order latent trait estimation in a long test, but decrease the classification accuracy and the estimation quality of the structural parameters. When statements are allowed to load on two distinct attributes in paired comparison items, the specific-attribute condition produces better a parameter estimation than the overlap-attribute condition. Finally, an empirical analysis related to work-motivation measures is presented to demonstrate the applications and implications of the new model.

Keywords
diagnostic classification model, forced-choice format, pairwise comparison, Bayesian estimation

1University of Taipei, Taiwan

Corresponding Author:
Hung-Yu Huang, Distinguished Professor, Department of Psychology and Counseling, University of Taipei, No.1, Ai-Guo West Road, Taipei, 10048, Taiwan.
Email: hyhuang@go.utaipei.edu.tw
Noncognitive tests are typically designed to measure individuals’ trait levels of psychological variables. They involve typical behavior performance, such as personality, interests, attitudes, and values, and predominantly use Likert-type items in a rating scale to provide measurement outcomes of self-report questionnaires. When the collected data are fit to a psychometric model, such as item response theory (IRT) models (Lord, 1980), concerns arise regarding the assessment of continuous latent traits and absolute traits, which reflect the benefits and ultimate purpose of normative tests. However, normative test scores are more susceptible to motivated distortion for applicants than for nonapplicants when the test has consequences (e.g., high-stake evaluations or job hiring situations) due to socially desirable responses or attempts to create a positive impression (Christiansen et al., 2005; Donovan et al., 2003; Rosse et al., 1998). In addition, there is little information about within-person differentiation if a person shows a consistent tendency to endorse certain categories on a rating scale (e.g., the neutral option on a 5-point rating scale) throughout all items (Matthews & Oddy, 1997).

In contrast to normative tests, the forced-choice (FC) items used in ipsative tests provide an alternative approach by directing a respondent to make judgments among a set of limited options and to choose the one that is most representative of himself or herself relative to others. Evidence shows that social desirability and response biases (e.g., extremity, acquiescence, and the halo effect) can be effectively controlled because the options that measure different traits are matched to be equally acceptable and attractive in terms of their desirability or extremity (Brown, 2016; Brown & Maydeu-Olivares, 2013; Christiansen et al., 2005; Stark et al., 2005). Well-known examples of the application of FC item formats to noncognitive measurement include the Edward Personal Preference Schedule (Borislow, 1958), the Gordon Personal Profile and Inventory (Hausknecht, 2010), the Occupational Personality Questionnaire (SHL, 2013), and the Kuder Occupational Interest Survey (Kuder & Diamond, 1979). Although FC items have been found to be more advantageous than single-stimulus items used in normative tests and are commonly used in personality and interest measurement, the nature of ipsative scoring remains controversial. The outcome measures are obtained by summing the frequency of the response choices for each statement and measuring different latent traits across the FC items; thus, person-centered measures hinder between-person comparisons and may result in mathematical dependency among multiple trait scales (Cornwell & Dunlap, 1994; Meade, 2004).

To meet the practical demands of providing scoring comparisons between and within individuals in ipsative tests, a variety of IRT models have been proposed for FC data analysis based on either the ideal-point approach or the dominance approach (Andrich, 1995; Brown & Maydeu-Olivares, 2011; Morillo et al., 2016; Stark et al., 2005; W.-C. Wang et al., 2017). Furthermore, FC item formats administered to examinees range from choosing the most preferred statement in pairwise-comparison items to ranking statements from most preferred to least preferred in ranking items (de la Torre et al., 2012; Hontangas et al., 2015; Joo et al., 2018; W.-C. Wang et al., 2016).
These formats have also been applied to computerized adaptive testing (CAT; Chen et al., 2020; Joo et al., 2020; Stark et al., 2012).

In some cases, such as clinical diagnostic settings, the preferred approach to classifications of test takers in assessments of personality or psychological disorders is to simply locate them on a continuum from low to high levels of latent traits. That is, individuals can be diagnosed and classified as either having or not having a certain type of personality or a specific disorder in a single analysis without additionally determining the thresholds on a latent continuum to define diagnostic classification through a latent continuous trait model (e.g., IRT models). Therefore, diagnostic classification models (DCMs, also known as cognitive diagnosis models; Rupp et al., 2010) have been developed to characterize the relationship of the latent discrete variables (also called attributes or skills) to the observed responses in a test and can provide diagnostic information as to whether test takers possess a measured attribute (e.g., a certain type of personality, disorder, or interest) rather than calibrating along a latent trait continuum.

There is abundant research and literature on the applications of DCMs to educational testing situations because DCMs allow for the efficient collection of finer-grained feedback on examinees' strengths and weaknesses in cognitive skills, and remedial instruction interventions can be provided immediately by adjusting teaching strategies (Huang, 2017; Lee et al., 2011; Leighton & Gierl, 2007; Ravand, 2016). In addition to the application of DCMs to educational and achievement tests, Templin and Henson (2006) proposed a disjunctive version of a DCM, the deterministic input noisy or gate (DINO) model, and applied this model to diagnose the degrees of a psychological gambling disorder. With the exception of their research, most studies have used traditional exploratory clustering methods to classify individuals into different classes in terms of different personality types (e.g., Fals-Stewart et al., 1994; Solomon et al., 2001), and a few studies have applied DCMs to personality assessment (e.g., Revuelta et al., 2018).

Although DCMs and other latent class models have been applied in personality and noncognitive tests, single-stimulus formats, such as dichotomous items and Likert-type items (which may be dichotomized for simplicity; see Templin & Henson, 2006), are commonly used for testing and are susceptible to social desirability and fakability. There are few latent discrete variable models in the literature that involve administering FC items and classifying multiple latent noncognitive attributes without using post hoc clustering methods or a cutoff point score on a latent continuum. Therefore, the main purpose of this study is to develop a new class of multiple classification latent class models for FC items on noncognitive tests, which represents the most significant contribution of this study to psychometric theory and testing practices.

The following discussion focuses on pairwise-comparison items in which respondents compare statements in an FC item to determine which statement is preferred. Respondents can rank more than two statements from the most preferred to the least preferred, and the ranking data can then be reduced to multiple paired comparisons.
This study is structured as follows. First, the classic comparative judgment theorem for pairwise comparison under the IRT framework is briefly reviewed, and a new class of pairwise-comparison models for measuring the latent discrete variables in FC items is introduced. Second, simulations are designed to evaluate the efficiency of the proposed model under a variety of manipulated conditions using Bayesian estimation, and the results are summarized from the parameter recovery. Third, the empirical data that were collected from the FC items to measure work motivation are used to demonstrate the applications and implications of the new models. Finally, this article presents overall conclusions of the results and provides suggestions and directions for future studies.

Existing IRT Approaches for FC Items

According to the law of comparative judgment (Thurstone, 1927), comparative statements involve unique psychological values or utilities that reflect respondents’ preferences. The favorable choice between two statements is determined by the difference of the two utilities. Let \( t_A \) and \( t_B \) be the utility parameters for statements \( A \) and \( B \) in an FC item. A latent response function can be specified for each utility parameter to represent the impacts of the underlying latent constructs on the desirability of a statement, which is given by

\[
t_A = f(\theta_{d_A}; \xi_A)
\]

and

\[
t_B = f(\theta_{d_B}; \xi_B)
\]

where \( \theta_{d_A} \) and \( \theta_{d_B} \) are the latent trait parameters for dimensions \( d_A \) and \( d_B \) that govern statements \( A \) and \( B \), respectively, and \( \xi_A \) and \( \xi_B \) are the fixed-effect statement parameter vectors (e.g., the intercept and slope parameters in a linear function) for statements \( A \) and \( B \), respectively. When the latent response function for each utility parameter is assumed to be a linear combination of the latent trait parameter and the statement parameter with error variance, the Thurstonian item response theory (TIRT) model can be formulated (Brown, 2016; Brown & Maydeu-Olivares, 2011, 2013). Subsequently, the response to the pairwise-comparison item, \( y_{ij(A>B)} \), is coded as one if statement \( A \) is preferred over statement \( B \) (i.e., \( t_A \geq t_B \)) and zero otherwise (i.e., \( t_A < t_B \)). Finally, the probability of preferring statement \( A \) to statement \( B \) in item \( i \) for person \( j \) can be expressed as

\[
P(y_{ji(A>B)} = 1 | \theta) = \Phi \left[ \frac{f(\theta_{d_A}; \xi_A) - f(\theta_{d_B}; \xi_B)}{\sqrt{\psi_A^2 + \psi_B^2}} \right]
\]
where $\Phi$ is the cumulative standard normal distribution function (i.e., the probit link function), $\psi^2_A$ and $\psi^2_B$ are residual variances for utility $A$ and $B$, respectively, and the other variables are defined as above. Figure 1(A) visualizes the conceptual diagram of the three-dimensional TIRT model when an FC item is composed of two statements that measure different dimensions using linear (between $\theta$ and $t$) and nonlinear (between $t$ and $y$) response functions.

The choice between statements in an FC item involves a discriminal process that can be represented by the discriminal response function for each statement (Andrich & Luo, 2019). In TIRT, the discriminal response function is assumed to be monotonic when using a cumulative approach; that is, when a statement’s utility is higher, it is more likely that this statement will be selected. However, a discriminal response does not always follow a cumulative response pattern, as TIRT assumes. Rather, the highest selection probability may occur when the statement matches the person’s option (i.e., ideal point), and an unfolding model with single-peaked response patterns can be applied to the discriminal response functions of statements. Following the unfolding discriminal response function, the multi-unidimensional pairwise preference (MUPP) model has been developed for pairwise-comparison items (Stark et al., 2005) and has been extended to the fit of ipsative ranking items (de la Torre et al., 2012; Hontangas et al., 2015; Joo et al., 2018). In addition, Morillo et al. (2016) replaced the unfolding item response function used in the MUPP model with a dominance item response function and created an alternative FC model that is almost equivalent to the TIRT model. If the measurement of specific objectivity is desired, then the Rasch ipsative model is readily available to analyze pairwise-comparison and ranking items using the dominance approach, as used in TIRT (W.-C. Wang et al., 2016, 2017).

