International Encyclopedia of the Social and Behavioral Sciences, 2nd edition

MS. 25084: Personality Assessment, Forced-Choice

Author: Anna Brown

Contact information:

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

Keywords: forced-choice format, ipsative data, single-stimulus format, multidimensional IRT, comparative judgments, absolute judgments, dominance models, ideal-point models, unfolding

Abstract (50-100 words):

Instead of responding to questionnaire items one at a time, respondents may be forced to make a choice between two or more items measuring the same or different traits. The forced-choice format eliminates uniform response biases, although the research on its effectiveness in reducing the effects of impression management is inconclusive. Until recently, forced-choice questionnaires were scaled in relation to person means (ipsative data), providing information for intra-individual assessments only. Item response modeling enabled proper scaling of forced-choice data, so that inter-individual comparisons may be made. New forced-choice applications in personality assessment and directions for future research are discussed.

See also: Personality Assessment, Faking and; Personality Assessment, New Approaches to.

MS. 25084: Personality Assessment, Forced-Choice

Personality assessment is largely reliant on respondent-reported information. Objectively scored personality tests are perhaps the most widely used assessment mode. Such tests typically consist of multiple items, each intended to measure one personality trait. The most popular way of gathering responses on test items is by presenting them one at a time, as single stimuli ("single-stimulus" response format). Alternatively, two or more items at a time may be presented, forcing respondents to make a choice between them ("forced-choice" response format). Although appealing in many applications due to their ability to elicit finer differentiation between stimuli, forced-choice questionnaires have been controversial due to substantial scaling problems they pose. While it is easy to infer relative standings on different personality traits within one person, it has proven difficult to infer absolute trait standings for inter-personal comparisons from forced-choice data.

In this article, I consider first the different types of forced-choice formats, followed by a short discussion of their potential advantages in reducing response biases. I then turn to the measurement models that have been used to scale forced-choice responses, staring with the classical scoring model leading to ipsative scores, and followed by new model-based approaches based on Item Response Theory (IRT). Models that have been applied to real-world forced-choice personality assessments are briefly described. Next, I evaluate the merits of different forced-choice questionnaire designs in providing accurate assessments of personality traits. Finally, I discuss the direction for future research with forced-choice measures in cross-cultural, educational and industrial-organizational psychology.

Methods for Collecting Forced-Choice Questionnaire Data

Forced-choice data may be collected in many ways. Forced-choice questionnaires typically consist of multiple "blocks", presenting two or more items at the same time, and respondents are asked to indicate their preferences for the items within each block. Items within blocks may measure the same traits ("unidimensional" forced choice or UFC) or different traits ("multidimensional" forced choice or MFC). Table 1 summarizes the main types of forced-choice designs; each briefly described in this section.

Table 1. Methods for collecting forced-choice data.

Block size	*Binary preference*	*Graded preference*	*Proportional preference*
$n = 2$	Paired comparison	Graded paired comparison	Proportional paired comparison
$n \geq 3$	Ranking; Partial ranking; Ranking with ties (Q-sort)		"Proportion-of-total-amount" task

Binary Preference

The simplest forced-choice task involves a pair of items, or block of size $n = 2$. Respondents are asked to indicate their preference for one item, for instance, to indicate which of the two statements is most true of them:

	most true of me
A. I pay attention to details	X
B. I change my mood a lot	

Examples of assessments using forced-choice pairs are the Tailored Adaptive Personality Assessment System (TAPAS), the Navy Computerized Adaptive Personality Scales (NCAPS), the Edwards Personal Preferences Schedule (EPPS) and the Myers-Briggs Type Indicator (MBTI).

When a block consists of three or more items ($n \geq 3$), respondents may be asked to rank statements according to the extent they are true of respondents, or to indicate the top rank only, or the top and bottom ranks, for example:

	rank order	most / least true	most true
A. I pay attention to details	2		
B. I change my mood a lot	4	least	
C. I have a good word for everyone	3		
D. I catch on to things quickly	1	most	most

Examples of tests using rankings are the Occupational Personality Questionnaire (OPQ32i and OPQ32r), the Personality and Preference Inventory (PAPI), the Customer Contact Styles Questionnaire (CCSQ 7.2), the Gordon Personal Profile Inventory (GPP-I), the Survey of Interpersonal Values (SIV), the DiSC Classic, and the Kolb Learning Style Inventory.

In ranking with ties, also known as Q-sort (Block, 1961), respondents have to assign items to categories, complying with a pre-defined distribution. For example, respondents are asked to sort 10 items into five piles, according to the extent to which the items describe respondents' personality, and only allowing a certain number of items in every pile:

very untrue of me	somewhat untrue of me	neutral	somewhat true of me	very true of me
1	2	4	2	1
Number of items in pile				

California Adult Q-Sort, Riverside Behavioral Q-sort and Riverside Situational Q-sort are all examples of this format. Q-sorts are not rating tasks but pure forced choice, because assignments to categories are constrained so that respondents must perform direct comparisons between items. Even if no items are "very true" of the respondent, he/she must assign the prescribed number of items to that category. The obtained responses are essentially rankings with ties – a group of items ranked first, a group of item ranked second, etc.

