Item Response Models for Forced-Choice Questionnaires: A Common Framework

Anna Brown

University of Kent

Author Note

Anna Brown, PhD, Lecturer in Psychological Methods and Statistics, School of Psychology, University of Kent.

Correspondence should be addressed to Anna Brown, School of Psychology, University of Kent, Canterbury, Kent CT2 7NP, United Kingdom. E-mail: A.A.Brown@kent.ac.uk

Acknowledgements

I am grateful to Alberto Maydeu-Olivares for his continuous support and helpful comments on an earlier draft of this paper.

Abstract

In forced-choice questionnaires, respondents have to make choices between two or more items presented at the same time. Several IRT models have been developed to link respondent choices to underlying psychological attributes, including the recent MUPP (Stark, Chernyshenko & Drasgow, 2005) and Thurstonian IRT (Brown & Maydeu-Olivares, 2011) models. In the present article, a common framework is proposed that describes forced-choice models along three axes: 1) the forced-choice format used; 2) the measurement model for the relationships between items and psychological attributes they measure; and 3) the decision model for choice behavior. Using the framework, fundamental properties of forced-choice measurement of individual differences are considered. It is shown that the scale origin for the attributes is generally identified in questionnaires using either unidimensional or multidimensional comparisons. Both dominance and ideal point models can be used to provide accurate forced-choice measurement; and the rules governing accurate person score estimation with these models are remarkably similar.

Keywords: forced choice, ipsative data, Thurstonian choice model, unfolding model, Bradley-Terry model, dominance model, ideal point model

Item Response Models for Forced-Choice Questionnaires: A Common Framework

The most popular way of gathering responses to personality and similar items is to ask respondents to evaluate one item at a time, independently of other items (*single-stimulus* format). An alternative way is to ask respondents to choose between several items presented at the same time, for example to indicate which statement describes them best – "I am relaxed most of the time" or "I do things according to a plan". Regardless of whether all or none of the items is evaluated favorably by a respondent, he/she will be forced to make a choice (hence the name – *forced-choice* format). With the single-stimulus formats, respondents make *absolute judgments* about every item. With the forced-choice formats, respondents engage in *comparative judgments*, assessing relative merits of the items.

Comparative judgments may be preferred to absolute judgments in many contexts. Because it is impossible to endorse all items, comparative judgments are effective whenever differentiation between responses is desired. For example, in situations where strong "halo" effects are present, or when respondents are likely to acquiesce, or use a limited range of a rating scale, forcing choice may result in more usable data (Bartram, 2007; Chan, 2003; Cheung & Chan, 2002; Maydeu-Olivares & Böckenholt, 2008). Forced-choice formats remove uniformly elevated or decreased judgments across all items, therefore eliminating rater effects such as leniency / severity (Cheung & Chan, 2002). It has even been argued that combining equally desirable items in the same block prevents respondents from endorsing the desirable items and rejecting the undesirable ones, thus reducing socially desirable responding (Christiansen, Burns & Montgomery, 2005; Jackson, Wroblewski & Ashton, 2000; Martin, Bowen & Hunt, 2001; Vasilopoulos et al., 2006). Finally yet importantly, direct comparisons between items remove the need for any response categories or rating scales, which often yield idiosyncratic interpretations and biases (e.g. Schwarz et al., 1991).

Despite these potential advantages, until recently the use of forced-choice questionnaires has been controversial because their classical scoring yielded *ipsative* data (from Latin *ipse*, or relative to self). As the name suggests, the ipsative scores are scaled in relation to the person mean, and while suitable for intra-individual assessments, they are not suitable for inter-individual comparisons. The psychometric problems of ipsative data are well described in the literature; see, for instance, Clemans (1970) for a full account of ipsative mathematics, and Meade (2004) for evidence of empirical problems when ipsative data are used for selection decisions.

These pitfalls of ipsative data were unfortunate since the main objective of psychometric questionnaires is to scale the objects of assessment (typically people) on some attributes so that they can be compared to the scale origin and to each other. In an attempt to infer proper measurement from forced-choice data, at least six Item Response Theory (IRT) models have been developed. These are the Zinnes-Griggs (1974) model, Andrich's (1989) IRT model for unfolding preference data and the Hyperbolic Cosine Model for pairwise preferences (Andrich, 1995), the Multi-Unidimensional Pairwise-Preference (MUPP) model (Stark, Chernyshenko & Drasgow, 2005), a multidimensional unfolding approach by McCloy, Heggestad and Reeve (2005), and the Thurstonian IRT model (Brown & Maydeu-Olivares, 2011). Yet, many researchers and test users are confused about the merits of these models. For example, there is much confusion around the ability to infer proper measurement from forced-choice data using different item types (see Drasgow, Chernyshenko & Stark, 2010; also Brown & Maydeu-Olivares, 2010). Unfortunately, despite a vast literature on choice models, including some excellent reviews (e.g. McFadden, 2001; Böckenholt, 2006), most works focus on models that are stimulus-centric, not person-centric. The main motivation for the former models is to establish properties of the stimuli (for example, in marketing or economics), while the latter models aim to infer measurement of individual

differences. Although relying on the same theories of choice behavior as stimulus-centric models do (for example, Thurstone's law of comparative judgment or Luce's choice axioms), person-centric applications necessitate the use of specialized forced-choice designs, item types, and measurement models. These issues have not been given enough attention in the psychometric literature, nor have they been treated in a consistent fashion so that cross-model comparisons could be made.

The present article aims to fill this gap by providing a common framework for describing models for forced-choice questionnaire data. The framework relies on the notion of a psychological value, or utility (Thurstone, 1929), elicited by every item presented to respondents. Utilities are assumed to vary between respondents, to be underlain by some psychological attributes we intend to measure, and to have a random error component. Then, the different models proposed for forced-choice data can be classified along three axes: 1) the forced-choice format used (i.e. whether the items are presented in pairs, or larger ranking blocks); 2) how the relationship between the utilities and the attributes is specified (for instance, assuming a dominance or an ideal point process); and 3) how the utilities are linked to the observed choices. Using the proposed framework, fundamental properties of forced-choice data for measurement of individual differences on psychological attributes are discussed. Specifically, the conditions for identification of the scale origin, which was the main challenge for ipsative data, are laid out. Merits of the existing IRT models for forced-choice questionnaires are discussed based on these fundamental properties and the article concludes with a discussion and suggestions for innovation.

Three Axes to Classify Forced-Choice Models

Item response models for forced-choice questionnaire data differ because they may use different: (1) experimental designs to gather data (forced-choice formats); (2)

measurement models to describe the relationship between the items and the psychological attributes they measure; and (3) decision models to explain choices between items. These are the basic axes necessary to describe any forced-choice model.

Questionnaires may use different **forced-choice formats** to gather data; for example, respondents may be asked to make a choice between two alternatives, or perform full or partial ranking on a larger set of alternatives. Because questionnaire items serve as indicators of some higher order dimensions of interest, a **measurement model** is needed to describe the relationships between the items and psychological attributes they measure. A **decision model** is needed to describe the process that respondents adopt when making choices between items. The decision model postulates the mechanism for preferring one item to the other. Options available within each of the three axes are described below.

1. Forced-Choice Formats

Forced-choice questionnaires typically consist of many individual *choice* tasks. The choice tasks within a questionnaire will be referred to as "*blocks*". Each block may involve two or more items.

The simplest forced-choice block constitutes a pair of items $\{i, k\}$, out of which respondents are asked to select one item according to some instruction (for instance, select a statement which describes own behavior most accurately). There are only two possible outcomes of such comparison – either the first item is preferred, or the second is preferred.

Thus, the response of person *j* to a pair of items {*i*, *k*} can be described using a *binary* random variable[1]:

$$y_{j\{i,k\}} = \begin{cases} 1, & \text{if item } i \text{ is preferred} \\ 0, & \text{if item } k \text{ is preferred} \end{cases}. \quad (1)$$

A forced-choice block may involve three or more items at once. Respondents may be asked to rank order *n* items (i.e. assign them ranking positions from 1 to *n*). Any given rank ordering of *n* items is fully described by $\tilde{n} = n(n-1)/2$ dummy variables, representing binary comparisons between all possible pairs of items. For example:

Ranking				Dummy variables					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
1	3	4	2	1	1	1	1	0	0

Alternatively, respondents may be asked to indicate only the most preferred item in a block, resulting in a *partial ranking*, because outcomes of some pairwise comparisons are not collected by design. Another common design is asking respondents to indicate the most and the least preferred items, for example:

Partial ranking				Dummy variables					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
most			least	1	1	1	1	.	0

[1] Here, the standard coding procedure in the Thurstonian choice literature (Maydeu-Olivares & Böckenholt, 2005) is adopted. It is important to note that at the point of coding, no assumptions are made about the underlying distributions, decision mechanisms, etc.

Another format is *ranking with ties*, also known as Q-sort (Block, 1961). Q-sorts typically involve all questionnaire items in one giant "block", in which respondents assigns the individual items to few rank-ordered groups (for more detail, see Brown & Maydeu-Olivares, in press). Consequently, the ranks of items in the same group are tied, so that information on the relative preferences within groups is not gathered by design. The dummy coding described above applies to Q-sorts. As can be seen, all types of data gathered in forced-choice tasks may be described using binary variables.

2. Measurement Models for Item-Attribute Relationships

In questionnaires, items serve as indicators of a number of psychological attributes, so that any item evaluations by respondents depend on these attributes. Evaluations of individual items (*absolute judgments*), however, are not observed in forced-choice questionnaires. Only outcomes of *comparative judgments* (e.g. preferred – not preferred) are observed. Thus, measurement models describing relationships between the items and the attributes must rely on **latent** evaluations of items. Thurstone's notion of item utility has been widely used in the literature to describe unobserved item evaluations, and it will be used here.

Thurstone proposed that any presented object elicits a psychological value, or *utility* in a respondent, which he described as "the affect that the object calls forth" (Thurstone, 1929; p. 160). Depending on the nature of assessment, the utility of a questionnaire item might represent the degree of attractiveness of a concept described by the item, likeness to the respondent's own behavior or personality, the level of agreement with an opinion etc.

People vary in their utilities for an item. This variation can be partitioned into two sources – individual differences on personal attributes that the item measures, and all other person-by-item interactions. In person-centric applications, the former is the focus of modelling, and the latter is considered a random error. Thus, the item utility can be described

using a latent variable, with the systematic part being a function of personal attributes and/or fixed characteristics of items (e.g. Andersen, 1976), and the random part representing all other item-by-person interactions. In this *random utility* model, the utility of item *i* for person *j*, t_{ji}, is the sum of the systematic part $\overline{t}_{ji}$ and the random part ε_{ji}

$$t_{ji} = \overline{t}_{ji}\left(\boldsymbol{\theta}_j\right) + \varepsilon_{ji}, \tag{2}$$

where $\overline{t}_{ji}$ is a function of psychological attributes $\boldsymbol{\theta}_j = \left(\theta_{j1}, \theta_{j2}, ..., \theta_{jd}\right)'$ of person *j* (hence the person subscript) and some fixed item properties (hence the subscript *i*). It is assumed that the random errors are independent of the person attributes and among each other.

Specific models may be postulated to describe the general relationships (2). Here we consider the most popular measurement models – linear factor analysis models (Brady, 1989; McDonald, 1999) and ideal point models (Takane, 1987; Brady, 1989). This list may be extended; for instance, one could adopt the Wandering Vector model (De Soete & Carroll, 1983) etc.

Linear Factor Analysis (LFA) models

In linear factor models, the utility of person *j* for item *i* is described as a linear function of the item mean μ_i, *d* personal attributes $\boldsymbol{\theta}_j = \left(\theta_{j1}, \theta_{j2}, ..., \theta_{jd}\right)'$ or common factors, which are weighted by factor loadings $\lambda_{i1}, \lambda_{i2}, ..., \lambda_{id}$, and the unique factor ε_{ji}

$$t_{ji} = \mu_i + \sum_{a=1}^{d} \lambda_{ia}\theta_{ja} + \varepsilon_{ji}. \tag{3}$$

Questionnaire items are most often designed to tap into one attribute only. In this case, the matrix of factor loadings Λ has only one non-zero entry in every row (has an *independent clusters basis*; McDonald, 1999).