The Forced-Choice Diagnostic Classification Model

FC items are typically developed to minimize intentional response distortion, which has been commonly observed in normative tests by equating response choices on attractiveness but differentiating them in terms of their validity. That is, each choice option (i.e., statement) in FC items is generally designed to represent a different latent trait, and theoretically, construct validity should be satisfied for these response options (Brown & Maydeu-Olivares, 2013, p. 36; Christiansen et al., 2005, p. 269). Following the FC methodology tradition (e.g., Brown & Maydeu-Olivares, 2011), we use the term “item” to indicate an FC item or a pairwise-comparison item, and each item is composed of two statements that measure different latent traits (e.g., latent continuous or binary variables) to be compared and chosen by individuals. Note that an item constructed with FC formats must have two or more statements, which is definitely different from Likert-type item formats where in most cases, a Likert-type item only involves one dimension and is designed to reflect respondents’ different levels of endorsement in terms of one specific statement (e.g., strongly disagree, disagree, agree, and strongly agree). Because comparative statements are
governed by distinct latent traits, an FC-format test is generally considered a multidimensional rather than unidimensional assessment.

Figure 1. Graphical Representation of (A) the TIRT Model and (B) the FC-DCM

Note. The observable item responses indicated by rectangles are assumed to follow a specific Bernoulli distribution. As, Bs, and Cs are statements that measure (A) the three respective latent continuous latent traits and (B) the three respective latent discrete attributes in the FC items. Solid and dotted arrows represent linear and nonlinear relationships, respectively. TIRT = Thurstonian item response theory; FC-DCM = forced-choice diagnostic classification model.

governed by distinct latent traits, an FC-format test is generally considered a multidimensional rather than unidimensional assessment.
A significant study of the five-factor personality format that uses FC items has demonstrated the application in which the statements that measure one of the five personality traits of neuroticism, extraversion, openness, agreeableness, and conscientiousness are compared and selected according to the respondent’s judgment within an item block (i.e., an FC item); choosing one statement means that another statement is not representative of the respondent (Brown & Maydeu-Olivares, 2011). It is rare that two comparative statements that measure the same dimension but are presented in opposite directions are judged by respondents (e.g., two statements that measure introversion and extraversion are simultaneously presented to be compared in an FC item), except for some applications in the literature (e.g., DeVito, 1985; Myers et al., 1998). Following the commonly used approach to constructing FC items, we assume that the comparative statements in an FC item measure distinct dimensions (or attributes) in both our proposed model and in the following empirical data analysis.

In the DCM framework, multiple latent continuous traits are replaced by multiple latent discrete attributes. These attributes can be thought of as dimensions because test takers’ responses to items are determined by their mastery of the attributes (Templin & Bradshaw, 2013). Note that the latent attributes can be defined in terms of multiple categories; for simplicity, this study constrains the attributes to latent binary variables. Figure 1(B) shows an illustrative example of a three-attribute DCM with FC formats in which the three attributes (denoted by \( a_k \)) of interest are to be measured and each attribute consists of several statements to be matched across attributes to form pairwise-comparison items. The responses to pairwise-comparison items are indicated by the rectangle symbols, and the probability of endorsing a prespecified statement preference pattern (e.g., preferring statement \( A \) over \( B \)) for each item can be determined by individual statement utility parameters and test takers’ mastery of comparative attributes (see below for more details). Because the measurement outcomes are binary variables, in contrast to linear factor models, the distribution of each response is assumed to follow a Bernoulli distribution using a nonlinear factor modeling approach with an identical-link function (de la Torre, 2011, p. 181; Maydeu-Olivares, 2005, p. 74). In addition, the statements are assumed to be fully representative of a prespecified set of attributes, and their effects on the response probability of an FC item are determined by each individual statement utility parameter.

As in traditional factor models, the associations between latent attributes (i.e., traits) should be taken into account. The DCM literature has documented several approaches to modeling the dependencies between latent attributes (Culpepper & Balamuta, In Press), and we revisit the relevant issues in the final discussion section. In our developed FC-DCM, we assume that the attributes are governed by a second-order latent continuous trait (\( \theta \)) and that the relationships among latent attributes can be fully characterized by the second-order latent trait under the higher-order DCM framework (de la Torre & Douglas, 2004). Again, because the attributes are latent binary variables, a nonlinear response function should be specified to relate \( \theta \) to \( \alpha \). Specifically, the mastery status of each attribute has a Bernoulli distribution, and the
corresponding probability can be transformed by using a logit-link function to relate to the combination of person (i.e., $\theta$) and attribute parameters (i.e., attribute discrimination and location parameters; see below).

The relationship between the higher-order latent trait and lower-order latent attributes can be modeled using the higher-order DCM approach proposed by de la Torre and Douglas (2004), which assumes a cumulative response process and uses a dominance IRT model for the probability curve of mastering an attribute. Therefore, the probability of test taker $j$ mastering attribute $k_A$ can be given by

$$P(\alpha_{jk_A} = 1|\theta_j, \delta_{0k_A}, \delta_{1k_A}) = \frac{\exp[\delta_{1k_A}(\theta_j - \delta_{0k_A})]}{1 + \exp[\delta_{1k_A}(\theta_j - \delta_{0k_A})]}$$ (4)

where $\theta_j$ is the level of the second-order latent trait for person $j$ and is assumed to follow a standard normal distribution and $\delta_{1k_A}$ and $\delta_{0k_A}$ are the attribute discrimination and location parameters, respectively, for attribute $k_A$. The probability of mastering other attributes ($k_B, k_C$, or others) can be formulated based on the same rationale.

Next, a Q-matrix that maps the latent attribute onto the observable FC response should be specified a priori. Two statements that are designed to measure two distinguishable attributes can be compared in an FC item to force test takers to endorse the one that is most preferred or most representative of themselves. An illustrative Q-matrix is presented in Appendix A in which 10 attributes underlie the responses to 30 pairwise-comparison items, with each attribute having five statements and each statement being exclusively accounted for by a single attribute. We use this Q-matrix exemplification to conduct simulations to evaluate the identifiability issue in the proposed FC-DCM, as described below. Note that a more complex Q-matrix that allows each statement to measure more than one attribute can be constructed straightforwardly to build a more general FC-DCM, although it is not shown here due to space constraints.

Let $y_{ji}$ be the response of person $j$ to pairwise-comparison item $i$, which is composed of two statements (e.g., statements $A_s$ and $B_s$) that measure two different attributes (e.g., $k_A$ and $k_B$). A value of one indicates that the former statement (i.e., statement $A$) is more preferred than the latter statement (i.e., statement $B$) by person $j$, and a value of zero indicates the opposite. Note that multiple statements can be designed to measure a single latent attribute, and a statement may be administered again in an FC item when appropriate. The probability of selecting statement $A_s$ over $B_s$ in item $i$ for person $j$ can be formulated conditional on the four attribute patterns of $(\alpha_{k_A}=0, \alpha_{k_B}=1)$, $(\alpha_{k_A}=1, \alpha_{k_B}=1)$, $(\alpha_{k_A}=0, \alpha_{k_B}=0)$, and $(\alpha_{k_A}=1, \alpha_{k_B}=0)$, which are given by

$$P(y_{ji}(A_s > B_s) = 1|\alpha_{k_A} = 0, \alpha_{k_B} = 1) = \eta_{i0}$$ (5)

$$P(y_{ji}(A_s > B_s) = 1|\alpha_{k_A} = 1, \alpha_{k_B} = 1) = \eta_{i1}$$ (6)

$$P(y_{ji}(A_s > B_s) = 1|\alpha_{k_A} = 0, \alpha_{k_B} = 0) = \eta_{i2}$$ (7)
Furthermore,

$$P(y_{i1}(A_i > B_j) = 1|a_{k_A} = 1, a_{k_B} = 0) = \eta_{iA_1} \times \alpha_{k_A} + \eta_{iB_1} \times (1 - \alpha_{k_B})$$  \hspace{1cm} (8)$$

where $\eta_{i0}$ is the probability of preferring statement $A_s$ to $B_s$ in item $i$ when $a_{k_A} < a_{k_B}$ and is analogous to the guessing parameter used in the DINA model (Junker & Sijtsma, 2001) because the required attribute pattern is reversed. Furthermore, $\eta_{i1}$ and $\eta_{i2}$ are the probabilities of preferring statement $A_s$ to $B_s$ in item $i$ when $a_{k_A} = a_{k_B}$ and can be constrained to .5 because no predominant attribute can be identified, whereas $\eta_{iA_1}$ and $\eta_{iB_1}$ are the two additive probabilities that represent the effect of attributes $k_A$ and $k_B$ on the probabilistic function for statements $A_s$ and $B_s$, respectively, when the predominant attribute is identical to the predominant statement and can be considered the statement utility parameters.

Taking the five personality traits test described above as an example (Brown & Maydeu-Olivares, 2011) and using the retrofitting approach that is commonly implemented in DCMs (see the “Demonstrative Example” section for more details), $\alpha_{k_A}$ can be represented as the attribute of conscientiousness, which underlies the two statements “I pay attention to details” (i.e., $A_1$) and “I am always prepared” (i.e., $A_2$), and $\alpha_{k_B}$ can be represented as the attribute of neuroticism, which underlies the two statements “I change my mood a lot” (i.e., $B_1$) and “I get upset easily” (i.e., $B_2$). Notably, the two statements governed by the same attribute are not indicative of different performance levels but are distinct and typical behaviors under each attribute, and in some cases, an opposite domain may be used to replace the original attribute (e.g., neuroticism is replaced by emotional stability, see Goldberg, 1992), where the behavioral statement should be scored inversely. When a person confronts two comparative statements in an FC item, the response of preferring statement $A_s$ (i.e., $A_1$ or $A_2$) to $B_s$ (i.e., $B_1$ or $B_2$) is coded as one, where the first item is designed to compare $A_1$ with $B_2$ and the second item is designed to compare $A_2$ with $B_1$. If the given person has a dominant attribute of conscientiousness over neuroticism (i.e., $\alpha_{k_A} = 1$ and $\alpha_{k_B} = 0$), then he or she is more likely to have the probability of scoring one following the conditional probabilistic function shown in Equation 8 than to have other conditional probabilities under the different attribute mastery patterns shown in Equations 5 to 7. Based on the same logic, the probability of the positive response for other attribute profiles, that is, $(\alpha_{k_A}, \alpha_{k_B}) = (0, 1), (1, 1), \text{or} (0, 0)$, can be specified according to the three respective conditional probabilities of Equations 5 to 7.