Assessments using paired comparisons and ranking tasks yield binary preference data, since the only information collected for any given pair of items is whether one item was preferred to the other or not. For full rankings, binary choices are known for every pairwise comparison; that is, in the example above, it is known that item A was preferred to items B and C, but not to item D. For partial rankings, such as those resulting from "most/least" choices, some outcomes of pairwise comparisons are not known; for instance, it is not known whether

item A was preferred to C or not. For rankings with ties, preferences for items within the same category are not known.

Graded Preference

It is also possible to gather information on the extent of preference for one item over another. For example, we may ask respondents to indicate to what extent one statement is more (or less) true of them than the other using ordered categories.

	much more true	slightly more true	slightly more true	much more true	
A. I pay attention to details				X	B. I change my mood a lot

With this format, graded (ordinal) preference information is available for every paired comparison.

Proportional Preference

To gather further quantitative information about the relative merits of items, respondents may be asked to distribute a fixed number of points between them. For instance, we may ask respondents to distribute 100 points between $n = 4$ statements according to the extent the statements describe them:

	Points (100 in total)
A. I pay attention to details	40
B. I change my mood a lot	0
C. I have a good word for everyone	10
D. I catch on to things quickly	50

This format captures yet more information about the extent of preference than graded comparisons. It is not only known, for example, that A was preferred to B more than it was preferred to C, but how much more. Proportions of the total amount are collected and therefore it may be inferred, for instance, that preference for A (40 points) was 4 times stronger than preference for C (10 points).

Advantages of the Forced-Choice Formats

Comparative judgments employed in forced-choice questionnaires can have substantial advantages over absolute judgments. Firstly, forced choice makes it impossible to endorse all items indiscriminately (so-called “acquiescence” bias). It is also impossible to elevate or reduce ratings across all items (“leniency / severity” effects), or provide uniformly extreme or middle ground ratings (“extremity / central tendency” responding). Overall, the forced-choice formats eliminate any systematic response sets that apply uniformly across items (Cheung and Chan, 2002).

Secondly, forced choice tackles the problem with lack of differentiation in ratings (so-called "halo" effects). Halo effects are particularly problematic in personality assessments involving external raters (such as spouses, colleagues or bosses) who often have overgeneralized perceptions of different characteristics of the assessment target based on one important dimension. Forcing choice between various characteristics of the assessment target facilitates finer nuances of judgment and reduces halo effects, enhancing the quality of data.

Thirdly, binary preferences do not require any rating scales since items are compared directly. This is an advantage since test takers do interpret verbal and non-verbal anchors provided with the rating scale differently. Furthermore, Maydeu-Olivares and Böckenholt (2008) argue that comparing items directly may be cognitively simpler than rating them, particularly when there are many rating categories with few or poor verbal anchors.

Finally, the use of forced-choice formats in personality assessments has been largely motivated by attempts to reduce socially desirable responding. It has been thought from conception of forced-choice personality measures that combining equally desirable items in the same block would reduce socially desirable responding compared to single-stimulus formats, where all desirable items can be easily endorsed and all undesirable ones can be rejected. Extreme forms of socially desirable responding often referred to as "faking good", are particularly concerning in high stakes personality assessments, where interest in the use of forced-choice questionnaires has been growing. Over the years, evidence for superiority of forced choice in high stakes (e.g. Christiansen et al., 2005; Jackson et al., 2000) as well as against it (Feldman and Corah, 1960; Heggestad et al. 2006) has been published. Findings are inconclusive for many reasons; including lack of control for differences in questionnaire designs and testing contexts as well as technical challenges in modelling forced-choice data (see Section 3). Good methodology is essential to move this research forward, but most importantly, good understanding of test takers' cognitions when completing personality assessments in high stakes. While test takers' cognitions have been studied with single-stimulus measures (e.g. Robie et al., 2007), there is a clear gap in our understanding of such cognitions in forced-choice assessments.

Scaling of Forced-Choice Responses

While the forced-choice formats have been shown to eliminate or reduce many types of response biases, their use in personality assessment has been very controversial. The reason is that classical scoring methods, when applied to forced-choice questionnaires, yielded tests scores inappropriate for inter-individual comparisons. Recently, this problem has been overcome by the use of item response modelling.

Classical Scoring Model and Ipsative Scores

In the classical scoring scheme for single-stimulus items, more points are awarded to items that respondents endorse to a higher degree. The same logic has been applied to the forced-choice questionnaires. If item A is preferred to item B, it ought to be awarded more points. For instance, in a "percentage-of-the-total-amount" task, the points that respondents award to items become item scores. Because all respondents have the same number of points to distribute, their total score for each block is the same (for instance, 100).