The LFA model has proven to be a reasonably good approximation when questionnaire items represent decisively positive or negative standing on the attribute continuum. Examples of such items are "I keep my paperwork in order" and "My paperwork is always in disarray", respectively, to indicate Orderliness. As the Orderliness score increases, the utility for the first item should increase, and for the second item decrease – the utility judgment therefore reflects a *dominance* response process.

Ideal Point (IP) models

The term "*ideal point*" was coined by Coombs (1960) based on the original work of Thurstone (1928). Thurstone argued that the psychological value for a statement such as "Fire arms should not belong in private hands" is the highest for a person with this exact level of the attitude towards Militarism. For this person, statements representing either higher or lower levels of Militarism should have lower utilities. Coombs called this point of maximum utility on the attribute continuum a person's "ideal point". In ideal point models, it is assumed that people and items can be represented by points in the same attribute space, and that a person's utility for an item increases as the distance between the ideal point (person location) and the item locations decreases.

Originally suggested for attitude items, the ideal point models have been recently considered for wider use (Drasgow, Chernyshenko & Stark, 2010). For instance, the utility of personality item "My attention to detail is about average" is expected to be highest for people with average "Orderliness", and lower for people with either extremely high or low Orderliness. For such items, dominance-type models are inappropriate.

Formally, a multidimensional ideal point model describes the utility of person *j* for item *i* as the linear function of the item mean μ_i, the minus distance between the person location (ideal point) $\boldsymbol{\theta}_j = \left(\theta_{j1}, \theta_{j2}, ..., \theta_{jd}\right)'$ and the item location $\boldsymbol{\delta}_i = \left(\delta_{i1}, \delta_{i2}, ..., \delta_{id}\right)'$, and the random error ε_{ji}

$$t_{ji} = \mu_i - D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_i\right) + \varepsilon_{ji}. \tag{4}$$

Two main types of ideal point models have been proposed based on how the person-item distance is defined. Shepard (1957), Coombs (1960) and other authors defined the distance between the ideal point and the item location using Euclidean geometry:

$$t_{ji} = \mu_i - \left(\sum_{a=1}^{d}\left(\theta_{ja} - \delta_{ia}\right)^2\right)^{1/2} + \varepsilon_{ji}. \tag{5}$$

Other authors, notably Takane (1987) and Brady (1989), used the squared Euclidean distance,

$$t_{ji} = \mu_i - \frac{1}{2}\sum_{a=1}^{d}\left(\theta_{ja} - \delta_{ia}\right)^2 + \varepsilon_{ji}. \tag{6}$$

Takane (1996) argued for the use of the more algebraically tractable squared Euclidean distance form on the grounds of its (slightly) better fit to empirical data. He was also concerned about the limited range in which comparisons between one-dimensional items of the Euclidean distance form discriminate (see discussion following expression (35) further in this article).

The common feature of the basic ideal point models discussed so far is that only one item property – item location – matters for the utility judgment. Unlike in LFA models, no distinction is made between abilities of items to measure their respective attributes, i.e. all

items are considered equally good indicators of the attributes. This assumption may be too strong for many psychological assessments. To allow items differ in the abilities to measure their respective attributes, one needs to introduce item-by-attribute-specific weights,

$$t_{ji} = \mu_i - \left(\sum_{a=1}^{d} w_{ia}^2 \left(\theta_{ja} - \delta_{ia} \right)^2 \right)^{1/2} + \varepsilon_{ji}, \tag{7}$$

$$t_{ji} = \mu_i - \frac{1}{2} \sum_{a=1}^{d} w_{ia}^2 \left(\theta_{ja} - \delta_{ia} \right)^2 + \varepsilon_{ji}, \tag{8}$$

using the Euclidean distance of the squared Euclidean distance respectively. Zero weights make some attributes irrelevant to the item utility judgment, easily accommodating the most common independent-clusters design. Positive[2] but differing weights would allow the attributes to influence the utility judgment to different extents.

3. Decision Models for Choice Behavior

Absolute evaluations of items are the basis of measurement of the underlying attributes. To model forced-choice questionnaire data, however, the decision mechanism by which respondents make **comparative** judgments must be considered. Presumably, comparative judgments are underpinned by absolute judgments about each item under comparison. This is indeed the case in popular decision theories. The oldest and best-known model for choice behavior is Thurstone's (1927) law of comparative judgment. Other influential models are Coombs's (1950) unfolding preference model, Luce's (1959) choice axioms, Tversky's (1972) "elimination by aspect" theory and others. The former three have been applied to modelling forced-choice questionnaire data.

---

[2] Negative weights in IP models do not make sense conceptually; hence, we use the squared values.

Thurstone's Law of Comparative Judgment

Using the notion of item utility, Thurstone (1927) argued that the relative utilities of items at the time of comparison determine the choices between them. That is, person *j* prefers item *i* to item *k* if his/her utility for *i* is greater[3] than his/her utility for *k*:

$$y_{j\{i,k\}} = \begin{cases} 1, & \text{if} \quad t_{ji} \geq t_{jk} \\ 0, & \text{if} \quad t_{ji} < t_{jk} \end{cases}. \tag{9}$$

When comparing three or more items, person *j* assigns ranks according to the relative order of his/her utilities for the items; that is, the utilities of items ranked 1, 2 ,.., *n* must be ordered so that $t_{j1} \geq t_{j2} \geq \ldots \geq t_{jn}$. It follows that for every pairwise combination of items within a ranking block, inequalities (9) must hold.

It is easy to see that in (9), the *difference of utilities,*

$$y^{*}_{j\{i,k\}} = t_{ji} - t_{jk}, \tag{10}$$

is the latent response tendency, which is mapped onto the discrete response scale of observed outcome $y_{j\{i,k\}}$ via a threshold process

$$y_{j\{i,k\}} = \begin{cases} 1, & \text{if} \quad y^{*}_{j\{i,k\}} \geq 0 \\ 0, & \text{if} \quad y^{*}_{j\{i,k\}} < 0 \end{cases}. \tag{11}$$

Thurstone's simple law provides a flexible mechanism for explaining choice behavior in many contexts by adopting suitable operationalization of utility. Despite the deterministic

[3] The equality sign in (9) is arbitrary because the utilities are continuous variables and two utilities can never take on exactly the same value (Maydeu-Olivares & Böckenholt, 2005).

rule of utility maximization, probabilistic modeling of preference responses is easily achieved by considering utility judgment a random process, as in (2). To describe the probability of preferring one item to another conditional on the attributes, an appropriate link function is chosen depending on the assumed distribution of the utility random parts ε_{ji}. Thurstone's assumption of normality of errors (and consequently normality of their difference) demands the normal ogive (probit) link function. The usual assumption of independent errors, once the attributes have been controlled for, leads to the additive variance for the error of the utility difference, $\text{var}(\varepsilon_i - \varepsilon_k) = \text{var}(\varepsilon_i) + \text{var}(\varepsilon_k)$. Denoting error variances for utilities of items *i* and *k* as ψ_i and ψ_k respectively, the probability of preferring *i* to *k* is

$$P\left(y_{j\{i,k\}} = 1 \middle| \boldsymbol{\theta}_j\right) = \Phi\left(\frac{\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right) - \bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)}{\left(\psi_i^2 + \psi_k^2\right)^{1/2}}\right). \tag{12}$$

Coombs's Unfolding Preference Model

In his choice model, Coombs (1950) assumed the proximity, or ideal point, process underlying evaluations of individual items. Thus, given the choice between two stimuli, a person will prefer the stimulus located nearer to his/her own position (ideal point) on the attribute continuum. Then, the preference ordering of *n* items should correspond to the inversed order of the items' distances from the person position (Coombs, 1950). The preferred item can lie to the left or to the right of the person's ideal point, as long as it is closer to the ideal point than the other item. Any preferential rank ordering, therefore, is the same as the orders of the item locations "folded" at the person location. In the unidimensional case that Coombs considered, he operationalized the person-item distance as the absolute difference between person and item locations (i.e. Euclidean distance).

Bennett and Hays (1960) generalized Coombs's theory to multiple attributes. Let $\boldsymbol{\theta}_j = \left(\theta_{j1}, \theta_{j2}, ..., \theta_{jd}\right)'$ be the location of person *j* in a space defined by *d* attributes, and $\boldsymbol{\delta}_i = \left(\delta_{i1}, \delta_{i2}, ..., \delta_{id}\right)'$ be the location of item *i* in the same space. Given choice between items *i* and *k*, the person will prefer the item with the smallest distance $D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_i\right)$ from own location:

$$y_{j\{i,k\}} = \begin{cases} 1, & \text{if} \quad D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_i\right) \le D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_k\right) \\ 0, & \text{if} \quad D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_i\right) > D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_k\right) \end{cases}. \tag{13}$$

Because in (13) each item "has one and only one scale position for all individuals" and that "each individual has one and only one scale position for all [items]" (Coombs, 1950, p.146), the model has a major drawback – it is fully deterministic. Given choice between two items, the person will prefer the item closer to own location with probability 1. Such a model is likely to fit forced-choice questionnaire data poorly. To describe empirical preferential choices, various models have been proposed that introduce random processes in the judgments of distance, ideal point or item location (e.g., Zinnes & Griggs, 1974).

The relation to Thurstone's model. It is easy to see that the unfolding preference model (13) is a special case of Thurstone's model (9). Indeed, if we conceptualize the person-item distance as minus the utility – Zinnes and Griggs (1974) called it "*disutility*" – the two models make identical predictions. Thus, the unfolding preference model is a special case of the law of comparative judgment, where the utility takes the form $t_{ji} = -D\left(\boldsymbol{\theta}_j, \boldsymbol{\delta}_i\right)$. When this deterministic expression for the item utility is modified by adopting the random ideal point model (4), Coombs's decision model yields the probability of preferring one item to another of the form (12). It is essentially a Thurstonian probabilistic model, where the ideal point measurement model is adopted implicitly.

Luce's Choice Axioms and Bradley-Terry Model

Luce (1959) took a top-down approach to describing individual choice behavior and postulated a set of general assumptions (axioms). His axiom of *independence from irrelevant alternatives* (or independence of context) states that the ratio of probabilities of choosing item *i* and choosing item *k* must be independent of the set to which these items belong. Applied to forced-choice modelling, there exists a ratio scale π_{ji}, representing person *j*'s *response strength* associated with item *i* (Luce, 1977), such that the probability of choosing item *i* from block S (or ranking item *i* first) is proportional to response strength of *i*:

$$P\left(y_{j\{i,S\}}=1\right)=\frac{\pi_{ji}}{\sum_{k\ \text{in}\ S}\pi_{jk}}. \tag{14}$$

A special case of this mathematical model for choice tasks with two alternatives (blocks of size $n = 2$) was proposed earlier by Bradley and Terry (1952). They postulated that if there exists a ratio scale π_{ji}, representing person *j*'s *true ratings* of the items, then the probability of preferring item *i* to item *k* is proportional to the items' true ratings: $\pi_{ji}/(\pi_{ji}+\pi_{jk})$.