Note that unlike additive DCMs, such as the generalized DINA (G-DINA) model (de la Torre, 2011), where the change in the probability of endorsement is a result of having a given attribute, the FC-DCM is formulated to quantify the effects of the two distinct attributes. Because each pairwise-comparison item is composed of two statements and each statement has unique characteristics represented by their utility levels, the endorsement probability for the predominant attribute pattern in a given item, as shown in Equation 8, is expressed as an additive formulation of the two statement utility parameters. The major difference between the traditional additive DCMs and
our proposed FC-DCM is that in the FC-DCM, the additive effect of each statement arises when the respondent’s mastery attribute pattern is identical to the predominant attribute pattern that an item specifies and when the effects of mastering the predominant attribute and not mastering the inferior attribute are quantified for the respective statements. In contrast, traditional additive DCMs consider and quantify only the parameters for the presence of the attributes for each item. The two statement utility parameters can thus be separately estimated to represent the effects of different attributes on the comparative statements when the predominant attribute (i.e., $\alpha_{k_A} = 1$) is present and the inferior attribute (i.e., $\alpha_{kB} = 0$) is absent in the FC-DCM.

By combining the four probability functions listed in Equations 5 to 8, the full conditional probabilistic function can be expressed as

$$P(y_{ji(A_j > B_s)} = 1|\alpha) = \eta_{i0}^* + (.5 - \eta_{i0}^*) \times I(\alpha_{k_A} \geq \alpha_{kB}) + \left[ \eta_{iA_j}^* \times \alpha_{k_A} + \eta_{iB_s}^* \times (1 - \alpha_{kB}) \right] \times I(\alpha_{k_A} > \alpha_{kB})$$  

(9)

where $I(\cdot)$ is the indicator function that returns a value of one if the described condition is satisfied and zero otherwise, $\eta_{i0}^* = \eta_{i0}$, and $\eta_{iA_j}^* + \eta_{iB_s}^* = \eta_{iA_j} + \eta_{iB_s} - .5$. In the above demonstration (Equations 5 to 9), an FC item is composed of two comparative statements, and each statement is postulated to load on a distinct attribute; therefore, the FC-format measurement is multidimensional. Because each statement reflects a unidimensional nature (i.e., a statement is allowed to measure one attribute) and an FC item involves two statements to reflect a multidimensional nature, Stark et al. (2005) thus identified such response formats as “multiunidimensional pairwise preferences” (p. 187).

In some cases, a statement may load on more than one attribute simultaneously, and multidimensionality can occur within statements such that a complex Q-matrix should be developed for each paired statement in the FC items. It is not uncommon in practice for the indicators in psychological or personality tests to contribute to more than one dimension (e.g., Super, 1942). Therefore, the FC-DCM is not limited to FC items with two comparative attributes but can be extended to fit FC items with multiple comparative attributes.

To map statements to attribute $k$ ($k = 1, \cdots, K$) in a diagnostic FC test, the entry $q_{k(s)}$ in the Q-matrix can be used to indicate whether attribute $k$ is required by statement $s$, where $q_{k(s)} = 1$ if attribute $k$ must be mastered to choose statement $s$ as the most representative statement and $q_{k(s)} = 0$ otherwise. When a statement is allowed to measure more than one attribute, that is, $\sum_k q_{k(s)} > 1$, and the two statements of $A_s^*$ and $B_s^*$ are present for comparison, we can extend the above FC-DCM to a general form by relating the statements’ corresponding attributes and mapping the two vectors $q_{(A)}$ and $q_{(B)}$ to the probabilistic function. Note that $q_{(A)}$ and $q_{(B)}$ are not exclusive and may share some common attributes. In the DCM context, item responses can be assumed to follow either a conjunctive modeling approach (i.e., the presence of all required attributes results in a higher success probability) or a disjunctive modeling
approach (i.e., the mastery of any subset of attributes has the same success probability), which can apply to the FC-DCM when a statement requires multiple attributes. A latent binary variable can be defined to classify individuals into groups in terms of conjunctive or disjunctive assumptions for the two statements (the person subscript is omitted for clarity) as follows:

\[
\eta_{\text{A}} = \prod_{k=1}^{K} \alpha_k q_k(A) \tag{10}
\]

and

\[
\eta_{\text{B}} = \prod_{k=1}^{K} \alpha_k q_k(B) \tag{11}
\]

for the conjunctive modeling approach; and

\[
\xi_{\text{A}} = 1 - \prod_{k=1}^{K} (1 - \alpha_k)^q_k(A) \tag{12}
\]

and

\[
\xi_{\text{B}} = 1 - \prod_{k=1}^{K} (1 - \alpha_k)^q_k(B) \tag{13}
\]

for the disjunctive modeling approach.

As a result, the full conditional probability of selecting statement \(A_s^*\) over \(B_s^*\) in item \(i\) for person \(j\) can be modeled as

\[
P(y_{ji}(A_i^* > B_i^*)) = \eta_{\text{A}} + (.5 - \eta_{\text{B}}) \times I(\xi_{\text{A}} \geq \xi_{\text{B}}) + \left[\eta_{\text{A}} \times \xi_{\text{A}} + \eta_{\text{B}} \times (1 - \xi_{\text{B}})\right] \times I(\xi_{\text{A}} > \xi_{\text{B}}) \tag{14}
\]

where the parameters are interpreted as in Equation 9. The FC-DCM is substantially flexible such that researchers can customize their models in practice to allow a statement to load on more than two attributes simultaneously or to extend the proposed FC-DCM with a conjunctive or disjunctive approach. In some cases, it may happen that some statements follow the conjunctive function and others follow the disjunctive function; for example, some personality types may be compensatory, while other types belong to the noncompensation category based on substantive theory, which leads to the formulation of a hybrid FC-DCM. When a hybrid FC-DCM applies to such data, model fit evaluation can be implemented to determine the best-fitting model according to fit statistics. Various fit statistics and methods have been
developed in the literature (e.g., Chen et al., 2013; de la Torre, 2011). The way that these model-data fit techniques work in the hybrid FC-DCM warrants further investigation that may go beyond the scope of this study.

Issues of identifiability for the newly proposed FC-DCM address whether the model parameters can be consistently estimated under a fitted DCM and its corresponding Q-matrix. Several conditions for model identifiability in a variety of DCMs have been developed and elaborated (e.g., Chen et al., 2015; Fang et al., 2019; Xu & Zhang, 2016). As noted by Fang et al. (2019), however, the identifiability conditions are substantially difficult to derive from unified theorems because model structures and functional forms depend largely on the specified Q-matrix and the selected DCM. To remain within the scope of this study and avoid a mathematically complicated derivation, we referred to previous works to seek possible identifiability results and conducted simulations to evaluate the consistency of the model parameter estimates. The relevant literature recommended that both conditions of a complete Q-matrix (i.e., a $K$-dimensional identity matrix, where $K$ is the number of attributes) and at least three items for any given attribute fulfill the identifiability result (Fang et al., 2019; Hartz, 2002; Ravand & Baghaei, 2020; Xu & Zhang, 2016), which may apply to the proposed FC-DCM. Because a pairwise-comparison item always measures at least two attributes, the condition of having a complete Q-matrix is not accessible; thus, the second condition is considered to construct an appropriate Q-matrix in the FC-DCM. FC-DCM identifiability is evaluated by a series of systematic simulations, which are described in the following section.

Method

Simulation Design

Two simulation studies, namely, one involving between-statement multidimensionality (i.e., each statement measures one attribute) and the other involving within-statement multidimensionality (i.e., each statement measures two attributes), were conducted to evaluate the parameter recovery in the proposed FC-DCM using a Bayesian estimation under a variety of manipulated conditions. In both studies, the FC-DCM for the pairwise-comparison items was used to generate the responses, and three common factors were manipulated as the (a) numbers of attributes of 5 and 10, (b) test lengths of 30 and 60, and (c) sample sizes of 500 and 1,000. The particular factors manipulated for the two respective studies are as follows. In addition to the above three factors, in the first simulation study, an attribute was designed to have 5 or 10 statements, and each statement was allowed to be presented repeatedly in the FC items. The two statements in an FC item were assembled for all replicated data by randomly selecting two statements that measure two different attributes among 25 (5 attributes $\times$ 5 statements), 50 (5 attributes $\times$ 10 statements and 10 attributes $\times$ 5 statements), and 100 (10 attributes $\times$ 10 statements) statements. The second simulation study (i.e., within-statement multidimensionality) used the conjunctive FC-DCM and additionally designed two conditions where the overlap-attribute condition
allowed the two statements in an FC item to measure a common attribute and the specific-attribute condition caused each of the statements in an FC item to measure distinct attributes. The FC items were generated by randomly assembling two statements for each replicated datum. Each statement was allowed to be shown once in a test because each statement loaded on two attributes simultaneously.

The statement utility parameters (e.g., $\eta^*_iA$ and $\eta^*_iB$) were generated uniformly between .15 and .25, and the intercept parameters (i.e., $\eta_i0$) were all set to .1 to ensure that the probability of selecting statement $A_i$ over $B_i$ was extremely low when a respondent had a reversed required attribute mastery pattern and relatively high when a respondent had the required attribute mastery pattern. The second-order latent trait parameters were assumed to follow a standard normal distribution, and the mastery status of the first-order attributes was determined by Equation 4. The location parameters (i.e., $\delta_i$) were set to $-2$, $-1$, 0, 1, and 2 with equally spaced values of 1 for the 5-attribute condition and set to values between $-2$ and 2 with equally spaced values of .4 for the 10-attribute condition. The discrimination parameters (i.e., $d_i$) were generated following a lognormal distribution with a mean of zero and a variance of .06.