	Scale	Person X	Person Y
A. I pay attention to details	Conscientiousness	40	5
B. I change my mood a lot	Neuroticism	0	25
C. I have a good word for everyone	Agreeableness	10	60
D. I catch on to things quickly	Openness	50	10
Total		100	100

When the item scores are then added to make up scale scores, the total score (the sum of all scale scores) on a classically scored forced-choice test is the same for everyone. The same logic applies to classical scoring of all forced-choice designs. For example, in a "most/least" partial ranking of four items, the most preferred item is given 2 points, the next two items are given 1 point each, and the least preferred item is given 0 points. The total amount of 4 points per block is the same for everyone, and the total score on the test is a constant.

This type of data is called "ipsative" (from Latin "ipse" – he, himself) because the scale scores are relative to self. Indeed, Person X in the example above obtained 50 points for Openness in one block, and this is his highest score in that block. However, because respondents have to indicate their preferences even if they agree with all items or disagree with all of them, this high score is only high relatively to other scores in the same block. It is possible that the absolute standing on Openness for Person X is in fact lower than that for Person Y, who only obtained 10 points for the same item. Ipsative scores are useful for intra-individual assessments because they provide valuable information of relative standings on personality traits for one person. They can be used in personal development, feedback, counselling etc. They, however, cannot be used for inter-individual comparisons.

The fact that the total score on the test is constant, despite different compositions of scale scores within that total for different people, causes substantial psychometric problems well described in the literature. As Clemans (1966), Hicks (1970) and others showed, ipsative scores cannot be analyzed in the same way as normative scores. The problems can be summarized as follows:

1) Ipsative scale scores always correlate negatively on average, even if the personality traits they measure correlate positively;
2) Ipsative scores cannot be factor analyzed using maximum likelihood method, because their variance-covariance matrix is of a reduced rank. Principal Components analysis can be applied to ipsative scores; however, the results are difficult to interpret since the components tend to contain conflicting rather than convergent traits (Cornwell and Dunlap, 1994);
3) Ipsative scores distort criterion-related validity estimates since their correlations with any external measure must sum to 0, creating spuriously positive and negative correlations;
4) Internal consistency reliability cannot be applied since the assumption of consistent coding is not met in ipsative scores (Brown and Maydeu-Olivares, 2013).

Different remedies have been suggested to alleviate the impact of ipsative constraints in classical scoring. One such remedy is increasing the number of measured traits in forced-choice questionnaires using the MFC format. It has been argued that when 30 or so traits are assessed, interpersonal comparisons can be performed meaningfully (Baron, 1996). Another way of releasing the ipsative constraint is to present items indicating negative trait standings (for

example, “I miss important details”) as well as items indicating positive trait standings (for example, “I pay attention to details”). Selecting the former statement will take points away from the Conscientiousness score, and selecting the latter will add points. Hence, the total score will show some variation. However, because both types of items indicating Conscientiousness are scored relatively to items indicating other traits, variation of the total score is still partially constrained (hence the name, “partially ipsative” data). Partially ipsative data and ipsative data arising from MFC questionnaires measuring many traits are less problematic; however, their psychometric problems are not eliminated but merely reduced.

The troubles with ipsative data stem from the fact that the implicit scoring model bears no relation to the psychological process used in forced-choice judgments (Meade 2004) – namely, relative positions on items are treated as if they were absolute positions. Adding relative positions of items together cannot possibly constitute a scale score that reflects person’s absolute standing on a trait.

Item Response Modeling of Forced-Choice Questionnaire Data

To overcome the problems of ipsative scores, any scoring protocol for forced-choice data must consider the true meaning of item responses. To this end, the response process involved in comparative judgments must be modelled. There are many theoretical models for individual choice behavior. The oldest and the best known is the law of comparative judgment of Louis Thurstone (1927). Other influential models are Coombs’s (1950) unfolding preference model, and Luce’s (1959) choice axioms. These theories have been adopted for modelling of forced-choice questionnaire data; in addition, Luce’s choice axioms have been applied by Andrich (1989) to provide explicit probability of binary choice.

IRT models for forced-choice questionnaire data can be classified according to three criteria: (1) block size that can be modelled, (2) dimensionality of comparisons (whether items measuring different traits can be modelled), and (3) the measurement model thought to underlie absolute judgments about questionnaire items.

Two types of measurement models have been utilized in forced-choice modelling: linear factor analysis models and ideal-point models. Linear factor analysis models assume that person *p*’s tendency to endorse item *i* (psychological value or item “utility”, as named by Thurstone) measuring personality trait *a* is a linear function of his/her trait score (LFA models allow items measuring multiple traits; a special case of factorially simple items is given here for simplicity).

$$\text{utility}_{pi} = \text{mean}_i + \text{loading}_i \, \text{trait}^a_p + \text{error}_{pi} \, . \qquad (1)$$

The factor analysis model assumes that for items indicating positive trait standings (for example, “I pay attention to details” indicating Conscientiousness), the item utility will increase as the trait score increases. For items indicating negative trait standings (for example, “I miss important details”), decrease in item utility is expected with an increase in trait score. This so-called “dominance” model is a fair representation of response process for most personality items that tend to represent extreme positive or negative standings on personality traits.