The strong ratio properties assumed for item true ratings ("response strength" in Luce's terminology) led Bradley (1953) to a very convenient operationalization of them as the exponential function of item utility $\pi_{ji} = \exp(t_{ji})$. However, in applications oriented toward measurement of individual differences, "strict utility" models are replaced by random utility models such as (2). Then, the conditional probability of preferring *i* to *k* is a logistic function of the systematic parts of the item utilities

$$P\left(y_{j\{i,k\}}=1\middle|\boldsymbol{\theta}_j\right)=\frac{\exp\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)\right)}{\exp\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)\right)+\exp\left(\bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)}=\frac{1}{1+\exp\left(-\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)-\bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)\right)}. \tag{15}$$

More generally, McFadden (1973) showed that when the utility judgments are independent, the probability of selecting item *i* from block S is a logistic function of utilities, obtained by using $\pi_{ji} = \exp(t_{ji})$ in expression (14). Because in random utility models (2), the errors are assumed independent after the utility judgments are conditioned on the personal attributes, Luce's axiom of independence from irrelevant alternatives applies. Thus, the probability of ranking item *i* first in block S is a generalization of the logistic response function (15) to blocks of size $n > 2$:

$$P\left(y_{j\{i,S\}} = 1 \middle| \boldsymbol{\theta}_j\right) = \frac{\exp\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)\right)}{\sum_{k \text{ in } S} \exp\left(\bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)} = \frac{1}{\sum_{k \text{ in } S} \exp\left(-\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right) - \bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)\right)}. \quad (16)$$

The relation to Thurstone's model. Parallels between the notions of Luce's "response strength", Bradley and Terry's "true rating", and Thurstone's "utility" are obvious. However, Thurstone's theory placed no conditions on the probability of choosing item *i* from a block beyond implying that it must equal the probability that the utility of *i* is the largest of all other items' utilities. Luce's condition (14) is more restrictive in prescribing "response probability proportional to response strength" (Luce, 1977; p. 216). Nearly identical prediction to those of Luce's model (departing by .02 at most) can be generated with the most restrictive case of Thurstone's model (Case V), where the utilities are assumed identically and independently distributed with equal variances (Luce, 1959). The prediction becomes perfect if the differences of utility judgments are distributed logistically rather than normally. Therefore, Luce's theory is consistent with Thurstone's model of choice behavior (Luce, 1977), when describing choices between independent alternatives with equal variances.

When utility judgments are dependent, Luce's choice axiom has been shown to overestimate the joint probability of selection for two items eliciting similar utilities

(McFadden, 1976). In forced-choice questionnaires, this would be commonplace since utilities of items depend on underlying psychological attributes. This problem, however, is solved by the use of random utility models and conditional probabilities. Once utility judgments have been controlled for the attributes, the random parts are independent. Assuming in addition that the random parts are identically distributed as double exponential[4] (their differences are then distributed logistically), expressions (15) and (16) provide suitable models for selecting item *i* from two or more alternatives, respectively (Takane & de Leeuw, 1987).

Andrich's Forced Endorsement Model[5]

Andrich (1989; also 1995) was interested in deriving the explicit probability of preferring one item in a pair (block size $n = 2$) from the probabilities of endorsement of the individual items. For this, he suggested a possible discrete response process whereby a person implicitly reacts to each item, reaching four possible outcomes: endorsing both items, ($y_i = 1$, $y_k = 1$); not endorsing either, ($y_i = 0$, $y_k = 0$); endorsing *i* but not *k*, ($y_i = 1$, $y_k = 0$); or endorsing *k* but not *i*, ($y_i = 0$, $y_k = 1$). The latter two judgments make the outcome of a forced-choice task obvious. That is, judgment (1, 0) should lead to preference for *i*, and the outcome $y_{\{i,\,k\}}=1$. Conversely, judgment (0, 1) should lead to preference for *k*, $y_{\{i,\,k\}}=0$.

---

[4] Double exponential (or Gumbel; sometimes referred to as Weibull) distribution has the cumulative function $F(z)=\exp(-\exp(-z))$.

[5] Andrich did not give his decision model a name – the name 'forced endorsement model' is suggested by the author of this article. This universal decision model should not be mistaken for specific IRT models for unfolding preference data that Andrich (1989, 1995) developed.

However, judgments (0, 0) and (1, 1) are not admissible in a forced-choice task. Andrich suggested that in this case, respondents are forced to reconsider the evaluations of both items until one of the two admissible outcomes is reached. He further assumed that "the distribution of these unacceptable original outcomes (0, 0) and (1, 1) among the acceptable ones, (1, 0) and (0, 1), is such that the latter retain their original relative probabilities" (Andrich, 1989; p. 197). With this, the probability of person *j* preferring item *i* over item *k* is given by the probability of endorsing *i* and not endorsing *k*, divided by the total probability of either admissible outcome

$$P\left(y_{j\{i,k\}}=1\right)=\frac{P\left(\mathrm{y}_{ji}=1,\,\mathrm{y}_{jk}=0\right)}{P\left(\mathrm{y}_{ji}=1,\,\mathrm{y}_{jk}=0\right)+P\left(\mathrm{y}_{ji}=0,\,\mathrm{y}_{jk}=1\right)}. \tag{17}$$

Assuming endorsements of items *i* and *k* independent events, conditional on personal attributes the items measure, the probability of response (1, 0) is the product of probabilities of responses $\mathrm{y}_{ji} = 1$ and $\mathrm{y}_{jk} = 0$. Then, the probability of preferring item *i* to item *k* can be written explicitly through probabilities of absolute judgments about the items

$$P\left(y_{j\{i,k\}}=1\middle|\boldsymbol{\theta}_j\right)=\frac{P\left(y_{ji}=1\ \middle|\boldsymbol{\theta}_j\right)P\left(y_{jk}=0\ \middle|\boldsymbol{\theta}_j\right)}{P\left(y_{ji}=1\ \middle|\boldsymbol{\theta}_j\right)P\left(y_{jk}=0\ \middle|\boldsymbol{\theta}_j\right)+P\left(y_{ji}=0\ \middle|\boldsymbol{\theta}_j\right)P\left(y_{jk}=1\ \middle|\boldsymbol{\theta}_j\right)}. \tag{18}$$

The relation to Thurstone's and Bradley-Terry models. Andrich's explicit probability expression (18) involves the probabilities of endorsement / rejection of individual items under comparison, and thus its specific form depends on the chosen IRT model for absolute judgments. The strong assumption of proportional relative probabilities of outcomes (1,0) and (0,1), however, demands the logistic distribution of errors, and also equal error

variances[6]. Using the logistic link function to describe the probabilities of endorsements and non-endorsements for items *i* and *k*, and denoting $\bar{t}_{ji}$ the systematic part of the random utility t_{ji} as before, we obtain

$$P\left(y_{ji}=1 \mid \boldsymbol{\theta}_j\right) P\left(y_{jk}=0 \mid \boldsymbol{\theta}_j\right) = \frac{\exp\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)\right)}{1+\exp\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)\right)} \cdot \frac{1}{1+\exp\left(\bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)}, \tag{19}$$

$$P\left(y_{ji}=0 \mid \boldsymbol{\theta}_j\right) P\left(y_{jk}=1 \mid \boldsymbol{\theta}_j\right) = \frac{1}{1+\exp\left(\bar{t}_{ji}\left(\boldsymbol{\theta}_j\right)\right)} \cdot \frac{\exp\left(\bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)}{1+\exp\left(\bar{t}_{jk}\left(\boldsymbol{\theta}_j\right)\right)}. \tag{20}$$

With this, the conditional probability of preferring item *i* to item *k* simplifies to the Bradley-Terry logistic model (15).

Classes of Models for Forced-Choice Questionnaires

Having established the axes on which forced-choice models may differ, we can identify potential classes of models, and classify the existing models. The first axis was defined by the **forced-choice format**, where we distinguished between pairs of items and larger ranking blocks. For the second axis (**measurement model**), we considered two random utility models – LFA and IP models. Further classification is possible based on whether a unidimensional or multidimensional variety is employed. The third axis was defined by the **decision model** adopted. Here, four different decision theories yielded only two fundamental probabilistic models for binary choice –using the probit link (12) or the logit link (15). The former type can be labelled *Thurstonian*, and the latter *Bradley-Terry*. The three axes can be

[6] Ignoring these assumptions and using the normal ogive link function results in probabilities that are different from those predicted by Thurstone's model (12). Discrepancies depend on the combination of two utilities, and can be large. For normally distributed utilities, Thurstone's model provides better prediction.

crossed to yield *Thurstonian LFA, Thurstonian IP, Bradley-Terry LFA* and *Bradley-Terry IP* models for either pairs of items (block size $n = 2$) or larger ranking blocks (block size $n \geq 3$).

Despite the fact that choices in blocks of any size may be described using binary outcomes of all pairwise comparisons, the block size has implications for modeling. With item pairs (blocks of size $n = 2$), the probability of an observed response pattern on the whole questionnaire equals the product of probabilities of all pairwise responses. This is because once conditioned on the underlying attributes, responses to the pairs are independent[7]. With blocks of size $n \geq 3$, the probability of an item having a certain ranking position is the probability of simultaneous outcomes of $n - 1$ comparisons with the rest of items in the block[8]. However, this probability is not equal to the product of probabilities of $n - 1$ preferences. This is because preference judgments involving the same item are **not independent**, even after controlling for the attributes. In random utility models (2), pairs involving the same item i have the shared random part of the common item utility, ε_i. Consider the utility differences for two comparisons $\{i, k\}$ and $\{i, q\}$, with the error terms $\left(\varepsilon_i - \varepsilon_k\right)$ and $\left(\varepsilon_i - \varepsilon_q\right)$ respectively. The covariance of the two error terms is not zero:

$$\operatorname{cov}\left(\varepsilon_i - \varepsilon_k, \varepsilon_i - \varepsilon_q\right) = \operatorname{cov}\left(\varepsilon_i, \varepsilon_i\right) = \psi_i^2 . \tag{21}$$

---

[7] Unlike in paired comparison tasks, it is assumed that no items are repeated across the forced-choice questionnaire. This is common practice in questionnaire design.

[8] For the partial ranking design whereby only one "best" item must be chosen, the multinomial logistic model of McFadden (16) may be used to model choices within each block, if it can be assumed that error variances are all equal. The choices for different blocks are independent conditional on the personal attributes, and the probability of observed response pattern is the product of probabilities of block choices. Since the assumption of equal error variances is often untenable, this model will not be considered further.

The local dependencies between pairwise judgments in blocks of size $n \geq 3$ must be modeled. This can be done by explicitly incorporating the local dependencies in the covariance structure of utility differences as shown later in the article.

Table 1 classifies the existing IRT models based on the measurement and the decision model they adopt. Because only Thurstonian IRT models (Brown & Maydeu-Olivares, 2011) allow for the local dependencies arising in blocks of size $n \geq 3$, the "block size" axis is not included in the table. In the remainder of this section, the existing IRT models are briefly described; for detailed descriptions and the models' applications to psychological assessment, see Brown and Maydeu-Olivares (in press). The models are presented in the chronological order of development.

TABLE 1 NEAR HERE

Zinnes-Griggs Model

This model was developed for questionnaires made of pairs (blocks of size $n = 2$) of ideal point items measuring one attribute. Zinnes and Griggs (1974) assumed that the utility-attribute relationship was described by a random utility IP model with two error components ε_j and ε_{ji}, representing "noisy" perceptions of own ideal point θ_j and of the item location δ_i respectively,

$$t_{ji} = -\left|\left(\theta_j + \varepsilon_j\right) - \left(\delta_i + \varepsilon_{ji}\right)\right|. \tag{22}$$

It was therefore assumed that the items vary only in their locations and are equally good indicators of the latent attribute. Further assuming normality and homogeneity of errors,

Zinnes and Griggs (1974) showed that the conditional probability of pairwise preference is given by the probit function

$$P\left(y_{j\{i,k\}}=1\middle|\theta_j\right)=1-\Phi\left(a_{j\{i,k\}}\right)-\Phi\left(b_{\{i,k\}}\right)+2\Phi\left(a_{j\{i,k\}}\right)\Phi\left(b_{\{i,k\}}\right),$$
$$\text{where}\qquad a_{j\{i,k\}}=\left(2\theta_j-\delta_i-\delta_k\right)/\sqrt{3} \tag{23}$$
$$b_{\{i,k\}}=\delta_i-\delta_k.$$

Thus, the model belongs to the Thurstonian IP class (unidimensional).

Andrich's Squared Difference and Hyperbolic Cosine Models for Pairwise Preferences

Andrich developed two models to describe choices made in pairs (n = 2) of ideal-point items measuring one attribute. Both models assume logistically distributed errors with equal variances; hence, they belong to the Bradley-Terry IP class (unidimensional).