Designs similar to the FC items and diagnostic classification assessment conditions in our simulations can be observed in previous studies (e.g., Chen et al., 2020; H. Liu et al., 2013; Templin & Henson, 2006), and the specifications of the generated parameters in the simulations are similar to those documented in the literature (e.g., Hsu & Wang, 2015; Huang, 2020; Junker & Sijtsma, 2001). For both studies, each condition was replicated 30 times because the sampling variation was observed to be rather trivial if more than 30 replications were conducted.

In the FC-DCM, as shown in Equations 5 to 8, we assume that the response probability will equal .5 when no predominant attribute is unidentified. This constraint is set based on reasonable expectations of test takers’ response behaviors. However, this may not be the case in practice because response randomness may be observed due to the wording of the item stems or other item-related characteristics. Following the first simulation study, a supplementary simulation was conducted to allow the choice probability conditional on having an identical mastery status between the two statements (denoted by $\eta^*_i$) to be estimated freely in which the .5 probability for each item was replaced by the values generated from a uniform distribution between .4 and .6. In addition, the quality of the item parameters in DCMs has been found to have impacts on attribute classification accuracy (e.g., de la Torre et al., 2010; Huang, 2018). Therefore, we added another supplementary simulation to evaluate the effects of the low-quality statement utility parameters on model parameter estimation in which the statement utility parameters were generated uniformly between .05 and .15 and the intercept parameters were all set to .4 compared with the high statement quality used above. The two supplementary simulations were conducted in an ideal estimation condition in which the responses of 1,000 persons to 60 items with five statements that measure the five attributes were generated according to the FC-DCM. The results were compared with those of the first simulation study under the same manipulation conditions to provide a systematic comparison.
In theory, when the number of attributes specified is larger, the test can provide more information, and the analysis will be more computationally burdensome (Embretson & Yang, 2013; Ravand & Baghaei, 2020). In real testing situations, a moderate or relatively high number of attributes has often been considered in the literature (e.g., Lee et al., 2011; H. Liu et al., 2013), and a rule of thumb indicated by de la Torre and Minchen (2014) recommends a maximum of 10 attributes. This recommendation is the major reason that the 10-attribute condition was considered in our study. As the number of attributes increases, however, more persons and more items are needed and expected to render higher attribute reliability and more accurate parameter estimation. Thus, we chose the 10-attribute condition for the evaluation of model identifiability in the FC-DCM. The exemplary Q-matrix with 10 attributes and 30 items, which is shown in Appendix A, was illustrated to generate item responses via simulation by increasing the sample sizes from 50 to 5,000 with equally spaced increments of 50; as a result, 100 person groups were used to calibrate the simulated data. The Q-matrix specification satisfies the condition that each attribute is required by at least three items and simultaneously structures attributes as either predominant or inferior across different pairwise-comparison items. If the estimated parameters converge to the generated value as the sample size increases, then this can provide numerical evidence of FC-DCM identifiability (Xu & Zhang, 2016).

Analysis. All the parameters in the FC-DCM were calibrated using Just Another Gibbs Sampler (JAGS) (Plummer, 2003) with Bayesian methods. Prior distributions for each model parameter should be specified and then combined with a statistical model and data responses to produce the joint posterior distributions of parameters for model parameter calibration. For the item and statement utility parameters, a beta prior with both hyperparameters equal to 1 was set for both the intercept and statement utility parameters with a truncated distribution between 0 and .5. For the second-order structural parameters, a normal prior distribution with a mean of 0 and a variance of 4 was used for the location parameters, and a lognormal distribution with a mean of 0 and a variance of 1 was used for the discrimination parameters. The choices of the priors for the model parameters were similar to those in the DCM literature that use a Bayesian estimation for model parameter estimation (e.g., Huang, 2017; Huang & Wang, 2014). The JAGS codes for the first simulation study are available and are listed in Appendix B.

Some notes should be provided regarding the JAGS syntax as follows. As in DCMs modeled with a confirmatory approach, a measurement component and a structural component should be identified for the FC-DCM. First, with respect to the measurement component, the responses of the pairwise-comparison items are determined by the subject’s mastery status of the prespecified attributes and the statement utility parameters, whose formulation is similar to JAGS programming for a variety of DCMs in the literature (e.g., Zhan et al., 2019). Second, the structural component can be specified by signifying the relationship between latent attributes, and we implement an IRT model by assuming that a second-order latent continuous latent
trait accounts for the associations among latent attributes under the higher-order DCM framework (de la Torre & Douglas, 2004). Because the number of latent classes (the attribute mastery profiles) and the characteristics of each class are defined prior to the analysis, the FC-DCM is a confirmatory rather than an exploratory restricted latent class model. Therefore, the FC-DCM can be considered a variant of multiple classification latent class models under the higher-order factor-structure framework (see Templin & Henson, 2006).

Using the multivariate potential scale reduction factor (Brooks & Gelman, 1998) with three parallel chains, we discarded the first 5,000 iterations as burn-in and adopted the subsequent 10,000 iterations to produce the mean of the marginal posterior density for each model parameter as the parameter estimate. No multiple nodes or label switching were observed in the simulation and empirical analyses. The quality of the parameter estimation was evaluated by computing the bias and root mean square error (RMSE) for the model structural parameters and by computing the RMSE for the second-order latent trait parameter. In addition, the correct classification rate for each individual attribute (CCR-A) and attribute pattern (CCR-P) was computed to assess the recovery of the respondent’s attribute mastery status on a single attribute and the overall profile level. It is expected that the precision of parameter estimation will improve as the number of sample sizes, test lengths, attributes, and statements increases and that when each statement measures two attributes simultaneously, the specific-attribute condition can provide a better parameter estimation than the overlap-attribute condition because of the purified information provided in statements with distinct attributes.

Results

To respect the space constraints, this section summarizes the results by computing the means and standard deviations of the bias and the RMSE for the structural parameter estimates to assess the quality of the parameter estimation. Tables 1 and 2 show the bias and RMSE values for the comparative statements that measure one attribute in the FC items in the 5- and 10-attribute conditions, respectively. The proposed FC-DCM provided good parameter recovery, as indicated by the rather small bias values. With respect to the second-order attribute parameters (i.e., $d_0$ and $d_1$), a smaller RMSE value was associated with a longer test length, larger sample size, and fewer attributes. In addition, the number of statements designed for each attribute had little systemic difference for the 5- and 10-attribute conditions. Regarding the first-order structural parameters (i.e., $\eta_0^1$ and $\eta_2^1$), result patterns similar to the second-order attribute parameters were observed, but they appeared to have a better parameter recovery with smaller RMSE values.

The person parameter recovery was examined by inspecting the mean CCR-A and CCR-P and the mean RMSE values of the second-order latent trait estimate across 30 replications, which are listed in Table 3. The top half of Table 3 corresponds to the results of the first simulation study in which a statement was built to load on an
| Statement size | 5 | 10 |
|----------------|---|----|
| **Sample size** | 500 | 1,000 | 500 | 1,000 |
| **Parameter** | **TL** | **Criterion** | **Bias (SD)** | **RMSE** | **Bias (SD)** | **RMSE** | **Bias (SD)** | **RMSE** | **Bias (SD)** | **RMSE** |
| **Second-order** | | | | | | | | | | | |
| $\delta_0$ | 30 | $M$ (SD) | -.032 (.316) | .518 (.155) | -.048 (.185) | .391 (.180) | -.089 (.285) | .456 (.135) | .054 (.211) | .375 (.160) |
| | 60 | $M$ (SD) | .000 (.165) | .281 (.109) | .000 (.155) | .229 (.115) | .000 (.176) | .324 (.114) | .000 (.121) | .236 (.083) |
| $\delta_1$ | 30 | $M$ (SD) | .064 (.035) | .392 (.083) | -.074 (.074) | .285 (.061) | .012 (.140) | .382 (.094) | -.060 (.082) | .275 (.050) |
| | 60 | $M$ (SD) | .015 (.086) | .329 (.118) | .053 (.089) | .210 (.046) | -.053 (.071) | .323 (.071) | -.030 (.081) | .208 (.046) |
| **First-order** | | | | | | | | | | | |
| $\eta_0$ | 30 | $M$ (SD) | .026 (.010) | .059 (.012) | .014 (.007) | .044 (.012) | .030 (.011) | .064 (.015) | .014 (.007) | .042 (.011) |
| | 60 | $M$ (SD) | .020 (.009) | .050 (.013) | .012 (.006) | .039 (.008) | .021 (.009) | .053 (.012) | .011 (.006) | .035 (.008) |
| $\eta_3$ | 30 | $M$ (SD) | -.019 (.011) | .040 (.008) | -.015 (.010) | .035 (.008) | -.048 (.035) | .058 (.027) | -.044 (.037) | .054 (.029) |
| | 60 | $M$ (SD) | -.004 (.006) | .020 (.006) | -.003 (.003) | .015 (.006) | -.40 (.041) | .047 (.035) | -.039 (.041) | .045 (.036) |