Ideal-point models assume that person *p*’s utility for item *i* measuring trait *a* is a function of distance between his/her trait score and the item location (there are many forms of ideal-point models; an expression using the unweighted Euclidean distance is given here without loss of generality).

$$\text{utility}_{pi} = \text{mean}_i - \left|\text{trait}_p^a - \text{location}_i\right| + \text{error}_{pi}\,. \tag{2}$$

This family of models assumes that items and people can be placed on the same trait continuum. Each item represents some position on the trait, and can be thought of as an "ideal" item describing characteristics of persons with this level of the trait (hence the name "ideal point"). For instance, item "My paperwork is always in order" represents a very high Conscientiousness score, and item 'My attention to detail is about average' represents an average score. For the latter item, the relationship between the item utility and Conscientiousness trait is clearly not linear, because the utility for this item will peak around the average Conscientiousness score, and be lower for respondents with either high or low scores. The ideal-point response process was originally proposed for attitude measurement, but recently has been suggested for use in personality assessments (Drasgow et al., 2010).

Thurstonian IRT model

Scope. Brown and Maydeu-Olivares (2011) developed the Thurstonian IRT model to enable analysis of data arising from forced-choice tests measuring multiple traits with ranking blocks of any size. Item parameters, correlations between the latent traits and person trait scores can be estimated using this IRT model.

Items in each block may measure the same trait (UFC) or different traits (MFC), or any combination of the two. A linear factor analysis model is assumed to describe the relationship between the items and traits they measure (i.e. the model utilizes the most common type of personality items, dominance items).

Origins. This model is based on Thurstone's (1927) law of comparative judgment. Thurstone postulated that preference judgments are determined by utilities of items under comparison. Each item elicits a utility judgment (judgment of the item's psychological value), and person *p* will rank item *i* above item *k* if his/her utility for *i* is higher than for *k*:

$$\text{prefer item } \begin{cases} i & \text{if} \quad \text{utility}_{pi} \geq \text{utility}_{pk} \\ k & \text{if} \quad \text{utility}_{pi} < \text{utility}_{pk} \end{cases}. \tag{3}$$

Observed rank orders of items within a block, therefore, can be seen as manifestations of the order of corresponding item utilities. To describe the observed ranks in a block, all pairwise comparisons between items are considered. There are $n(n\text{-}1)/2$ non-redundant pairwise comparisons in a block of size *n*. For example, any given rank order of four items, A, B, C and D, is fully described by knowing outcomes of six paired comparisons: {A, B}, {A, C}, {A, D}, {B, C}, {B, D} and {C, D}. For any of these six pairwise comparisons, the utility maximization rule (3) applies.

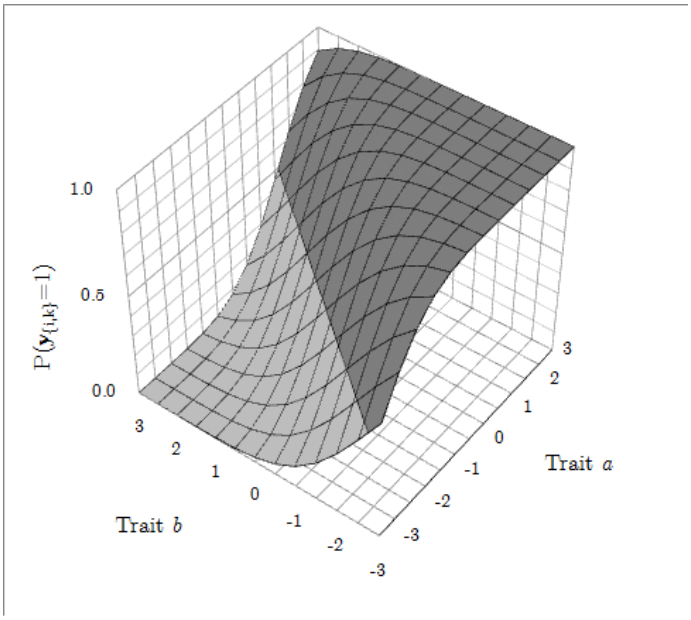
Model. Because preference for item *i* over item *k* is determined by the difference of their utilities, the utility difference is the central unit of analysis in Thurstonian modelling. To model responses to ranking blocks in relation to personality traits the items measure, Brown and Maydeu-Olivares (2011) assumed that person's utilities for items relate to the traits via a linear factor analysis model. Thus, the utility differences can be expressed as linear combinations of personality traits using (1). Observed binary preferences are thought to be dichotomized expressions of unobserved utility differences. With this, the probability $P(y_{p\{i,\,k\}}=1)$ of preferring

the first item in a pair {*i*, *k*} of items measuring different traits *a* and *b*, conditional on person *p*'s trait score is given by the cumulative standard normal function

$$P\left(y_{p\{i,k\}}=1\right)=\Phi\left(\frac{-\text{threshold}_{\{i,k\}}+\text{loading}_i\,\text{trait}_p^a-\text{loading}_k\text{trait}_p^b}{\sqrt{\text{var}(\text{error}_i)+\text{var}(\text{error}_k)}}\right). \quad (4)$$