In his first unfolding IRT model for pairwise preferences, Andrich (1989) assumed the simple (unit weights) squared person-item difference as the measure of distance, using a modified IP model (6), with $t_{ji}=-\left(\theta_j-\delta_i\right)^2+\varepsilon_{ji}$. This model can be named "Simple Squared Difference Model for Pairwise Preferences" or SSDMPP. The conditional probability for pairwise preference is

$$P\left(y_{j\{i,k\}}=1\middle|\theta_j\right)=\frac{1}{1+\exp\left(-2\left(\delta_i-\delta_k\right)\left(\theta_j-\frac{\delta_i+\delta_k}{2}\right)\right)}. \tag{24}$$

Andrich's (1995) second model used the absolute difference as the measure of distance. The resulting conditional probability for pairwise preference can be simplified using hyperbolic cosine, hence the model name – "Simple Hyperbolic Cosine Model for Pairwise Preferences" or SHCMPP.

$$P\left(y_{j\{i,k\}}=1\middle|\theta_j\right)=\frac{\cosh\left(\theta_j-\delta_k\right)}{\cosh\left(\theta_j-\delta_k\right)+\cosh\left(\theta_j-\delta_i\right)}. \quad (25)$$

Multi-Unidimensional Pairwise Preference (MUPP) Model

The MUPP (Stark, Chernyshenko & Drasgow, 2005) is a model for forced-choice pairs of dimension pure ideal point items measuring the same or different attributes. The MUPP model was derived by populating Andrich's explicit probability expression (18) with the binary version of the Generalized Graded Unfolding Model or GGUM (Roberts, Donoghue & Laughlin, 2000). The flexible GGUM enables the use of dimension pure items with varying discriminating power, varying locations and even varying maximal probability of endorsement. The resulting conditional probability of preference in the MUPP model is a logistic function; thus, the model belongs to the Bradley-Terry IP class (multidimensional).

McCloy-Heggestad-Reeve Unfolding Model

McCloy, Heggestad and Reeve (2005) adopted Coombs's unfolding preference model (13) to describe the process of responding to multidimensional forced-choice blocks compiled from dimension pure ideal point items. Because of its deterministic nature, the approach has not developed a formal IRT model; however, it can be used to enable pseudo-estimation of latent attributes. To this end, McCloy and colleagues suggested creating blocks of items with locations that are equal within the block and different between blocks, so that boundaries between person scores on different dimensions can be estimated using multidimensional unfolding. The person scores are then estimated as the midpoints of their lower and upper boundaries.

Thurstonian IRT Model

The Thurstonian IRT model (Brown & Maydeu-Olivares, 2011) was developed to enable analysis of data arising from forced-choice questionnaires using dominance items and blocks of any size. Items in each block may indicate the same or different attributes, or any mixture of the two. Moreover, dimension complex items can be modelled (Brown & Maydeu-Olivares, 2012). The conditional probability of pairwise preference is given by the probit link (12), with the LFA difference of systematic parts of the utilities (26). This model therefore belongs to the Thurstonian LFA class (multidimensional).

Fundamental Properties of Forced-Choice Measurement of Individual Differences

Probabilistic models for pairwise decisions, either using the probit link (12) or the logit link (15), depend on the difference of the systematic parts of item utilities, $\overline{t}_{ji} - \overline{t}_{jk}$. These quantities are central to forced-choice measurement; they determine how effective the item-pairs are in measuring their underlying attributes. This section looks at fundamental properties of the utility differences under LFA and IP models.

Utilities follow an LFA model

With the LFA model (3), the difference of systematic parts of item utilities is given by

$$\overline{t}_{ji} - \overline{t}_{jk} = \mu_i - \mu_k + \sum_{a=1}^{d} \left[\left(\lambda_{ia} - \lambda_{ka} \right) \theta_{ja} \right]. \tag{26}$$

In the common case of dimension pure items (i.e. each item measures only one attribute), the systematic difference of utilities for two items measuring **different attributes** θ_a and θ_b is

$$\overline{t}_{ji} - \overline{t}_{jk} = \mu_i - \mu_k + \lambda_{ia}\theta_{ja} - \lambda_{kb}\theta_{jb}, \tag{27}$$

and for two items measuring **the same attribute** it is

$$\overline{t}_{ji} - \overline{t}_{jk} = \mu_i - \mu_k + (\lambda_i - \lambda_k)\theta_j. \tag{28}$$

It follows from (28) that when factor loadings are equal, $\lambda_i = \lambda_k$, the response tendency does not depend on the person parameter θ_j at all; that is, unidimensional comparisons between items with equal factor loadings are uninformative (Maydeu-Olivares & Brown, 2010).

The unidimensional response tendency (28) in conjunction with either logit or probit link function will yield a familiar *s*-shaped curve, examples of which are presented in Figure 1. The two-dimensional response tendency (27) will yield the probability function describing a surface similar to the one presented in Figure 2.

FIGURES 1 AND 2 NEAR HERE

Utilities follow an IP model

With the ideal point model (4), the systematic difference of utilities is given by

$$\overline{t}_{ji} - \overline{t}_{jk} = \mu_i - \mu_k - \left(D(\boldsymbol{\theta}_j, \boldsymbol{\delta}_i) - D(\boldsymbol{\theta}_j, \boldsymbol{\delta}_k)\right). \tag{29}$$

The person-item distances can be measured by either the Euclidean (7) or the squared Euclidean (8) distance. In the general case, the distance formulae contain attribute- and item-specific weights.

Squared Euclidean distance IP model. It is convenient to rearrange the utility expression so that the fixed and random effects are clearly separated,

$$t_{ji} = \left[\mu_i - \frac{1}{2}\sum_{a=1}^{d} w_{ia}^2 \delta_{ia}^2 \right] + \sum_{a=1}^{d} w_{ia}^2 \delta_{ia} \theta_{ja} - \frac{1}{2}\sum_{a=1}^{d} w_{ia}^2 \theta_{ja}^2 + \varepsilon_{ji} = \\ = \tilde{\mu}_i + \sum_{a=1}^{d} w_{ia}^2 \delta_{ia} \theta_{ja} - \frac{1}{2}\sum_{a=1}^{d} w_{ia}^2 \theta_{ja}^2 + \varepsilon_{ji} \quad . \tag{30}$$

In the above, $\tilde{\mu}_i$ is the fixed intercept (expression in square brackets). Then the systematic difference of utilities is a quadratic function of person attributes

$$\bar{t}_{ji} - \bar{t}_{jk} = \tilde{\mu}_i - \tilde{\mu}_k + \sum_{a=1}^{d} \left(w_{ia}^2 \delta_{ia} - w_{ka}^2 \delta_{ka} \right) \theta_{ja} - \frac{1}{2}\sum_{a=1}^{d} \left(w_{ia}^2 - w_{ka}^2 \right) \theta_{ja}^2 , \tag{31}$$

unless the two items under comparison measure the same set of attributes and are equally discriminating on these attributes (i.e. their weights are equal within each attribute, $w_{ia}^2 = w_{ka}^2$). In this case, the quadratic terms θ_{ja}^2 disappear, and the resulting function is linear and identical to the one given by the linear factor model (26) with parameters

$\mu_i = \tilde{\mu}_i + \frac{1}{2}\sum_{a=1}^{d} w_{ia}^2 \delta_{ia}^2$ and $\lambda_{ia} = w_{ia}^2 \delta_{ia}$.

Brady (1989) and Tsai and Böckenholt (2001) noted this equivalence when considering a special case – the squared Euclidean distance IP model with unit weights. Essentially, from choices made between items measuring the same set of attributes, it is impossible to determine if the generating measurement model for the items was a LFA model or an IP model. In personality and similar assessments using forced-choice questionnaires, however, dimension complex items of that kind are never used. Attributes are usually measured with a set of items that form an independent clusters basis. For dimension pure items measuring **different attributes**, the difference of utilities is always a quadratic function of person attributes:

$$\bar{t}_{ji} - \bar{t}_{jk} = \tilde{\mu}_i - \tilde{\mu}_k + w_{ia}^2 \delta_{ia} \theta_{ja} - w_{kb}^2 \delta_{kb} \theta_{jb} - \frac{1}{2}\left(w_{ia}^2 \theta_{ja}^2 - w_{kb}^2 \theta_{jb}^2 \right). \quad (32)$$

However, for items measuring **the same attribute**, the systematic difference of utilities

$$\bar{t}_{ji} - \bar{t}_{jk} = \tilde{\mu}_i - \tilde{\mu}_k + \left(w_i^2 \delta_i - w_k^2 \delta_k \right) \theta_j - \frac{1}{2}\left(w_i^2 - w_k^2 \right) \theta_j^2 \quad (33)$$

is a linear function of person attribute θ_j when the item weights are equal, $w_i^2 = w_k^2$. Thus, for unidimensional ideal point items with unit weights, comparative data are indistinguishable from comparative data generated by an LFA model with parameters $\mu_i = \tilde{\mu}_i + \frac{1}{2}\delta_i^2$ and $\lambda_i = \delta_i$. Furthermore, when the item locations are the same, $\delta_i = \delta_k$, the response tendency in (33) does not depend on the person parameter θ_j at all; that is, unidimensional comparisons between items with equal weights and locations are not informative.

The unidimensional response tendency (33) with equal weights is a linear function of θ_j, and the conditional probability of preference is a familiar *s*-shaped curve (see Figure 1). When weights are not equal, the shape of the curve can vary from an *s*-shaped function to a Gaussian function depending on the item locations, as illustrated in Figure 3. The two-dimensional response tendency (32) will yield the probability function describing a surface similar to the one presented in Figure 4. As can be seen, the ideal point response process to individual items involved in comparisons cause a "number of peaks and valleys" (Drasgow, Chernyshenko, and Stark 2009; page 74) in the response surface.

FIGURES 3 AND 4 NEAR HERE

Euclidean distance IP model. For dimension pure items, the systematic difference of utilities of items measuring **different attributes** is

$$\overline{t}_{ji} - \overline{t}_{jk} = \mu_i - \mu_k - \left(w_{ia}^2 \left| \theta_{ja} - \delta_{ia} \right| - w_{kb}^2 \left| \theta_{jb} - \delta_{kb} \right| \right), \tag{34}$$

and the systematic difference of utilities of items measuring **the same attribute** is

$$\overline{t}_{ji} - \overline{t}_{jk} = \mu_i - \mu_k - \left(w_i^2 \left| \theta_j - \delta_i \right| - w_k^2 \left| \theta_j - \delta_k \right| \right). \tag{35}$$

When comparing unidimensional items with equal weights (i.e. $w_i^2 = w_k^2$), the systematic difference of utilities is a piecewise linear function of θ_j, which discriminates only between the location parameters of the two items. When the item locations are the same, $\delta_i = \delta_k$, the response tendency does not depend on the person parameter θ_j – identical to IP models using squared Euclidean distance.

The unidimensional IP response tendency (35) with equal weights will yield a piecewise *s*-shaped response function, examples of which are presented in Figure 5. As can be seen, the comparison is only discriminating between the item locations. In contrast, the Zinnes-Griggs version (22) of the IP measurement model yields smooth functions discriminating in the whole range of the attribute (see Figure 6). Two-dimensional IP functions akin to (34) yield a surface, similar to the one presented in Figure 4.

FIGURES 5 AND 6 NEAR HERE

Fitting Forced-Choice Models

Mean and Covariance Structure of Utility Differences

Forced-choice models may be formulated in terms of analysis of mean and covariance structures (Takane, 1987; Brady, 1989; Maydeu-Olivares & Böckenholt, 2005). This allows accounting for all dependencies between observations, either due to underlying attributes (and their co-variation in multidimensional models), or due to the local dependencies between pairwise choices in blocks of size $n \geq 3$. This approach is adopted in Thurstonian IRT models and briefly summarized here; for more details see Brown and Maydeu-Olivares (2012).