**Note.** TL = test length; RMSE = root mean square error.
Table 2. Statistical Summary of the Structural Parameter Recovery for Statements Measuring 1 Attribute From 10 Attributes

| Parameter | Statement size | Sample size | TL | Criterion | Bias | RMSE | Bias | RMSE | Bias | RMSE | Bias | RMSE |
|-----------|----------------|-------------|-----|-----------|------|------|------|------|------|------|------|------|
|           |                | 500         | 1,000 | 500       | 1,000 |       |       |       |       |       |       |       |
| Second-order |               |             |       |           |      |      |      |      |      |      |      |      |
| δ₀        | 30 M (SD)      | -.024 (.406) | .572 (.077) | -.041 (.362) | .514 (.160) | .055 (.404) | .604 (.097) | .009 (.387) | .567 (.130) |       |      |
|           | 60 M (SD)      | -.042 (.232) | .381 (.158) | .007 (.138) | .259 (.108) | -.011 (.204) | .393 (.093) | -.005 (.155) | .271 (.122) |       |      |
| δ₁        | 30 M (SD)      | -.056 (.203) | .478 (.143) | -.082 (.127) | .344 (.067) | -.025 (.164) | .498 (.147) | -.128 (.076) | .388 (.106) |       |      |
|           | 60 M (SD)      | -.059 (.120) | .303 (.041) | -.083 (.080) | .202 (.029) | -.063 (.086) | .256 (.034) | -.029 (.053) | .180 (.028) |       |      |
| First-order |               |             |       |           |      |      |      |      |      |      |      |      |
| η₀        | 30 M (SD)      | .051 (.049) | .072 (.046) | .037 (.040) | .060 (.042) | .053 (.053) | .075 (.049) | .041 (.043) | .063 (.045) |       |      |
|           | 60 M (SD)      | .032 (.033) | .056 (.036) | .020 (.020) | .043 (.028) | .031 (.031) | .056 (.035) | .021 (.022) | .044 (.030) |       |      |
| ηₛ        | 30 M (SD)      | -.039 (.021) | .055 (.018) | -.036 (.020) | .053 (.016) | -.053 (.024) | .063 (.023) | -.051 (.024) | .063 (.022) |       |      |
|           | 60 M (SD)      | -.017 (.013) | .037 (.010) | -.012 (.010) | .032 (.008) | -.035 (.020) | .053 (.017) | -.031 (.017) | .049 (.015) |       |      |

Note. TL = test length; RMSE = root mean square error.
attribute in pairwise-comparison items. With respect to CCR-A and CCR-P, as expected, a long test length resulted in higher classification accuracy, whereas the reverse effect was observed as the number of attributes increased from 5 to 10. There were few clear systematic patterns between the small and large sample sizes and between the small and large statement sizes. A long test length was also associated with a more precise second-order latent trait estimation when 10 attributes were used than when 5 attributes were used. Although the second-order latent trait parameters appeared to be recovered poorly due to their larger RMSE values, this result was not surprising because the precision of the second-order latent trait estimation largely depends on the amount of information that the attributes provide. The deteriorated estimation would be expected to be mitigated when a larger number of attributes is used to produce more test items (Hsu & Wang, 2015; Huang, 2020).

The second simulation study allowed each comparative statement to measure two attributes. The recovery of the structural parameters of the model is summarized in Tables 4 and 5 for the 5- and 10-attribute conditions, respectively. Generally, parameter recovery patterns similar to the first simulation study were observed for both attribute conditions. Except for some cases in the 10-attribute condition, the bias values were very small. As expected, for most simulation conditions, a large sample size and long test length produced smaller RMSE values and yielded a better parameter recovery. The specific-attribute condition appeared to provide a more precise estimation than the overlap-attribute condition, which indicates that the comparative

Table 3. Statistical Summary of the Person Parameter Recovery for the Simulated Data

| Attribute size | 5 | 10 |
|---------------|---|----|
| Sample size   |   |    |
| Test length   |   |    |
| Statement size |   |    |
| 5             |   |    |
| M CCR-A       | .909 | .909 | .912 | .912 | .858 | .858 | .934 | .934 |
| M CCR-P       | .739 | .911 | .752 | .914 | .295 | .623 | .299 | .621 |
| M RMSE (i)    | .924 | .927 | .917 | .934 | .934 | .806 | .934 | .798 |
| 10            |   |    |
| M CCR-A       | .909 | .948 | .911 | .946 | .856 | .930 | .860 | .935 |
| M CCR-P       | .742 | .914 | .748 | .911 | .299 | .612 | .296 | .628 |
| M RMSE (i)    | .906 | .939 | .918 | .925 | .935 | .814 | .927 | .801 |
| Statement condition |   |    |
| Overlap attribute |   |    |
| M CCR-A       | .851 | .904 | .856 | .901 | .787 | .860 | .795 | .866 |
| M CCR-P       | .585 | .752 | .599 | .746 | .167 | .373 | .179 | .395 |
| M RMSE (i)    | .933 | .906 | .939 | .903 | .959 | .879 | .951 | .859 |
| Specific attribute |   |    |
| M CCR-A       | .876 | .905 | .879 | .907 | .833 | .909 | .836 | .910 |
| M CCR-P       | .646 | .750 | .658 | .746 | .271 | .535 | .270 | .544 |
| M RMSE (i)    | .913 | .895 | .914 | .907 | .890 | .784 | .884 | .787 |

Note. CCR-A = correct classification rate for each attribute; CCR-P = correct profile classification rate for the attribute pattern; RMSE = root mean square error.
Table 4. Statistical Summary of the Structural Parameter Recovery for Statements Measuring Two Attributes From Five Attributes

| Statement condition | Overlap attribute | Specific attribute |
|---------------------|-------------------|-------------------|
|                     | Sample size       |                   |                   |
|                     | TL 500 1,000      | 500 1,000         | 500 1,000         |
| Parameter           | Bias RMSE         | Bias RMSE         | Bias RMSE         |
|                     |                   |                   |                   |
| Second-order        |                   |                   |                   |
| $\delta_0$          | 30 $M (SD)$       | -.158 (.389) .573 (.137) | -.086 (.383) .502 (.197) | -.093 (.342) .529 (.154) | .000 (.242) .356 (.101) |
|                     | 60 $M (SD)$       | -.034 (.387) .566 (.130) | .014 (.351) .490 (.208) | .000 (.530) .545 (.269) | .000 (.383) .508 (.208) |
| $\delta_1$          | 30 $M (SD)$       | .004 (.364) .559 (.168) | .039 (.215) .516 (.244) | .062 (.294) .563 (.227) | -.008 (.266) .449 (.180) |
|                     | 60 $M (SD)$       | .017 (.244) .511 (.163) | -.027 (.070) .431 (.169) | -.106 (.174) .511 (.334) | -.110 (.162) .365 (.262) |
| First-order         |                   |                   |                   |
| $\eta_0$            | 30 $M (SD)$       | .065 (.011) .098 (.010) | .048 (.010) .080 (.009) | .031 (.009) .062 (.010) | .019 (.006) .048 (.009) |
|                     | 60 $M (SD)$       | .053 (.015) .086 (.015) | .031 (.009) .063 (.012) | .019 (.009) .050 (.010) | .013 (.007) .039 (.008) |
| $\eta_0^*$          | 30 $M (SD)$       | -.042 (.024) .054 (.019) | -.033 (.022) .048 (.016) | -.024 (.021) .038 (.013) | -.015 (.020) .031 (.011) |
|                     | 60 $M (SD)$       | -.034 (.021) .047 (.017) | -.024 (.020) .039 (.013) | -.018 (.019) .032 (.011) | -.012 (.019) .027 (.010) |

Note. TL = test length; RMSE = root mean square error.
Table 5. Statistical Summary of the Structural Parameter Recovery for Statements Measuring 2 Attributes From 10 Attributes

| Statement condition | Overlap attribute | Specific attribute |
|---------------------|-------------------|-------------------|
| Sample size         | 500               | 1,000             | 500               | 1,000             |
| Parameter           | TL  | Criterion | Bias      | RMSE     | Bias      | RMSE     | Bias      | RMSE     |
|                     |     |           |           |          |           |          |           |          |
|                     | 30  | M (SD)    | -.343 (.357) | .597 (.201) | -.203 (.354) | .562 (.172) | -.169 (.293) | .500 (.139) | -.067 (.245) | .465 (.123) |
|                     | 60  | M (SD)    | -.212 (.432) | .618 (.209) | -.135 (.298) | .481 (.172) | -.070 (.244) | .439 (.143) | -.006 (.164) | .318 (.127) |
|                     | 30  | M (SD)    | .124 (.330)  | .704 (.237) | .062 (.375)  | .686 (.540) | .077 (.297)  | .617 (.337) | -.037 (.158) | .424 (.128) |
|                     | 60  | M (SD)    | -.134 (.175) | .471 (.099) | -.134 (.069) | .286 (.033) | -.037 (.195) | .415 (.196) | -.037 (.090) | .231 (.030) |
| Second-order        |     |           |           |          |           |          |           |          |
|                     |     |           |           |          |           |          |           |          |
|                     |     |           |           |          |           |          |           |          |
| First-order         |     |           |           |          |           |          |           |          |
|                     |     |           |           |          |           |          |           |          |
| Note. TL = test length; RMSE = root mean square error.
statements that measure distinct attributes can result in purified diagnostic information between comparison attributes and are recommended for use when pairwise-comparison items are characterized by within-statement multidimensionality. Furthermore, the 10-attribute condition was more vulnerable than the 5-attribute condition to the use of overlap-attribute statements for the location and discrimination parameters because the bias values were slightly larger in the 10-attribute condition with overlap-attribute statements, as shown in Table 5.

The bottom half of Table 3 shows the mean CCR-A and CCR-P for evaluating the attribute mastery recovery and the mean RMSE values for assessing the second-order latent trait recovery in the second simulation study. For most of the manipulated conditions, the mean percentages of the correct attribute classifications increased as the test length increased from 30 to 60 items and when specific-attribute statements were used, whereas the mean correct classification rates decreased as the number of attributes increased from 5 to 10. With respect to the \( \theta \) parameter, in general, both a long test length and the use of specific-attribute statements were associated with more precise latent trait estimates. It was also found that the RMSE values uniformly decreased as the number of attributes increased to 10 when two comparative statements were designed to measure distinct attributes. Finally, the sample size appeared to have a minimal effect on both attribute classification and latent trait estimation.