For items with positive factor loadings, the probability of preferring item *i* to item *k* increases when the score on the trait measured by item *i* increases and the score on the trait measured by item *k* decreases. The response function (4) defines a surface, an example of which is presented in Figure 1, where the probability of preferring one item to another is plotted against two traits. The surface illustrates that the dominance process of evaluating individual items results in a monotonic relationship between the person's trait scores and the choice probability.

Figure 1. Example Thurstonian IRT response function for a pair of items measuring two different traits.



When items measuring the same trait are being compared in a forced-choice block, the conditional probability of preferring item *i* to item *k* by person *p* (4) simplifies to

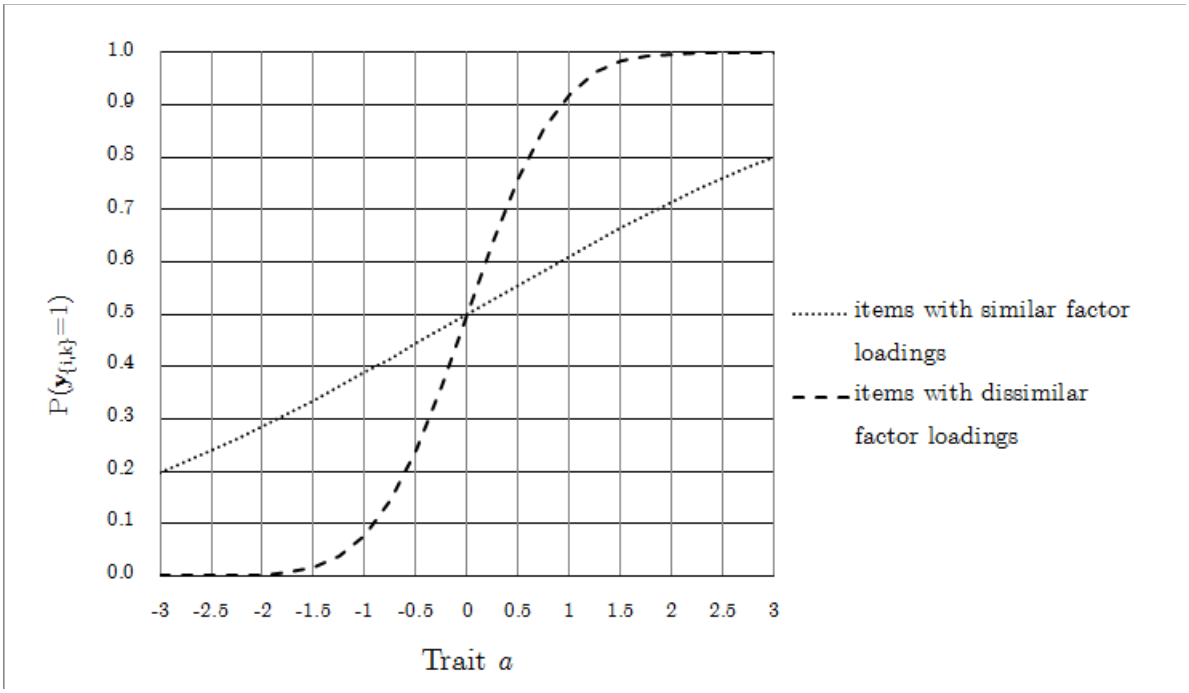
$$P\left(y_{p\{i,k\}}=1\right)=\Phi\left(\frac{-\text{threshold}_{\{i,k\}}+\left(\text{loading}_i-\text{loading}_k\right)\text{trait}_p^a}{\sqrt{\text{var}(\text{error}_i)+\text{var}(\text{error}_k)}}\right). \quad (5)$$

It follows from (5) that two items with similar factor loadings measuring the same trait will yield a pairwise comparison with a near-zero pairwise factor loading; thus it provides very little information for measurement of the latent trait. In other words, items that are likely to have

very similar valuations by the same person are ineffective as a pair for determining the absolute standing on the trait. This is a fundamental property of comparative judgments rather than of a specific model; this feature is shared by all forced-choice models, as the reader shall see.

The unidimensional response function is a curve, examples of which are presented in Figure 2. The figure illustrates that two items with very different factor loadings yield a highly informative comparison, whereas two items with very similar factor loadings (albeit highly discriminating in their own right) yield an uninformative comparison.

Figure 2. Example Thurstonian IRT response functions for two pairs of items measuring the same trait.