Because it is assumed that respondents are sampled randomly from the population of interest, all person-specific quantities (the utilities, the attributes and the errors) can be treated as random effects. These effects have some distributions over the population so they can be written in terms of random variables, omitting the person subscript *j*. A questionnaire with *p* forced-choice blocks elicits $p\tilde{n}$ pairwise comparisons, where $\tilde{n} = n(n-1)/2$ is the number of pairwise comparisons in each block of size *n*. Then the latent utility differences for each pairwise comparison can be written in matrix form as

$$\mathbf{y}^{*} = \mathbf{A}\mathbf{t}, \tag{36}$$

where $\mathbf{y}^{*}$ is a ($p\tilde{n}$) vector of latent differences $y^{*}_{\{i,k\}}$ described by equation (10); $\mathbf{t}$ is an (pn) vector of item utilities t_i, and $\mathbf{A}$ is a ($p\tilde{n} \times pn$) block-diagonal matrix of contrasts. Each block in $\mathbf{A}$ represents pairwise contrasts between items in a forced-choice block. When $n = 2$, each block in $\mathbf{A}$ is $\begin{pmatrix} 1 & -1 \end{pmatrix}$, whereas when $n = 3$, and $n = 4$, they are, respectively,

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \text{and} \qquad \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{37}$$

Given the assumption of normally distributed utilities, the differences of utilities are also normally distributed, so that only means and covariances are needed to describe them. The aim is to express the utility differences through the item parameters and person attributes, depending on the measurement model adopted.

When the utilities follow the LFA model (3), we have

$$\mathbf{t} = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{38}$$

where $\boldsymbol{\mu}_t$ is a (pn) vector of utility means, and $\boldsymbol{\Lambda}$ is a ($pn \times d$) matrix of factor loadings. The linear factor model assumes that the common factors $\boldsymbol{\theta}$ have the covariance matrix $\boldsymbol{\Phi}$; and that the unique factors $\boldsymbol{\varepsilon}$ are uncorrelated with the common factors and with each other so that their covariance matrix $\boldsymbol{\Psi}^2$ is diagonal. Then, the mean and covariance structure of the utility differences is given by:

$$\boldsymbol{\mu}_{y^*} = \mathbf{A}\boldsymbol{\mu}_t = -\boldsymbol{\gamma}, \qquad \boldsymbol{\Sigma}_{y^*} = \mathbf{A}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2\right)\mathbf{A}', \tag{39}$$

where $\boldsymbol{\gamma}$ is a ($p\tilde{n}$) vector of thresholds replacing the pairwise differences of the utility means (i.e. means of individual items are not of interest and are not estimated).

When the utilities follow the more algebraically tractable squared Euclidean distance IP model (30), the utilities may be presented in matrix form as follows

$$\mathbf{t} = \tilde{\boldsymbol{\mu}}_t + \Delta_w \boldsymbol{\theta} - \frac{1}{2}\mathbf{W}\boldsymbol{\theta}^{(2)} + \boldsymbol{\varepsilon}\,. \tag{40}$$

Here, $\tilde{\boldsymbol{\mu}}_t$ is a (*pn*) vector of fixed parameters (expression in square parentheses in (30)); $\boldsymbol{\Delta}_w$ is a (*pn* × *d*) matrix of the weighted item locations; **W** is a (*pn* × *d*) matrix of the weights; and $\boldsymbol{\theta}^{(2)}$ is a (*d*) vector of squared attributes $\boldsymbol{\theta}^{(2)} = \left(\theta_1^2, \theta_2^2, ..., \theta_d^2\right)'$. Because (40) involves both the linear and the quadratic terms of the attributes, the covariance structure of the utility differences will include the covariance matrix of vector $\boldsymbol{\theta}$, the covariance matrix of $\boldsymbol{\theta}^{(2)}$, and the covariance matrices of $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{(2)}$, and of $\boldsymbol{\theta}^{(2)}$ with $\boldsymbol{\theta}$. This rather complex structure cannot be reduced to the simple LFA structure (39) obtained by Takane (1987) and Brady (1989) for IP models with unit weights. This would amount to assuming equal weights for all items on all attributes – clearly an impossibility for forced-choice questionnaires involving dimension pure items measuring different attributes.

Identifying the Scale Origin

The problem of ipsative data is interpersonal incomparability of scores caused by non-identification of the scale origin for any of the measured attributes. The scale origin for the attributes cannot be identified because the total test score is constrained in ipsative scoring (Brown & Maydeu-Olivares, 2013). To ensure proper measurement of individual differences in IRT models, the scale origin must be identified for all measured attributes.

Böckenholt (2004) shows that it is impossible to identify the scale origin of item **utilities** t_{ji} or t_{jk} from their observed difference $t_{ji} - t_{jk}$ for any given person *j*. This is because the addition of any constant *c* to each of the utilities will result in the identical difference value, $(t_{ji} + c) - (t_{jk} + c) = t_{ji} - t_{jk}$. Therefore, from a single comparison between two items, it

is impossible to determine the absolute standing on item utilities; only the relative standings on the utility scale may be determined. The same is true for a block of *n* items – adding a constant *c* to all item utilities will yield identical pairwise utility differences.

Fortunately, the focus of measurement in forced-choice questionnaires is not the item utilities but the **attributes** that underlie them. When a comparison $\{i, k\}$ is made between items measuring the same attribute, the systematic difference of item utilities under the LFA measurement model is described by (28). Clearly, the attribute score is uniquely identified from this expression, unless the two factor loadings are equal, $\lambda_i = \lambda_k$. The same applies to unidimensional squared Euclidean ideal point models (33) with equal weights, because of their equivalence to LFA models for preference data. The attribute score is uniquely identified, unless the two item locations are equal, $\delta_i = \delta_k$. For the Euclidean distance IP model (35) with equal weights, scale identification is ensured in the interval between the two item locations, where the comparison is informative. To conclude, informative unidimensional comparisons (comparisons between items with different factor loadings or locations) ensure an identified scale origin of the attribute.

When multidimensional comparisons of dimension pure items are made, the difference of item utilities under the LFA measurement model is described by (27). As this is a linear equation with two unknowns, more information is needed to identify the scale of the attributes. In forced-choice questionnaires, attributes are measured with multiple items, which provide the additional information we need. For simplicity of illustration, consider two pairwise comparisons, $\{i, k\}$ and $\{q, r\}$, each measuring attributes θ_a and θ_b. From two systematic differences of item utilities,

$$\begin{aligned} \overline{t}_{ji} - \overline{t}_{jk} &= \mu_i - \mu_k + \lambda_{ia}\theta_{ja} - \lambda_{kb}\theta_{jb} \\ \overline{t}_{jq} - \overline{t}_{jr} &= \mu_q - \mu_r + \lambda_{qa}\theta_{ja} - \lambda_{rb}\theta_{jb} \end{aligned}, \tag{41}$$

the scales of the two attributes are uniquely identified, unless the factor loadings for the pair $\{i, k\}$ are a linear combination of the factor loadings for the pair $\{q, r\}$.

More generally, the vector of utilities under the LFA measurement model is described by (38), and the vector of latent utility differences is

$$\mathbf{At} = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{A}\boldsymbol{\Lambda}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\varepsilon}\,. \tag{42}$$

If the ($p\tilde{n} \times d$) matrix $\mathbf{A}\boldsymbol{\Lambda}$, which is the product of the ($p\tilde{n} \times pn$) design matrix $\mathbf{A}$ and the ($pn \times d$) matrix of factor loadings $\boldsymbol{\Lambda}$, is of full rank, the attributes' means and their covariance matrix $\boldsymbol{\Phi}$ can be identified. While the design matrix $\mathbf{A}$ is of reduced rank (Maydeu-Olivares, 1999; Böckenholt, 2004), leading to non-identification of the scale origin for the item utilities, the product $\mathbf{A}\boldsymbol{\Lambda}$ is generally of full rank (d) unless the factor loadings have special properties. Consider, for example, a questionnaire consisting of three blocks of size $n = 3$, with the following factor loading matrices for the utilities and the latent differences of utilities, respectively

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \\ \hdashline \lambda_4 & 0 & 0 \\ 0 & \lambda_5 & 0 \\ 0 & 0 & \lambda_6 \\ \hdashline \lambda_7 & 0 & 0 \\ 0 & \lambda_8 & 0 \\ 0 & 0 & \lambda_9 \end{pmatrix}, \qquad \mathbf{A}\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & -\lambda_2 & 0 \\ \lambda_1 & 0 & -\lambda_3 \\ 0 & \lambda_2 & -\lambda_3 \\ \hdashline \lambda_4 & -\lambda_5 & 0 \\ \lambda_4 & 0 & -\lambda_6 \\ 0 & \lambda_5 & -\lambda_6 \\ \hdashline \lambda_7 & -\lambda_8 & 0 \\ \lambda_7 & 0 & -\lambda_9 \\ 0 & \lambda_8 & -\lambda_9 \end{pmatrix}. \tag{43}$$

An example set of nine positive factor loadings {1, 1, 1; 1, 1.2, 0.5; 1, 0.7, 2.5} will yield the matrix $\mathbf{A}\boldsymbol{\Lambda}$ that is of full rank (rank = 3), leading to the identified scale origin of the three measured attributes. Combining equal in magnitude but different in sign factor loadings

in each block, for instance {1, -1, 1; 1, 1, -1; -1, 1, 1}, will also result in identified attribute scales. However, equal factor loadings within **every block**, $\lambda_1 = \lambda_2 = \lambda_3$, $\lambda_4 = \lambda_5 = \lambda_6$ and $\lambda_7 = \lambda_8 = \lambda_9$, will yield a degenerate matrix $\mathbf{A\Lambda}$ (its columns sum to 0), and non-identified scales for the attributes. Another example of a degenerate solution is equal factor loadings within **every attribute**, $\lambda_1 = \lambda_4 = \lambda_7$, $\lambda_2 = \lambda_5 = \lambda_8$ and $\lambda_3 = \lambda_6 = \lambda_9$. To conclude, unless LFA items with special properties are used (for instance, equal factor loadings within every block), the scale origin of the attributes is identified in forced-choice questionnaires with either unidimensional or multidimensional comparisons.

When dimension pure IP items measuring different attributes are compared, the difference of utilities is described by (32) (without loss of generality, the squared Euclidean distance can be used). For simplicity of illustration, we can also assume weights equal unity. Consider two pairwise comparisons, $\{i, k\}$ and $\{q, r\}$, measuring two attributes θ_a and θ_b,

$$\begin{aligned} \overline{t}_{ji} - \overline{t}_{jk} &= \tilde{\mu}_i - \tilde{\mu}_k + \delta_{ia}\theta_{ja} - \delta_{kb}\theta_{jb} - \frac{1}{2}\left(\theta_{ja}^2 - \theta_{jb}^2\right) \\ \overline{t}_{jq} - \overline{t}_{jr} &= \tilde{\mu}_q - \tilde{\mu}_r + \delta_{qa}\theta_{ja} - \delta_{rb}\theta_{jb} - \frac{1}{2}\left(\theta_{ja}^2 - \theta_{jb}^2\right) \end{aligned} . \qquad (44)$$

When the item locations for pair $\{i, k\}$ are identical to the locations for pair $\{q, r\}$, $\delta_{ia} = \delta_{qa}$ and $\delta_{kb} = \delta_{rb}$, the two equations above become identical with respect to the attributes (thetas), and their scales cannot be identified. Different item locations between the pairs, even when the locations within the pair are the same, for instance $\delta_{ia} = \delta_{kb} = 1$ and $\delta_{qa} = \delta_{rb} = -1$, would yield identified scales of the attributes. This simple illustration is in agreement with McCloy and colleagues (2005), who informally demonstrated that blocks of multidimensional comparisons with dimension pure ideal point items can produce uniquely identified scale origins for the measured attributes when items with equal locations are

combined in the same block, but the locations vary widely between blocks (i.e. locations of items within attributes vary).

Given that the general IP measurement model includes weights, and both linear and quadratic terms of the attributes, the conditions for identifying the scale origin are necessarily more complex than the simple matrix-rank condition for LFA models. To the author's knowledge, the conditions for identifying the scale origin for multidimensional IP models with weights have not been developed.