The results of the two supplementary simulations are summarized in Table 6 and provide a comparison to the first simulation study in which the same manipulated conditions were considered (i.e., 1,000 persons’ responses to 60 items with five statements measuring five attributes). When the \( \eta_i^* \) parameter was not constrained to

| Condition            | Unconstrained | Low quality |
|----------------------|---------------|-------------|
| Parameter            | Bias (SD)     | RMSE (SD)   | Bias (SD) | RMSE (SD) |
| Second-order         |               |             |           |           |
| \( \delta_0 \)       | \(.000 (.141)\) | \(.242 (.119)\) | \(.000 (.201)\) | \(.263 (.119)\) |
| \( \delta_1 \)       | \(.009 (.066)\) | \(.188 (.038)\) | \(.722 (.206)\) | \(1.005 (.332)\) |
| First-order          |               |             |           |           |
| \( \eta_{i0} \)      | \(.009 (.012)\) | \(.034 (.015)\) | \(-.041 (.021)\) | \(.076 (.023)\) |
| \( \eta_{i1} \)      | \(-.005 (.007)\) | \(.019 (.008)\) | \(.004 (.008)\) | \(.030 (.008)\) |
| Criterion            | Person parameter recovery | | |
| \( M \) CCR-A        | \(.944\)      | \(.805\)    |           |           |
| \( M \) CCR-P        | \(.918\)      | \(.381\)    |           |           |
| \( M \) RMSE (\( \theta \)) | \(.954\) | \(.982\)    |           |           |

Note. CCR-A = correct classification rate for each attribute; CCR-P = correct profile classification rate for the attribute pattern; RMSE = root mean square error; for both simulation conditions, the responses of 1,000 persons to 60 items with five statements measuring five attributes were generated.
equal .5 and the probability randomness was considered conditional on the identical mastery status of the two statements, as shown on the left-hand side of Table 6, it was clearly found that the parameter recovery was not compromised and was comparable with that of the first simulation study in which the .5 probability was constrained. With respect to the low-quality statement parameter condition, however, the right-hand side of Table 6 shows that the precision of the parameter estimation deteriorated for both the model and person parameters. In particular, the correct attribute classification rates declined substantially compared with the high-quality statement parameters used in the first simulation study, which is consistent with the findings in previous studies (de la Torre et al., 2010; Huang, 2018).

Finally, the consistency of the estimator with respect to the statement and population parameters (which are indicative of the first- and second-order parameters, respectively) was evaluated in a set of simulations with different sample sizes. If the parameter estimates recover to their generated values as the sample size increases, then this can provide numerical evidence for FC-DCM identifiability (Xu & Zhang, 2016). As shown in Figure 2, all estimation curves present similar consistency patterns, and the mean RMSEs across the estimates for each of the four types of parameters (i.e., $\delta_0$, $\delta_1$, $\eta_{00}$, and $\eta_{11}$ from subplots A to D, respectively, in Figure 2) appear to approach zero in a substantial number of respondents (i.e., 5,000 persons). Because the estimation precision of the second-order population parameters depends not only on the sample size, but also on the test length (Hsu & Wang, 2015; Huang, 2017), it is not surprising that the $\delta_0$ and $\delta_1$ parameters did not show estimated consistency patterns as well as the $\eta_{00}$ and $\eta_{11}$ parameters. Accordingly, the example result illustrates that even in the most extreme form (i.e., 10 attributes and 30 items), the parameters in the FC-DCM can be estimated consistently and may be identifiable when the Q-matrix is constructed appropriately to specify that each attribute is predominant in three items and inferior in another three items. Similar identifiability results for other conditions (e.g., 10 attributes and 60 items) can be expected because more items are added to load on the attributes and to provide more estimation information, although they are not presented here due to space constraints.

**Demonstrative Example**

Nine broad features that relate to the work environment were developed to measure the degree of applicants’ work motivation. Respondents were asked to rank the nine features according to their judgment of the importance of the feature in an ideal job environment (Yang et al., 2010). The ranking data were collected based on 1,080 respondents. Maydeu-Olivares and Brown (2010) used a TIRT model to fit the transformed binary data with 36 paired comparison items and found that the TIRT provided acceptable model-data fit when an underlying latent trait (i.e., work motivation) was assumed to govern these work features. In this demonstrative analysis, nine job features, (1) a supportive environment, (2) challenging work, (3) career progression, (4) ethics, (5) personal impact, (6) personal development, (7) social
Figure 2. Changes in the Values of the Mean RMSE as the Sample Size Is Increased. (A) Attribute Location Parameter, (B) Attribute Discrimination Parameter, (C) Item Intercept Parameter, (D) Statement Utility Parameter. Note. RMSE = root mean square error.
interaction, (8) competition, and (9) work security, are treated as measuring distinct attributes, and the proposed FC-DCM is used to fit the FC data with pairwise-comparison items.

Note that the work-motivation assessment data were collected under a traditional test theory framework that aims to create a single score for a large content domain, and they were analyzed with the TIRT model to achieve the measurement goals (Maydeu-Olivares & Brown, 2010) rather than being used for the purposes of diagnostic classification. In such situations, retrofitting DCMs to data responses obtained from the assessment that are not initially developed for diagnostic purposes serves as a plausible approach to obtain diagnostic information and has been used widely in applied settings (e.g., de la Torre, 2011; Huang, 2017; Lee et al., 2011; Templin & Henson, 2006). Although the analysis results based on the TIRT appear to show that a unidimensional latent trait underlies the item responses, evidence from the cumulative literature has indicated that few data sets can reflect pure unidimensionality (Thissen, 2016) and a larger construct can be theoretically broken down into a variety of subdomains and attributes (R. Liu et al., 2017). It is not uncommon for the data from a unidimensional IRT-based test to fit a DCM by identifying attributes in terms of test specifications or item characteristics in retrofitting contexts (Ravand & Baghaei, 2020). Therefore, the nine features were considered to be different diagnostic attributes, and these features were selected repeatedly to form pairwise-comparison items. Specifically, an attribute was measured multiple times, and more information was attainable through multiple comparisons across attributes. For more detailed procedures on retrofitting in DCMs, interested readers are advised to refer to the study performed by R. Liu et al. (2017).

Because one statement that represents each latent attribute was repeatedly compared in the paired comparison items and a higher-order structure is assumed to account for the relationship between attributes in the fit of the FC-DCM, four research questions arise regarding measurement invariance with respect to the discrimination and statement utility parameters, which generates four variants of the FC-DCM to be compared with respect to the model-data fit. When imposing equivalence constraints on both the discrimination parameters across attributes and the statement utility parameters across paired comparison items, a restricted version of the FC-DCM can be formulated and is denominated as the restricted FC-DCM. When only the invariance assumption of the discrimination parameters is maintained across attributes, the restricted FC-DCM extends to the FC-DCM with discrimination invariance. In contrast, the FC-DCM with statement utility invariance arises by constraining only the statement utility parameter of the same attribute to be equal across paired comparison items.

Finally, the saturated FC-DCM emerges by allowing both the discrimination and statement utility parameters to be estimated freely across attributes and comparative items and serves as the most general formulation of the FC-DCM in the fit of the data. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were computed and compared for the selection of the better-fitting model. In
addition, posterior predictive model checking was conducted to provide evidence of the model-data fit in an absolute sense by assessing the plausibility of the replicated data against the observed data during numerous iterations based on the classic item difficulty (i.e., the proportion of endorsement of each item) that has previously been used in the Bayesian scheme for evaluating the fit of DCMs (Huang, 2017).

As with the analysis procedure used in the simulation study, the results showed that the restricted FC-DCM produced a better fit to the data due to its smaller AIC and BIC values (AIC = 24,990 and BIC = 25,270) than the FC-DCM with statement utility invariance (AIC = 25,010 and BIC = 25,320), followed by the FC-DCM with discrimination invariance (AIC = 25,290 and BIC = 25,880) and the saturated FC-DCM (AIC = 25,310 and BIC = 25,930). The mean absolute difference of the proportion of item endorsement obtained between the replicated data and the observed data over the interactions across items was close to zero, which indicates that the restricted FC-DCM yielded an absolutely appropriate fit to the data.

Under the best-fitting FC-DCM, the common attribute discrimination parameter was estimated to be .291, the item intercept parameters were estimated to be between .003 and .019 ($M = .008$), and the statement utility parameters were all estimated to be approximately .249. Regarding the attribute location parameters, the parameters were estimated to be –2.525, –2.462, –2.172, –2.340, .752, –4.064, .837, 4.938, and 4.379 for the nine attributes, which suggests that the applicants who were higher on the latent continuum that represents work motivation perceived the importance of the “competition” attribute in the work environment, whereas those lower on the latent continuum treated the “personal development” attribute as a significant feature in their ideal job. Corresponding to the higher-order structural parameter estimation, it was found that the proportions of applicants who judged the nine attributes as important job characteristics were .674, .670, .651, .662, .445, .764, .439, .140, and .218, which indicates that the “personal development” and “competition” attributes represented the most and least important working characteristics for applicants’ work motivation, respectively. Furthermore, the correlations between each pair of attributes were estimated to be between –.359 and .152 ($M = –.117$); this range covers negative and positive values due to the ipsative measurement property (Aitchison, 1986) and is similar to the findings of the dimension correlations that use IRT-based FC models (e.g., Bürkner et al., 2019; Chen et al., 2020).