To enable parameter estimation, the pairwise preference model (4) is embedded in a generalized Structural Equation Modeling (SEM) framework. Responses to all ranking blocks (coded as binary outcomes of pairwise comparisons using (3)) are modelled simultaneously, each outcome serving as an indicator of common factors (latent personality traits), resulting in a single measurement model with binary outcomes (IRT model). The model takes care of dependencies that exist in blocks of size $n > 2$, where pairwise comparisons involving the same item share the same factor loading and the same error term. The model estimates item parameters according to (4) – factor loading and error variance for every item, and threshold for every pairwise comparison. In addition, correlations between the latent traits are estimated.

After the item parameters have been estimated, person trait scores can be estimated by either maximum likelihood method or Bayesian estimation. It has been shown that Thurstonian item response modelling approach overcomes the problems of ipsative data (Brown and Maydeu-Olivares, 2013).

Applications. Applications of this flexible model have included re-analysis of existing forced-choice questionnaire data and development of new measures. Re-analysis of the Customer Contact Styles Questionnaire data (CCSQ) demonstrates the advantages of IRT modeling for personality assessment; specifically, interpersonal comparability of person trait scores estimated

by the IRT method as opposed to the classical method resulting in ipsative scores (Brown and Maydeu-Olivares, 2013). The development of a new IRT scored version of the Occupational Personality Questionnaire (OPQ32r) illustrates how Thurstonian IRT modeling may be applied to re-analyze and re-develop an existing assessment tool, enhancing its strong features and transforming its scoring protocol (Brown and Bartram, 2009). The Thurstonian IRT model was also used to inform the development of a new measure, the Forced-Choice Five Factor Markers (Brown and Maydeu-Olivares, 2011).

Zinnes-Griggs model for unidimensional pairwise preferences

Scope. Zinnes and Griggs (1974) developed an IRT model to enable analysis of data arising from forced-choice tests consisting of simple paired comparisons (block size $n = 2$) of items measuring the same trait (UFC). Item parameters and trait scores for individuals can be estimated using this IRT model. An ideal-point process is assumed to describe the relationships between test items and the trait they measure.

Origins. This model is based on Coombs's (1950) preference decision theory. According to Coombs, when facing a choice between two items, the person will prefer the item closer to an "ideal" item representing own position on the trait ("ideal point"). Formally, given a choice between items *i* and *k*, each with their own location on the trait, person *p* will

$$\text{prefer item } \begin{cases} i & \text{if} \quad \left|\text{trait}_p - \text{location}_i\right| \le \left|\text{trait}_p - \text{location}_k\right| \\ k & \text{if} \quad \left|\text{trait}_p - \text{location}_i\right| > \left|\text{trait}_p - \text{location}_k\right| \end{cases}. \tag{6}$$

Coombs's model implicitly assumes an ideal-point response process for every item involved in comparison; that is, the psychological value or utility of an item to a person equals the inverse of the distance between the person trait score and the item location. Because of this inverse relationship between the distance to an item and the person's utility for it, Zinnes and Griggs called the item-person distance "disutility".

Model. Zinnes and Griggs (1974) modified the original deterministic version of Coomb's decision model to make it probabilistic. To this end, they consider "noisy perceptions" of item locations and the person's own ideal point at the time of comparison – three random variables distributed normally around their expected values. They showed that the probability $P(y_{p\{i,k\}}=1)$ of preferring the first item in a pair $\{i, k\}$ of items measuring the same trait, conditional on person *p*'s trait score is a one-dimensional IRT model, where $\Phi(\cdot)$ is a cumulative standard normal distribution function:

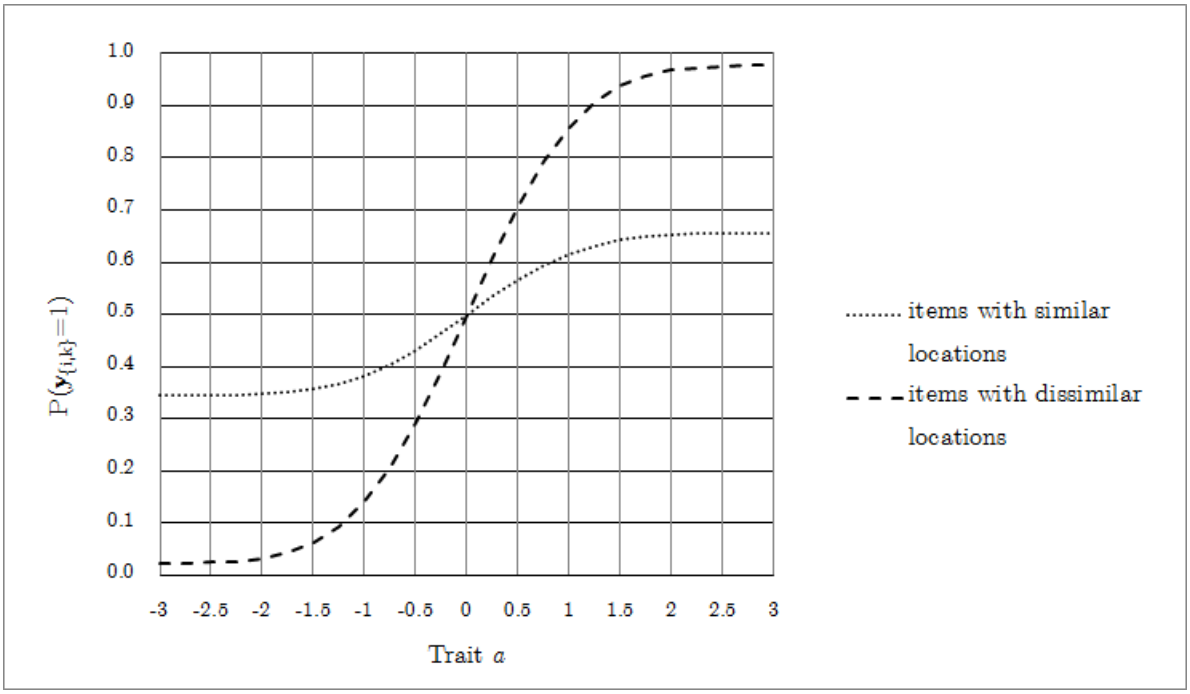
$$P\left(y_{p\{i,k\}}=1\right) = 1 - \Phi\left(a_{p\{i,k\}}\right) - \Phi\left(b_{\{i,k\}}\right) + 2\Phi\left(a_{p\{i,k\}}\right)\Phi\left(b_{\{i,k\}}\right), \tag{7}$$

$$\text{where} \quad a_{p\{i,k\}} = \left(2\cdot\text{trait}_p - \text{location}_i - \text{location}_k\right)/\sqrt{3}$$

$$b_{\{i,k\}} = \text{location}_i - \text{location}_k$$

In this simple model, the conditional probability depends only on the person's trait score and the item locations. It is therefore assumed that items vary only in their locations on the trait continuum, and thus they are equally good measures of the trait (equally discriminating).

Figure 3. Zinnes-Griggs response functions for two pairs of items measuring the same trait.



An example response function is illustrated in Figure 3, where the conditional probabilities of preferring the first item in a pair are plotted for two pairs of items, with different combinations of item location parameters. This figure shows that comparing items with similar locations results in a very "flat" function with a shallow slope; and comparing items with very dissimilar locations results in a function with a steep slope. Therefore, comparisons between items located closely on the same trait are non-informative. The same effect is observed when items with similar factor loadings are compared under the Thurstonian IRT model (see Figure 2).

Applications. This straightforward unidimensional model has been applied to create the Navy Computerized Adaptive Personality Scales (NCAPS).

Multi-Unidimensional Pairwise-Preference (MUPP) model

Scope. Stark and colleagues (2005) developed the MUPP model to estimate person trait scores from binary paired comparisons (block size $n = 2$) of items measuring the same trait (UFC) or different traits (MFC). Currently, item parameters cannot be estimated from the actual forced-choice data and are assumed known. The MUPP model assumes an ideal point response process for the items involved in comparisons.

Origins. The MUPP model adopts an approach to explaining preference judgments originally suggested by Andrich (1989). Andrich's motivation was to write an explicit expression for the probability of preferring one item to another through probabilities of accepting and rejecting the individual items. First, he postulated that the event of picking item *i* from the set {*i*, *k*} has the same probability as the event of picking the first of two alternatives: (1) endorsing *i* and rejecting *k*, or (2) endorsing *k* and rejecting *i*.

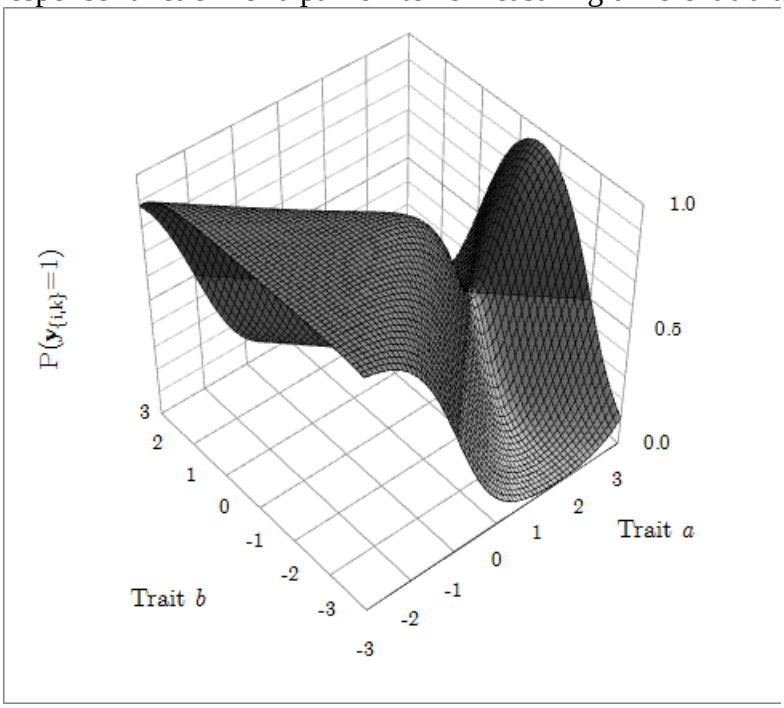
Second, Andrich assumed that acceptances and rejections of individual items are independent events, conditional on the traits the items measure, therefore the joint probability of endorsing *i* and rejecting *k* is the product of two probabilities, the probability of endorsing *i* and the probability of rejecting *k*. With this, the probability $P(y_{\{i,k\}}=1)$ of preferring the first item in a pair $\{i, k\}$ of items measuring different traits *a* and *b*, conditional on person *p*'s trait score is given by