Estimating Item Parameters

Mean and covariance structures of utility differences can be estimated using general-purpose SEM software. The observed binary outcomes of pairwise comparisons **y** are linked to the response tendencies **y*** through the threshold process (11), and the response tendencies depend on latent variables θ (person attributes). Estimation of these models involves integration, which can be performed in two ways – either by integrating over the latent differences of utilities, or over the latent attributes (Takane & De Leeuw, 1987). The dimensionality of integration in the former is large, but because the utility differences are assumed multivariate normal, estimation can proceed by pieces (using tetrachoric correlations). This limited information approach uses a generalized least squares (GLS) estimation. When integrating over the latent attributes, dimensionality of integration is smaller and the maximum likelihood estimation based on the full joint probabilities of response patterns can be used. However, one must be concerned about whether local independence conditional on the latent attributes is satisfied. As discussed previously, the local independence assumption holds when items are presented in pairs. However, when block size is $n > 2$, the local independence assumption does not hold. Even when block size is $n = 2$, the full information method might be too computationally demanding since forced-

choice questionnaires usually measure several attributes, thus necessitating multidimensional integration. Thus, limited information methods based on tetrachorics are the estimation method of choice for general mean and covariance structures arising from forced-choice questionnaires.

Theory for estimating LFA mean and covariance structures (39), aka Thurstonian IRT models, has been fully developed, including all necessary identification constraints (see Brown & Maydeu-Olivares, 2012). These models may be estimated with respect to item parameters and structural parameters (i.e. correlations between the latent attributes) using Mplus (L.K. Muthén & B.O. Muthén, 1998-2012), which conveniently combines all necessary features. Applications to date demonstrate successful re-analysis of existing forced-choice data using this approach (Brown & Maydeu-Olivares, 2013; Brown, 2009; Brown & Bartram, 2009-2011).

As for the IP models, theory for parameter specification and identification constraints has been developed for special (unidimensional) cases only. Specifically, item parameters (locations) in the Zinnes-Griggs model can be estimated by marginal maximum likelihood (MML) procedure (Zinnes & Griggs, 1974; Stark & Drasgow, 2002). In Andrich's (1989, 1995) models for unfolding pairwise preferences, the item locations can also be estimated by maximum likelihood. So far, no multidimensional IP models (i.e. MUPP model or McCloy-Heggestad-Reeve approach) have estimated the item parameters from actual forced-choice data, which is not surprising giving their complexity and non-monotone nature of response functions (see Figure 4). Currently these models may be used for person parameter estimation only, with the item parameters assumed known (in practice, they are established through single-stimulus item trials).

Estimating Person Parameters

For personality and similar assessments, differentiating between people on a set of attributes is the main objective of measurement. After the model parameters – the item parameters and the correlations between the latent attributes – have been estimated, the factor scores may be estimated. To this end, maximum likelihood estimation can be used. Alternatively, Bayesian estimation with the multivariate normal prior with the covariance matrix $\mathbf{\Phi}$ can be used, either maximizing the mean of the posterior distribution (expected a-posteriori or EAP), or its mode (maximum a-posteriori or MAP). The former can be used in applications with one or two measured attributes; the latter is recommended in applications with many measured attributes.

In questionnaires using blocks of size $n = 2$, responses are independent conditional on the attributes, thus making estimation straightforward. In questionnaires with blocks of size $n \geq 3$, however, structured local dependencies between pairwise comparisons involving the same item exist. So far, Bayesian estimation ignoring these local dependencies has been used in applications with good results (Brown & Maydeu-Olivares, 2011; 2012; 2013).

Conclusions and Discussion

This article provided a common framework for describing models for forced-choice questionnaire data, using three axes: 1) the forced-choice format (i.e. block size); 2) the measurement model describing the item utilities through a set of psychological attributes the questionnaire measures; and 3) the decision model explaining choices between individual items. The measurement model dictates the specific expression for the systematic differences of item utilities. The distributional assumptions about the utility random parts dictate the choice of the link function. Together, they determine the chosen set of formulae describing

the probabilities of pairwise preferences. By combining the three axes in different ways, models suitable for a wide range of forced-choice questionnaire data can be described.

Popular models for choice behavior have been used to describe decisions in forced-choice questionnaires. This article shows that the four decision models considered here yielded two types of IRT models for pairwise choices, with the conditional probabilities being either logit or probit functions of the systematic utility differences (in this article, they were denoted Thurstonian and Bradley-Terry types). While Coombs's unfolding model implicitly assumes an ideal point measurement model, and is a special case of Thurstone's law, the other decision models are open to the use of different measurement models. Thus, Thurstone' law of comparative judgment can be used in conjunction with either LFA or IP items, or indeed with items described by other measurement models. The same is true for the explicit probability expressions provided by Bradley and Terry, and Andrich – they can be used in conjunction with different measurement models for absolute judgments.

Choices between available options with respect to any of the three axes can be motivated by the necessity to find a model suitable for an existing forced-choice questionnaire, or by the opportunity to create a questionnaire from scratch. When modelling data collected using an existing questionnaire, the choice of options is narrowed by the nature of items (for example, assuming a dominance response process or an unfolding process), distributional assumptions, block size and the number of attributes measured. When creating a new questionnaire, the available options are vast, and the question arising is whether any of the options are superior to the rest.

The question of the optimal block size is one of finding a balance between the information captured by item comparisons, and questionnaire usability. Combining items in larger blocks increases the amount of information per item. While choice between $n = 2$ items yields only one piece of binary information, choice among $n = 3$ items yields $\tilde{n} = 3$ pieces,

and choice among $n = 4$ items yields $\tilde{n} = 6$ pieces of information. Although the item information is not additive due to the local dependencies between pairwise comparisons within the same block, gains are obvious. However, human capacity to process comparisons is limited. There is evidence that the use of large blocks increases cognitive complexity of choice tasks, and worsens the quality of data obtained, which may have adverse impact on less educated or people with lesser reading skills (Brown & Bartram, 2009-2011).

The question of the "best" measurement model to use in choice tasks was considered in this article by looking at the response tendencies determining choices – the utility differences under LFA and IP models. Although reflecting different response processes in absolute judgments, the two measurement models can yield identical or similar properties when it comes to comparative judgments. A number of important points concerning these properties are summarized in the article. First, IP models in which equally discriminating items measuring the same attribute are compared in the same block produce identical utility difference functions to LFA models. Therefore, if the objective is to measure psychological attributes using unidimensional comparisons, it does not matter which type of items are used – those assuming an LFA model, or those assuming an IP model. It is impossible to tell from comparative data which generating measurement model was assumed for the questionnaire items. Either measurement model will produce informative comparisons if basic rules of forced-choice block construction are met. These rules are remarkably similar for LFA and IP models. For LFA models, items must have very different factor loadings to yield well-discriminating unidimensional comparisons (Maydeu-Olivares & Brown, 2010); for IP models, items must have very different locations (Drasgow, Chernyshenko & Stark, 2009).

Second, the rules for identifying the attributes' scale origins and thus ensuring interpersonal comparability are remarkably similar for LFA and IP models. For unidimensional comparisons, items with different factor loadings in LFA models, and

different locations in IP models, ensure an identified scale origin of the attribute. For multidimensional LFA models, the matrix of factor loadings $\boldsymbol{\Lambda}$ must yield a full-rank matrix of contrast loadings $\mathbf{A}\boldsymbol{\Lambda}$. This, for instance, if not fulfilled when factor loadings within every block are equal, or when factor loadings within every attribute are equal. For multidimensional IP models, the locations of items within blocks may be equal; however, the locations of items within attributes must be different (McCloy, Heggestad & Reeve, 2005). As long as the attribute scales are identified, either LFA or IP models can be used for measuring psychological attributes using unidimensional or multidimensional comparisons.

It appears that these fundamental properties of forced-choice questionnaire data are not widely understood. The problems with identifying the scale origin of **utilities** from comparative judgments described by Böckenholt (2004) have translated into skepticism about the ability of forced-choice questionnaires to "recover" the absolute standings on **attributes**. In this article, it is shown that when the focus of measurement is the attributes underlying the items, the scale origin may be identified without any special remedies such as embedding a small number of unidimensional pairs into multidimensional forced-choice questionnaires advocated in Stark, Chernyshenko and Drasgow (2005; also Drasgow, Chernyshenko & Stark; 2009).

The conclusions about the fundamental similarity of measurement inferred from LFA and IP models are somewhat ironic, considering that there has been much debate about advantages of either measurement model in general, and their merits in forced-choice applications in particular (see Drasgow, Chernyshenko & Stark, 2010; and Brown & Maydeu-Olivares, 2010). In reality, the questionnaire developer has a choice – to either utilize simple-to-analyze and widely available LFA items, or more rare IP items, or indeed write new items under either model. Neither model is inherently superior for forced-choice

measurement; the choice of a measurement model should be based on other considerations. These include conceptual suitability of the item type to the type of assessment (e.g. comparing attitude statements might call for ideal point items, while comparing performance statements might call for dominance items); pragmatic considerations regarding item-writing practices (Huang & Mead, 2014; Brown & Maydeu-Olivares, 2010); and pragmatic modeling considerations (methods and software available, whether the item parameters can be estimated from forced-choice data etc.). Given the widespread confusion about the types of items that can be used in forced-choice questionnaires (see commentaries in response to Drasgow, Chernyshenko & Stark, 2010) it is hoped that this conclusion will inform well-motivated choices in applications.

Even a limited selection of decision models and measurement models included here yielded a wide range of forced-choice models possible in principle. Although the focus of this article was on simple **choice** formats requiring IRT models for binary data, psychometric modelling of comparative judgments does not have to stop here. A general SEM framework for estimating mean and covariance structure of utility differences described in this article allows extensions to ordered categorical or continuous data. Such extensions would enrich the current repertoire, and allow modeling graded comparisons data, pick any constant-sum data (aka compositional data; Chan, 2003; Böckenholt, 2006) – the range of potential future applications is vast. I look forward to new developments in this area.

References

Andersen, E. B. (1976). Paired comparisons with individual differences. *Psychometrika, 41(2)*, 141-157.

Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement, 13*, 193-296.

Andrich, D. (1995). Hyperbolic Cosine Latent Trait Models for Unfolding Direct-Responses and Pairwise Preferences. Applied Psychological Measurement, 20, 269-290.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*, 263-272.

Bennett, J. F. and Hays, W. L. (1960). Multidimensional Unfolding: Determining the Dimensionality of Ranked Preference Data. *Psychometrika, 25*, 27-43.

Block, J. (1961). *The Q-sort method in personality assessment and psychiatric research*. Springfield, IL: Charles C. Thomas.

Böckenholt, U. (2004). Comparative judgments as an alternative to ratings: Identifying the scale origin. *Psychological Methods, 9*, 453–465.

Böckenholt, U. (2006). Thurstonian-based analyses: Past, present and future utilities. *Psychometrika, 71 (4)*, 615–629.

Bradley, R. A. (1953). Some statistical methods in taste testing and quality evaluation. *Biometrics, 9*, 22-38.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39, 324–345.

Brady, H.E., 1989. Factor and ideal point analysis for interpersonally incomparable data. *Psychometrika, 54*, 181–202.

Brown, A. (2009). *Doing less but getting more: Improving forced-choice measures with IRT*. Paper presented at the 24th annual conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Brown, A. & Bartram, D. (2009-2011). *OPQ32r Technical Manual*. Surrey, UK. SHL Group.

Brown, A. & Maydeu-Olivares, A. (2010). Issues that should not be overlooked in the dominance versus ideal point controversy. *Industrial and Organizational Psychology, 3*, 489-493.

Brown, A. & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*, 460-502.

Brown, A. & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*: 1135–1147.

Brown, A. & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*, 36-52.

Brown, A. & Maydeu-Olivares, A. (in press). Modeling forced-choice response formats. In *Irwing, P., Booth, T. & Hughes, D. (Eds.), The Wiley Handbook of Psychometric Testing*. London: John Wiley & Sons.