In the TIRT analysis of the same data, Maydeu-Olivares and Brown (2010) found that the characteristics of a supportive environment, challenging work, and career progression were the three predominant indicators that relate to general work motivation and that the three features of social interaction, competition, and work security were the least strongly related to work motivation. Although there were some differences in the judgment of importance among these job features between the TIRT and FC-DCM, the observed similar result patterns for the two distinct methodologies may confirm the validity of the proposed FC-DCM when analyzing the work-motivation ranking data.
Conclusion

The merit of FC item formats for noncognitive assessments has been supported and evidenced by their ability to minimize not only uniform response biases but also extremely favorable responses compared with normative tests, as documented in the recent literature (e.g., Brown & Maydeu-Olivares, 2011; Christiansen et al., 2005, 2013; Stark et al., 2005). Most psychometric models developed for FC items are designed to calibrate respondents on latent continuous trait scales and rarely provide classification information about whether respondents have a certain type of measured variable unless a prespecified threshold is additionally imposed on a latent continuum. In contrast, the DCM facilitates the goal of obtaining diagnostic information regarding classification on multiple attributes but is still susceptible to intentional response biases through improved scores due to the use of single-stimulus items. In this study, based on Thurstone’s law of comparative judgment (Thurstone, 1927), we develop a new class of DCM for pairwise-comparison items (i.e., the FC-DCM) to combine the advantages of FC formats and the DCM by simultaneously controlling for motivated response distortion and providing both a classification of latent binary attributes and an overall evaluation of the latent continuous trait under the higher-order DCM framework. The proposed FC-DCM is flexible for application to practical settings such that a statement that represents one or more attribute(s) can be repeatedly selected for judgments on the pairwise-comparison items, between- and within-statement multidimensionality can be created where appropriate, and conjunctive or disjunctive versions of the FC-DCM can be developed and empirically examined based on a diversity of assumptions.

A series of simulation studies were conducted to assess the parameter recovery and the efficiency of the developed FC-DCM under a variety of manipulated conditions using Bayesian estimation. When each statement was allowed to measure an attribute in the first simulation study, it was observed that the improved model structural parameters were closely associated with a large number of respondents and test items and that the first-order structural parameters (i.e., the statement utility parameters) appeared to be recovered better than the second-order structural parameters (i.e., the location and discrimination parameters). Increasing the number of attributes from 5 to 10 decreases the precision of the model’s structural parameter estimation, whereas increasing the number of statements from 5 to 10 appears to have a less clear and systematic impact on the model’s structural parameter estimation. The major factors that influence the precision of the second-order latent trait estimation and the accuracy of the attribute classification are the number of test items and the measured attributes, with a longer test corresponding to better latent trait and attribute estimations and more attributes resulting in lower attribute classification accuracy but improving the latent trait estimation.

The second simulation study allowed each statement to load on two attributes and manipulated whether the two statements in a comparative item measured a common attribute (i.e., the overlap- and specific-attribute conditions). As in the first simulation, the results show that in most conditions, improved model structural parameters are attainable when a long test length and a large sample size are used and that the
first-order model parameters have a better parameter recovery than the second-order model parameters. In addition, the specific-attribute condition produces a better model parameter estimation than the overlap-attribute condition. The effect of the attribute size on the model parameter recovery is the same in the first and second simulation studies. For the person parameter recovery, similar to the first simulation study, both the test length and the number of attributes have a substantial impact on the latent trait and attribute estimation. Unsurprisingly, the specific-attribute condition provides a better parameter estimation than the overlap-attribute condition regardless of the item or person parameters; previous studies have indicated that a simple-structured Q-matrix that arranges each item to measure only one attribute can provide more differentiated information between two cognitive profiles than a complex-structured Q-matrix that arranges each item to measure more than two attributes (e.g., Huang, 2018; C. Wang, 2013). Accordingly, the FC-DCM developed for pairwise-comparison items can provide satisfactory model parameter recovery and accurate person attribute classification when the comparative statements are structured as measuring either a single attribute in a simple-structured Q-matrix or distinct attributes in a complex-structured Q-matrix.

The demonstration example was selected from a ranking task that recruited job applicants to order nine work environment features in terms of their importance in an ideal job, with the outcome measures used as an indicator of general work motivation. The ranking data were transformed into 36 paired comparison items. Maydeu-Olivares and Brown (2010) used the TIRT model to fit comparative data and found that the TIRT model that assumes a common underlying trait provided an acceptable model-data fit. In our analysis, the nine broad features were treated as latent attributes governed by a common second-order latent trait, and the same comparative data were fit to the FC-DCM. Four variants of the FC-DCM were proposed in terms of different assumptions of model parameter equivalence and compared with respect to the goodness of fit to the data. In the best-fitting model, the results revealed that job applicants tend to judge the “personal development” attribute as the most important feature in the work environment and the “competition” attribute as the least important feature. The ordering of the proportion of each attribute by the applicants in the FC-DCM appears to be similar to the ordering of the strength of the relationship between each feature and the common trait in the TIRT analysis. The main difference between the FC-DCM and TIRT analyses is that the FC-DCM can both classify respondents into different categories for each attribute with diagnostic information and provide an overall evaluation on a latent continuous scale with summative assessment purposes, while the TIRT merely satisfies the needs of the latter case. Note that the empirical analysis included one statement for each measured attribute. Although the number of statements used for the pairwise-comparison items appeared to be less influential in parameter recovery in the simulation studies, the results should be interpreted with caution, and more statements are recommended to increase the diversity of attribute measures.

To model the dependencies between attributes, we use the higher-order DCM and assume that a general latent trait determines the probability of mastering each of the
attributes for respondents when local independence is satisfied for attribute mastery conditional on the more broadly defined latent trait (de la Torre & Douglas, 2004). In addition to the major advantages of greatly reducing the estimation complexity compared with the saturated model and simultaneously obtaining an overall ability assessment and discrete attribute classification, the higher-order variable is analogous to the overall unidimensional IRT score and may be promising for retrofitting DCMs to IRT-based data, as is the case with our empirical demonstration (R. Liu et al., 2017). If some type of attribute hierarchy, such as psychologically sequential attribute mastery, is observed (see Leighton & Gierl, 2007), then a reduced saturated model may provide an alternative approach to modeling the relationships of attributes by constraining zero probability for the impermissible attribute profiles. Other methods, such as log-linear parameterization (Rupp & Templin, 2008) and tetrachoric parameterization (Hartz, 2002), are also available and can be directly applied to the new FC-DCM to account for attribute associations.

Directions for further development and the extension of the FC-DCM are provided for future research. First, although the FC-DCM together with a specified Q-matrix provides consistent parameter estimation in our study, further work on identification and estimation issues based on solid mathematical foundations is necessary to indicate the diverse conditions under which the model parameters are identifiable from the data. Second, dichotomously discrete latent variables are used as indicators of whether respondents possess a specific attribute. These are most commonly observed in the DCM literature and are implemented in our developed FC-DCM. In some cases, polytomous attributes that allow individuals to be divided into multiple categories may provide additional diagnostic information (Chen & de la Torre, 2013; Sun et al., 2013) and can be extended to the FC-DCM. Third, the FC formats were limited to comparative items in this study, and paired comparison outcomes were obtained by transforming the ranking data. However, not all paired comparison items can be converted into rankings due to a possible intransitive pattern (Maydeu-Olivares & Brown, 2010). In addition, FC models for ranking data have recently been proposed and provide a more efficient estimation than comparative data (Joo et al., 2018; W.-C. Wang et al., 2016). Establishing an FC-DCM for ranking data would be an encouraging and valuable area for future research. Fourth, from the perspective of testing efficacy, it is desirable to apply the FC-DCM to CAT by combining the advantages of FC formats, DCMs, and CAT. As observed in the simulation studies, the second-order latent trait is not estimated satisfactorily because a limited number of attributes are commonly used in DCMs such that the latent continuous trait cannot be precisely estimated (Hsu & Wang, 2015). With the popularity of computer technology, the response times of respondents can be routinely recorded and used as collateral information to facilitate the precision of person and item measures in higher-order DCMs (Huang, 2020; Zhan et al., 2018). How to implement response times to improve the parameter estimation in the FC-DCM and the effectiveness of incorporating additional information in the estimation process would be interesting topics for future research.
Appendix A

Table A1. Q-Matrix for 30 Items Measuring 10 Attributes, With Each Attribute Having Five Statements

| Item | attribute | | | | | | | | |
|------|-----------|---|---|---|---|---|---|---|---|
| 1    | 1-1       | 2-2 | | | | | | | |
| 2    | 1-1       | 2-2 | | | | | | | |
| 3    | 2-2       | 1-1 | | | | | | | |
| 4    | 2-2       | 1-1 | | | | | | | |
| 5    | 2-2       | 1-1 | | | | | | | |
| 6    | 1-1       | 2-2 | | | | | | | |
| 7    | 1-1       | 2-2 | | | | | | | |
| 8    | 2-2       | 1-1 | | | | | | | |
| 9    | 2-2       | 1-1 | | | | | | | |
| 10   | 1-3       | 2-4 | | | | | | | |
| 11   | 1-3       | 2-4 | | | | | | | |
| 12   | 1-3       | 2-4 | | | | | | | |
| 13   | 1-3       | 2-4 | | | | | | | |
| 14   | 2-4       | 1-3 | | | | | | | |
| 15   | 2-4       | 1-3 | | | | | | | |
| 16   | 2-4       | 1-3 | | | | | | | |
| 17   | 2-4       | 1-3 | | | | | | | |
| 18   | 2-4       | 1-3 | | | | | | | |
| 19   | 2-4       | 1-3 | | | | | | | |
| 20   | 2-4       | 1-3 | | | | | | | |
| 21   | 2-4       | 1-3 | | | | | | | |
| 22   | 2-4       | 1-3 | | | | | | | |
| 23   | 2-4       | 1-3 | | | | | | | |
| 24   | 2-4       | 1-3 | | | | | | | |
| 25   | 2-4       | 1-3 | | | | | | | |
| 26   | 2-4       | 1-3 | | | | | | | |
| 27   | 2-4       | 1-3 | | | | | | | |
| 28   | 2-4       | 1-3 | | | | | | | |
| 29   | 2-4       | 1-3 | | | | | | | |
| 30   | 2-4       | 1-3 | | | | | | | |

Note. The number listed on the left-hand side of the dashed sign specifies the attribute scheme, where 1 indicates a predominant attribute and 2 indicates an inferior attribute in a pairwise-comparison item. The number listed on the right-hand side of the dashed sign specifies the statement used for each attribute.
Appendix B
JAGS Codes for the FC-DCM in the First Simulation Study

# N is the number of persons
# T is the number of FC items
# Z is the number of attributes
# S is the number of statements for an attribute
# alpha indicates the respondent's attribute mastery status
# d0 is the intercept parameter
# sd is the statement utility parameter
# q_att identifies which attribute is measured by an FC item
# q_sta identifies which statement is administered for a given attribute
# r is the item response
# a and b are the discrimination and location parameters, respectively
# theta is the second-order latent trait
model{
  for (i in 1:N) {
    for (j in 1:T) {
      p[i,j] <- d0[j] + (0.5–d0[j])*step(alpha[i,q_att[j,1]]-alpha[i,q_att[j,2]])
        + (sd[q_sta[j,1],q_att[j,1]] + sd[q_sta[j,2],q_att[j,2]]*(1–alpha[i,q_att[j,2]]))
        *step(alpha[i,q_att[j,1]]–alpha[i,q_att[j,2]]–1)
      r[i,j] ~ dbern(p[i,j])
    }
    for (k in 1:Z) {
      pt[i,k] <- exp(a[k] *(theta[i]–b[k])/(1 + exp(a[k]*(theta[i]–b[k])))
      alpha[i,k] ~ dbern(pt[i,k])
    }
    theta[i] ~ dnorm(0, 1)
  }
  # priors
  for (j in 1:T) {
    d0[j] ~ dbeta(1,1)(0.5)
  }
  for (j in 1:S) {
    for (k in 1:Z) {
      sd[j,k] ~ dbeta(1,1)(0.5)
    }
  }
  for (k in 1:Z) {
    a[k] ~ dlnorm(0,1)
    b[k] ~ dnorm(0,0.25)
  }
  # data
  "N" = 500; "T" = 30; "Z" = 5; "S" = 5;
  q_att = structure(Data = c(1, 2, 2, 3, ....

(continued)
Declaration of Conflicting Interests

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

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the Ministry of Science and Technology (No. 109-2410-H-845-015-MY3).

ORCID iD

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

References

Aitchison, J. (1986). The statistical analysis of compositional data. Chapman and Hall.

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. (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

Borislow, B. (1958). The Edwards Personal Preference Schedule (EPPS) and fakability. Journal of Applied Psychology, 42(1), 22–27. https://doi.org/10.1037/h0044403

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. (2016). Item response models for forced-choice questionnaires: A common framework. Psychometrika, 81(1), 135–160. https://doi.org/10.1007/s11336-014-9434-9

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

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

Note. FC-DCM = forced-choice diagnostic classification model.
Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827–854. https://doi.org/10.1177/0013164419832063

Chen, C.-W., Wang, W.-C., Chiu, M.-M., & Ro, S. (2020). Item selection and exposure control methods for computerized adaptive testing with multidimensional ranking items. *Journal of Educational Measurement, 57*(2), 343–369. https://doi.org/10.1111/jedm.12252

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement, 37*(6), 419–437. https://doi.org/10.1177/0146621613479818

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*(2), 123–140. https://doi.org/10.1111/j.1745-3984.2012.00185.x

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association, 110*, 850–866. https://doi.org/10.1080/01621459.2014.934827

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267–307. https://doi.org/10.1207/s15327043hup1803_4

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville and Willson (1992). *Journal of Occupational and Organizational Psychology, 67*(2), 89–100. https://doi.org/10.1111/j.2044-8325.1994.tb00553.x

Culpepper, S. A., & Balamuta, J. J. (In Press). Inferring latent structure in polytomous data with a higher-order diagnostic model. *Multivariate Behavioral Research*. Advance online publication. https://doi.org/10.1080/00273171.2021.1985949

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the dina model. *Journal of Educational Measurement, 47*(2), 227–249. https://doi.org/10.1111/j.1745-3984.2010.00110.x

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicologia Educativa, 20*(2), 89–97. https://doi.org/10.1016/j.pse.2014.11.001

de la Torre, J., Ponsoda, V., Leenen, I., & Hontangas, P. (2012, April). *Examining the viability of recent models for forced-choice data*. Paper presented at the Meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

DeVito, A. J. (1985). Review of the Myers-Briggs type indicator. *Ninth Mental Measurements Yearbook*, 2, 1030–1032.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). As assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81–106. https://doi.org/10.1207/S15327043HUP1601_4

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*(1), 14–36. https://doi.org/10.1007/s11336-012-9296-y

Fals-Stewart, W., Bircher, G. R., Schafer, J., & Lucente, S. (1994). The personality of marital distress: An empirical typology. *Journal of Personality Assessment, 62*(2), 223–241. https://doi.org/10.1207/s15327752jpa6202_5
Fang, G., Liu, J., & Ying, Z. (2019). On the identifiability of diagnostic classification models. *Psychometrika, 84*(1), 19–40. https://doi.org/10.1007/s11336-018-09658-x

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* [Doctoral Dissertation]. University of Illinois.

Hausknecht, J. P. (2010). Candidate persistence and personality test practice effects: Implications for staffing system management. *Personnel Psychology, 63*(2), 299–324. https://doi.org/10.1111/j.1744-6570.2010.01171.x

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

Hsu, C.-L., & Wang, W.-C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement, 52*(2), 125–143. https://doi.org/10.1111/jedm.12069

Huang, H.-Y. (2017). Multilevel cognitive diagnosis models for assessing changes in latent attributes. *Journal of Educational Measurement, 54*(4), 440–480. https://doi.org/10.1111/jedm.12156

Huang, H.-Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification, 35*, 437–465. https://doi.org/10.1007/s00357-018-9265-y

Huang, H.-Y. (2020). Utilizing response times in cognitive diagnostic computerized adaptive testing under the higher-order deterministic input, noisy “and” gate model. *British Journal of Mathematical and Statistical Psychology, 73*(1), 109–141. https://doi.org/10.1111/bmsp.12160

Huang, H.-Y., & Wang, W.-C. (2014). The random-effect DINA model. *Journal of Educational Measurement, 51*(1), 75–97. https://doi.org/10.1111/jedm.12035

Joo, S. H., Lee, P., & Stark, S. (2020). Adaptive testing with the GGUM-RANK multidimensional forced choice model: Comparison of pair, triplet, and tetrad scoring. *Behavior Research Methods, 52*(2), 761–772. https://doi.org/10.3758/s13428-019-01274-6

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

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272. https://doi.org/10.1177/0146621012032064

Kuder, G. F., & Diamond, E. E. (1979). *Kuder occupational interest survey: General manual* (2nd ed.). Science Research Associates.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144–177. https://doi.org/10.1080/15305058.2010.534571

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Liu, H., You, X., Wang, W., Ding, S., & Chang, H.-H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification, 30*(2), 152–172. https://doi.org/10.1007/s00357-013-9128-5
Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), 357–383. https://doi.org/10.1177/0013164416685599

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

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*(2), 169–182. https://doi.org/10.1111/1468-2389.00057

Maydeu-Olivares, A. (2005). Linear item response theory, nonlinear item response theory, and factor analysis: A unified framework. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 73–100). Lawrence Erlbaum Associates.

Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*(6), 935–974. https://doi.org/10.1080/00273171.2010.531231

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using Ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*(4), 531–552. https://doi.org/10.1348/0963179042596504

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/10.1177/0146621616662226

Myers, I. B., Mary, H. M., Naomi, Q., & Allan, H. (1998). *MBTI handbook: A guide to the development and use of the Myers-Briggs type indicator consulting* (3rd ed.). Psychologists Press.

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

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment, 34*(8), 782–799. https://doi.org/10.1177/0734282915623053

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing, 20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Revuealta, J., Halty, L., & Ximénez, C. (2018). Validation of a questionnaire for personality profiling using cognitive diagnostic modeling. *The Spanish Journal of Psychology, 21*, Article E63. https://doi.org/10.1017/sjp.2018.62

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment testing and hiring decisions. *Journal of Applied Psychology, 83*(4), 634–644. https://doi.org/10.1037/0021-9010.83.4.634

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives, 6*(4), 219–262. https://doi.org/10.1080/1536360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

SHL. (2013). *OPQ32r technical manual* (version 1.0). SHL Group.
Solomon, A., Haaga, D. A. F., & Arnow, B. A. (2001). Is clinical depression distinct from subthreshold depressive symptoms? A review of the continuity issue in depression research. *Journal of Nervous and Mental Disease, 189*(8), 498–506. https://doi.org/10.1097/00005053-200108000-00002

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multidimensional pairwise-preference model. *Applied Psychological Measurement, 29*(3), 184–203. https://doi.org/10.1177/0146621604273988

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*(3), 463–487. https://doi.org/10.1177/1094428111444611

Sun, J., Xin, T., Zhang, S., & de la Torre, J. (2013). A polytomous extension of the generalized distance discriminating method. *Applied Psychological Measurement, 37*(7), 503–521. https://doi.org/10.1177/0146621613487254

Super, D. E. (1942). The Bernreuter Personality Inventory: A review of research. *Psychological Bulletin, 39*(2), 94–125. https://doi.org/10.1037/h0058418

Templin, J. L., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275. https://doi.org/10.1007/s00357-013-9129-4

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Thissen, D. (2016). Bad questions: An essay involving item response theory. *Journal of Educational and Behavioral Statistics, 41*(1), 81–89. https://doi.org/10.3102/1076998615621300

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286. https://doi.org/10.1037/h0070288

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement, 73*(6), 1017–1035. https://doi.org/10.1177/0013164413498256

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

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

Xu, G., & Zhang, S. (2016). Identifiability of diagnostic classification models. *Psychometrika, 81*(3), 625–649. https://doi.org/10.1007/s11336-015-9471-z

Yang, M., Inceoglu, I., & Silvester, J. (2010, January). *Exploring ways of measuring person-job fit to predict engagement*. Paper presented at the BPS Division of Occupational Psychology conference, Brighton, UK.

Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modeling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 71*(2), 262–286. https://doi.org/10.1111/bmsp.12114

Zhan, P., Jiao, H., Man, K., & Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics, 44*(4), 473–503. https://doi.org/10.3102/1076998619826040