Chan, W. (2003). Analyzing ipsative data in psychological research. *Behaviormetrika, 30,* 99-121.

Chan, W., & Bentler, P.M. (1998). Covariance structure analysis of ordinal ipsative data. Psychometrika, 63, 369–399.

Cheung, M.W.L, & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling, 9,* 55-77.

Christiansen, N., Burns, G., & Montgomery, G. (2005). Reconsidering the use of forced-choice formats for applicant personality assessment. *Human Performance, 18,* 267-307.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs,* 14.

Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review, 57,* 145-158.

Coombs, C. H. (1960). A theory of data. *Psychological Review, 67,* 143-159.

De Soete, G., & Carroll, J.D., 1983. A maximum likelihood method for fitting the wandering vector model. *Psychometrika, 48,* 553–566.

Drasgow, F., Chernyshenko, O.S., & Stark, S. (2009). Test theory and personality measurement. In *Oxford Handbook of Personality Assessment*, Butcher, J.N (ed.), p. 59-80. London: Oxford University Press.

Drasgow, F., Chernyshenko, O. S., & Stark, S. (2010). 75 years after Likert: Thurstone was right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3,* 465–476.

Huang, J., & Mead, A. D. (2014, July 7). Effect of Personality Item Writing on Psychometric Properties of Ideal-Point and Likert Scales. *Psychological Assessment*. Advance online publication. http://dx.doi.org/10.1037/a0037273

Jackson, D., Wroblewski, V., & Ashton, M. (2000). The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution? *Human Performance, 13,* 371–388.

Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York, NY: John Wiley.

Luce, R. D. (1977). The Choice Axiom after Twenty Years. *Journal of Mathematical Psychology, 15,* 215-233.

Martin, B. A., Bowen, C.-C. & Hunt, S.T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32,* 247-256.

Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika, 64,* 325-340.

Maydeu-Olivares, A. & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods, 10,* 285-304.

Maydeu-Olivares, A. & Böckenholt, U. (2008). Modeling subjective health outcomes: Top 10 reasons to use Thurstone's method. *Medical Care, 46,* 346-348.

Maydeu-Olivares, A. & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research, 45*, 935 - 974.

McCloy, R., Heggestad, E., Reeve, C. (2005). A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. *Organizational Research Methods, 8*, 222-248.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*. New York: Academic Press.

McFadden, D. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement, 5*, 363–390.

McFadden, D. (2001). Economic Choices. *The American Economic Review, 91 (3)*, 351-378.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organisational Psychology, 77*, 531-552.

Muthén, L.K. & Muthén, B.O. (1998-2012). *Mplus User's guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Roberts, J.S., Donoghue, J.R. & Laughlin, J.E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Schwarz, N., Knäuper, B., Hippler, H.J, Noelle-Neumann, E. & Clark, L. (1991). Rating scales numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570-582.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*, 325–345.

Stark, S., Chernyshenko, O. & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement, 29*, 184-203.

Stark, S. & Drasgow, F. (2002). An EM approach to parameter estimation for the Zinnes and Griggs paired comparison IRT model. *Applied Psychological Measurement, 26*, 208–227.

Takane, Y. (1987). Analysis of covariance structures and probabilistic binary choice data. *Communication and Cognition, 20*, 45-62.

Takane, Y. (1996). An item response model for multidimensional analysis of multiple choice data. *Behaviormetrika, 23*, 153-167.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review, 34, 273-286*.

Thurstone, L. L. (1928). Attitudes can be measured. *American journal of Sociology, 33*, 529-554.

Thurstone, L.L. (1929). The measurement of psychological value. In Smith and Wright (eds), *Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago*. Chicago: Open Court, 157-174.

Thurstone, L.L. (1931). Rank order as a psychophysical method. *Journal of Experimental Psychology, 14*, 187-201.

Tsai, R. C., & Böckenholt, U. (2001). Maximum likelihood estimation of factor and ideal point models for paired comparison data. *Journal of Mathematical Psychology, 45*, 795–811.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, Vol. 79(4), 281–299.

Vasilopoulos, N.L., Cucina, J.M., Dyomina, N.V., Morewitz, C.L. & Reilly, R.R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance*, *19*, 175-199.

Zinnes, J.L. & Griggs, R.A. (1974). Probabilistic, multidimensional unfolding analysis. *Psychometrika*, *39*, 327-350.

Table 1. Classification of existing IRT models for forced-choice questionnaires

		Decision model	
Measurement model		*Thurstonian*	*Bradley-Terry*
LFA	*Unidimensional*	Thurstonian IRT	
	Multidimensional	Thurstonian IRT	
IP	*Unidimensional*	Zinnes-Griggs	SSLMPP SHCMPP MUPP
	Multidimensional		MUPP

Note. LFA = Linear Factor Analysis; IP = Ideal Point. SSDMPP = Simple Squared Difference Model for Pairwise Preferences; SHCMPP = Simple Hyperbolic Cosine Model for Pairwise Preferences; MUPP = Multi-Unidimensional Pairwise Preference.

Figures

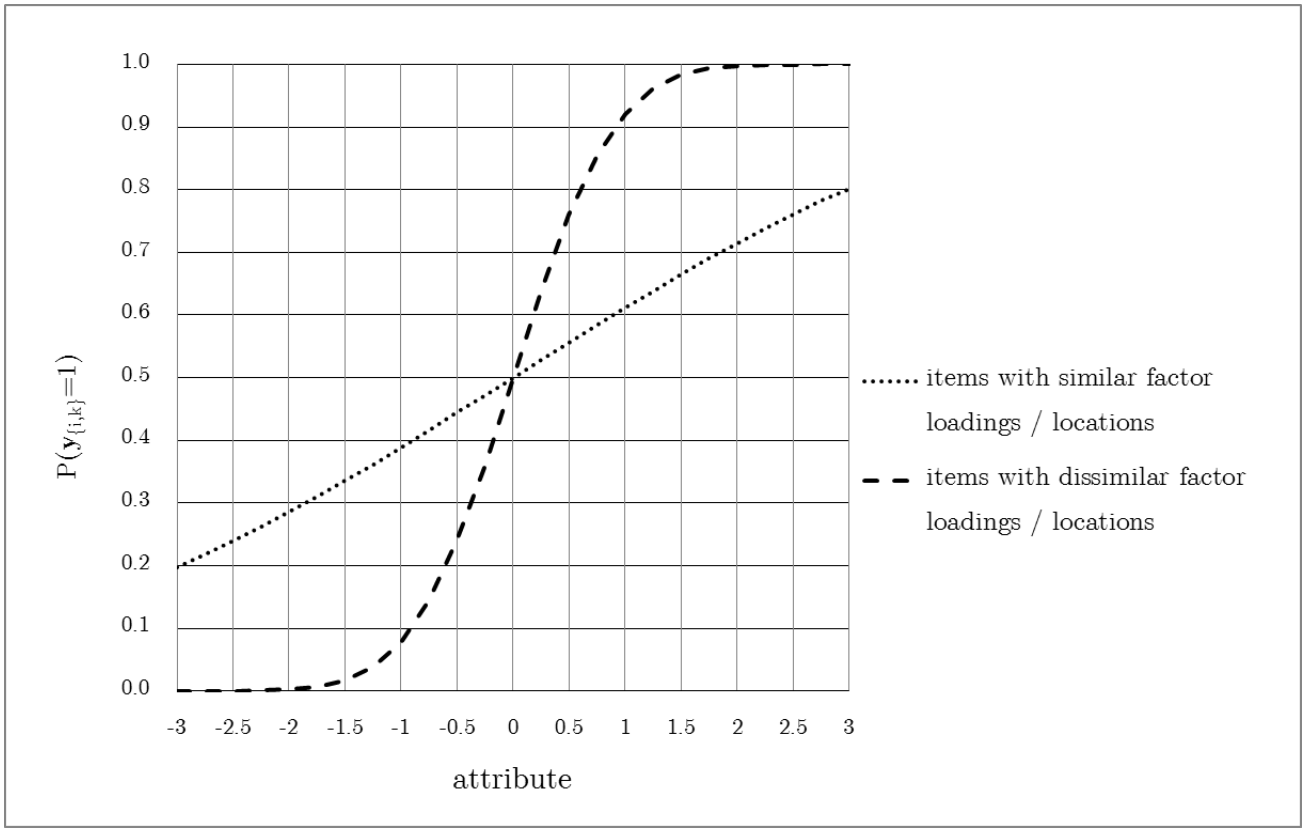

Figure 1. Response function for preference between two items measuring the same attribute under a LFA measurement model, and under a squared Euclidean distance IP model with equal weights.

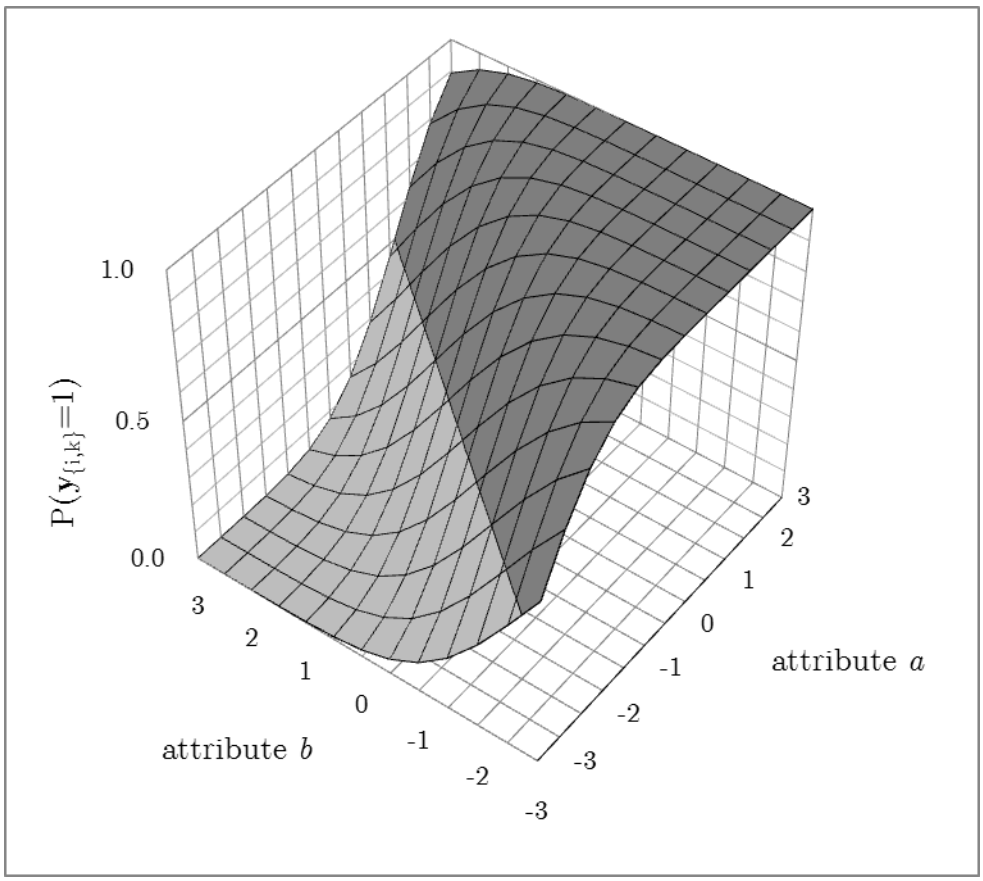

Figure 2. Response function for preference between two items measuring different attributes under a LFA measurement model.

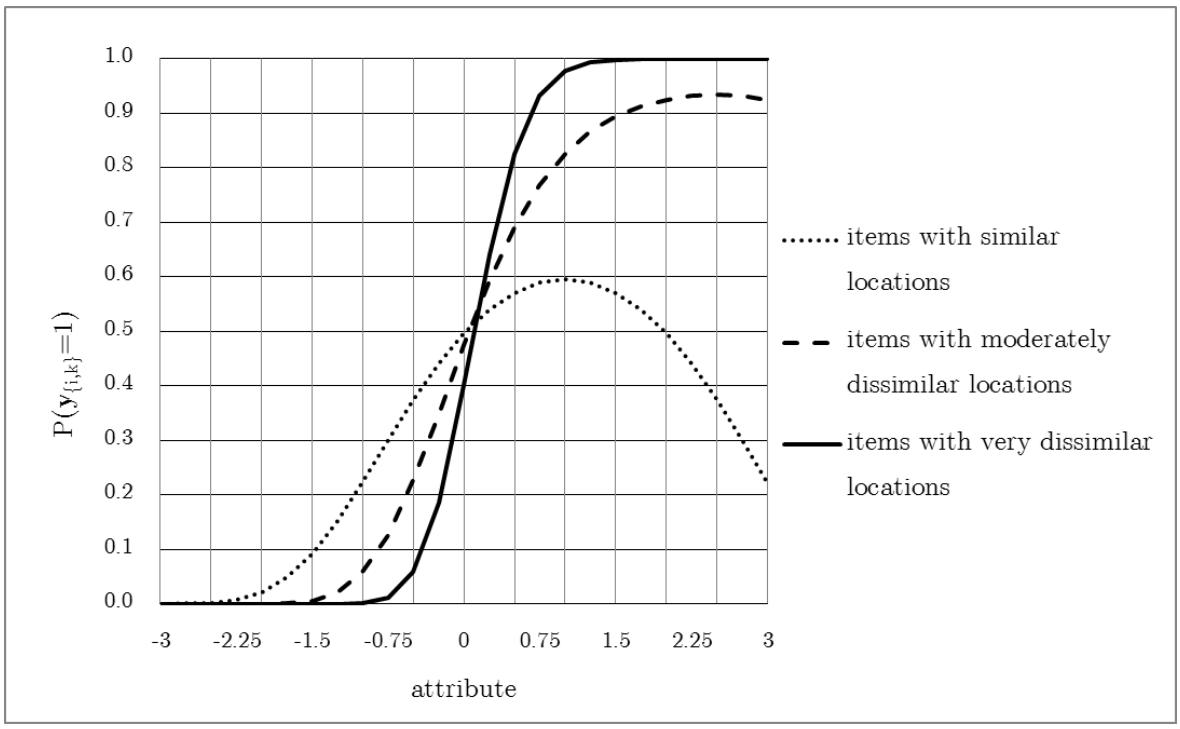

Figure 3. Response functions for preferences between two items measuring the same attribute under a squared Euclidean distance IP model with unequal weights.

Note. The first item has weight $w_i^2 = 1.5$, and the second has weight $w_k^2 = 1$.

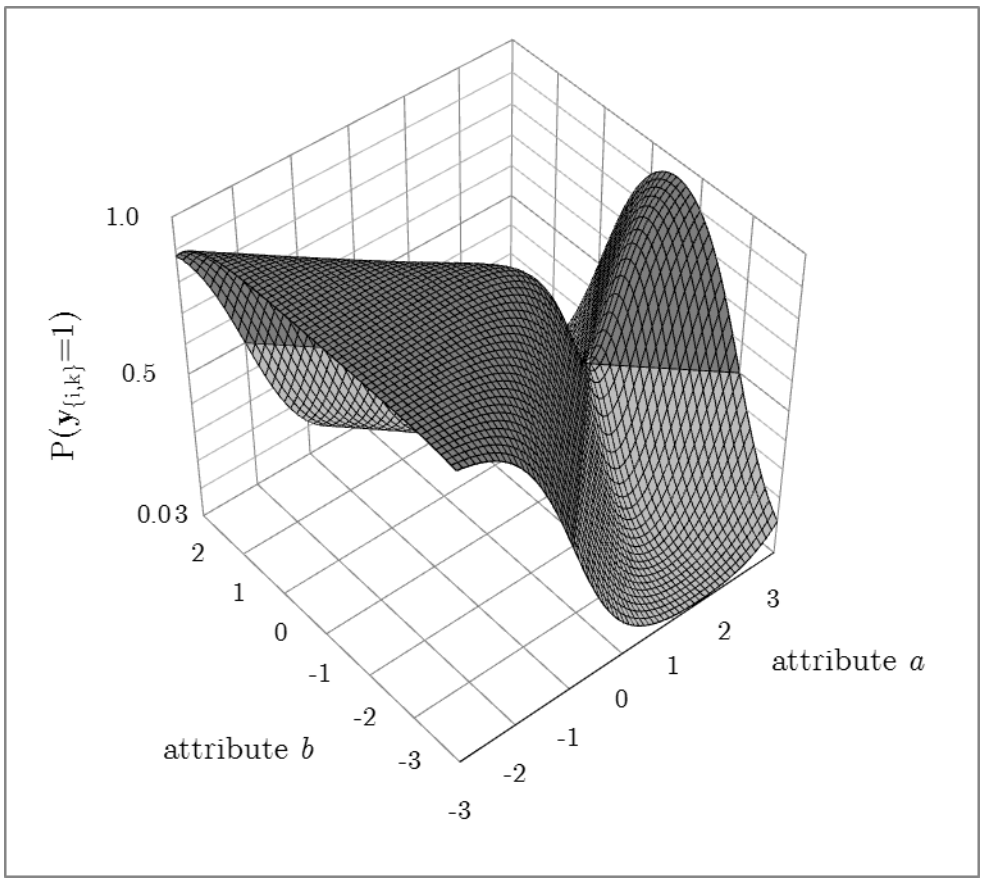

Figure 4. Response surface for preference between two items measuring different attributes under an IP measurement model.

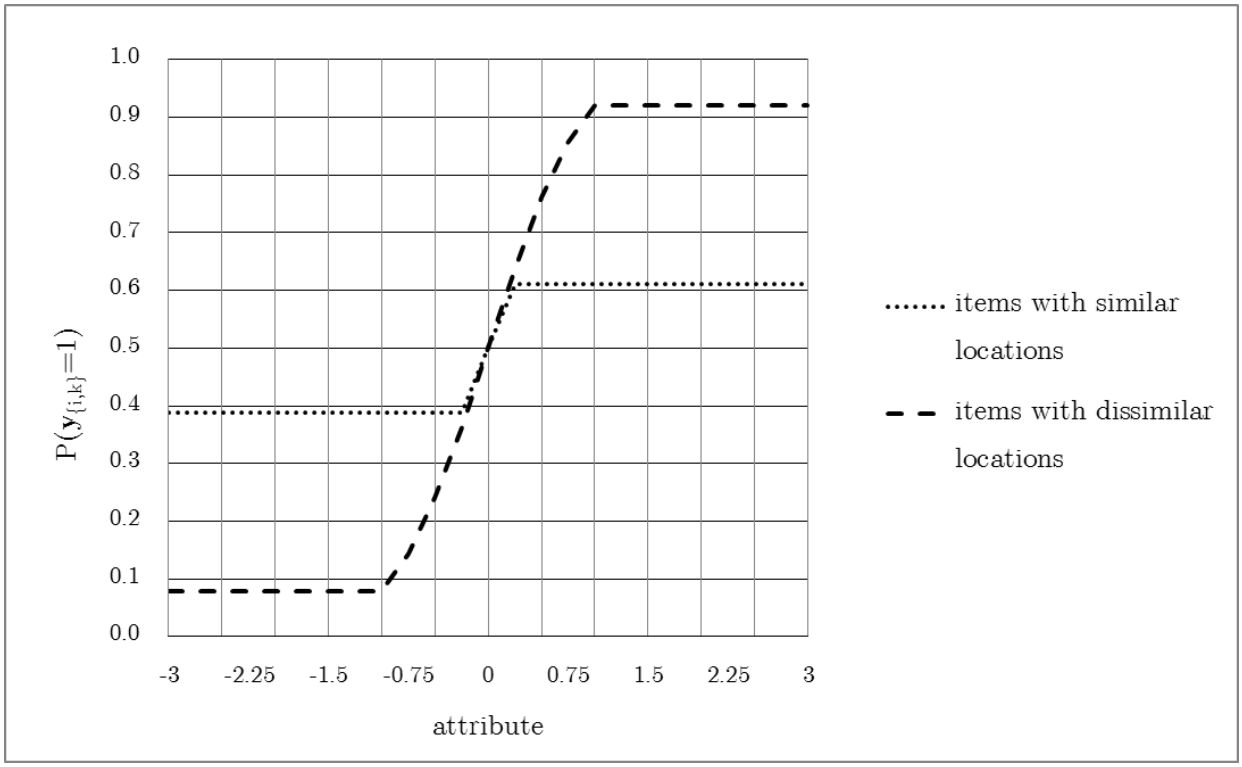

Figure 5. Response functions for preferences between two items measuring the same attribute under a Euclidean distance IP measurement model.

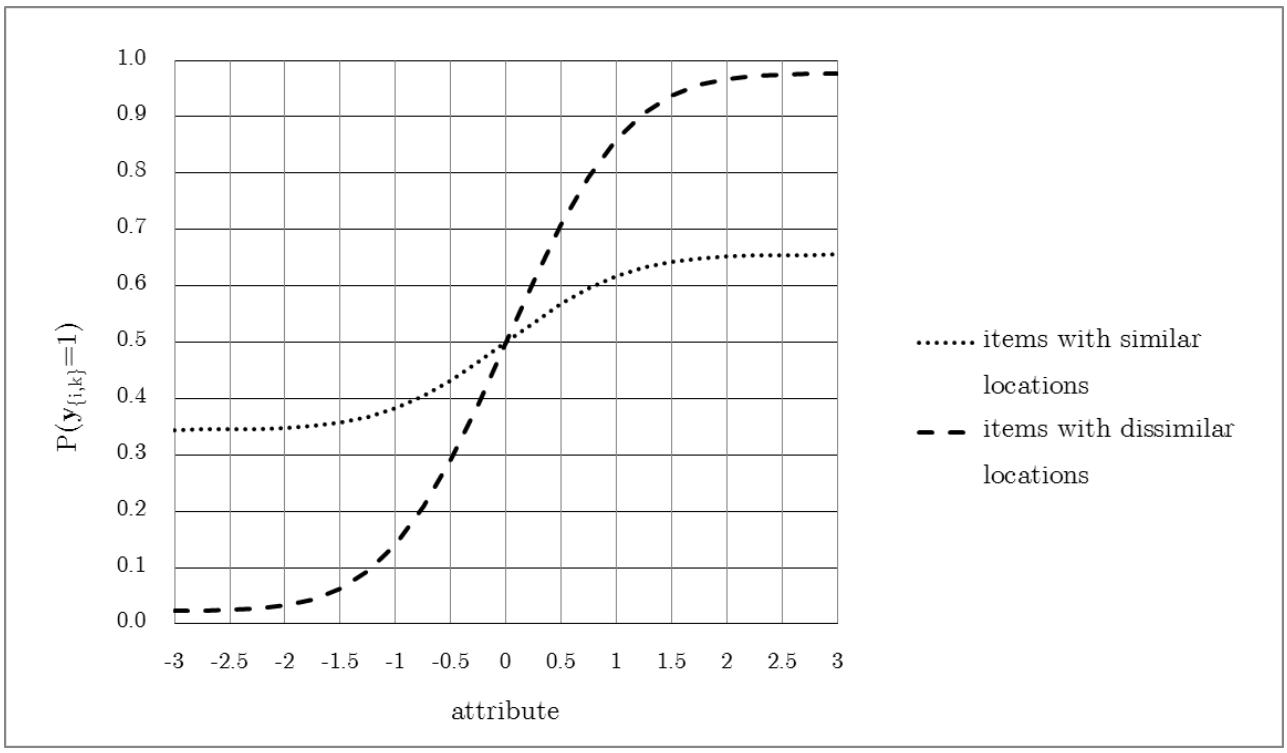

Figure 6. Response functions for preferences between two items measuring the same attribute under the Zinnes-Griggs version (22) of an IP measurement model.