$$P\left(y_{p\{i,k\}}=1\right)=\frac{P\left(\text{endorse } i\left|\text{trait}_p^a\right.\right)P\left(\text{reject } k\left|\text{trait}_p^b\right.\right)}{P\left(\text{endorse } i\left|\text{trait}_p^a\right.\right)P\left(\text{reject } k\left|\text{trait}_p^b\right.\right)+P\left(\text{reject } i\left|\text{trait}_p^a\right.\right)P\left(\text{endorse } k\left|\text{trait}_p^b\right.\right)}. \quad (8)$$

Model. Stark and colleagues (2005) use an ideal-point model to substitute the probabilities of accepting and rejecting individual items in general expression (8). Specifically, they advocate the use of a binary version of the Generalized Graded Unfolding Model or GGUM (Roberts, Donoghue, and Laughlin 2000). The GGUM is very flexible and allows items to differ in discrimination, locations and even in maximum probability of endorsement.

If items measuring different traits are compared, the item response function (8) in conjunction with the binary GGUM defines a surface, an example of which is presented in Figure 4. This surface is rather complex, with the ideal-point process of evaluating individual items resulting in a non-monotone relationship between the latent traits and the pairwise preference.

Figure 4. MUPP response function for a pair of items measuring different traits.



If items measuring the same trait are compared, the MUPP gives a response curve similar to those depicted in Figure 3. Just like under the Zines-Griggs model, comparisons between unidimensional items with similar locations are non-informative, and the same effect is observed

when items with similar factor loadings are compared under the Thurstonian IRT model (see Figure 2).

Applications. The MUPP model was used in the development of the Tailored Adaptive Personality Assessment System or TAPAS, a comprehensive and customizable assessment system measuring any subset of 23 personality facets deemed important for predicting job performance in civil and military organizations.

McCloy-Heggestad-Reeve unfolding model for multidimensional ranking blocks

Scope. McCloy and colleagues (2005) sketched a model for the process of responding to MFC blocks of any size, compiled from ideal-point items, to inform item selection and estimate person trait scores. Item parameters cannot be estimated from the actual forced-choice data and are assumed known.

Origins. This approach adopts an extension of Coombs's original one-dimensional unfolding model to the multidimensional case. The model predicts that a person will prefer an item that is located nearer to the person in multidimensional trait space than another item located further from the person in that space. The use of Coombs's unfolding decision model implicitly assumes the ideal-point response process for every item involved in comparisons, each reflecting changes in its own personality trait.

Model. Items are assumed equally good measures of the traits (equally discriminating), therefore the only differentiating feature for items is their locations. This and another simplifying assumption – that of no random influences on preference decisions – are used to inform creation of forced-choice designs that are effective for accurate estimation of trait scores. McCloy and colleagues show that assuming item locations known, an effective forced-choice measure can be assembled from blocks of items with locations that vary across the attribute space.

Discussion

Since conception of forced-choice personality measures, the biggest technical challenge was scaling of person's relative preferences to reflect their absolute standing on traits of interest. Classical scoring methods that simply counted the relative responses as absolute did not provide adequate measurement, making the forced-choice assessment method controversial. Modern scoring methods have the potential to provide proper scaling by taking to account the comparative nature of judgments. These methods link observed comparative decisions to unobserved personality traits, controlling for properties of items.

Despite the breakthroughs in scaling forced-choice questionnaire data, not every forced-choice design is guaranteed to deliver good measurement properties once an IRT model has been applied to it. Just like a single-stimulus questionnaire with poor items that do not provide information on target traits or with a poor rating scale, a forced-choice questionnaire may be incapable of providing accurate measurement of target traits whichever modern scoring model is applied. This is because the forced-choice format dictates its own rules of measurement.

One such rule was demonstrated in this article – comparisons between items that are likely to elicit similar value judgments in the same person are not informative for establishing the absolute trait position. For example, two items, "My paperwork is always in order" and "I am always on top of my paperwork", may be good measures of Conscientiousness in their own rights, but when put in a paired comparison, they will provide no information on the

Conscientiousness trait. This is because high scorers will agree with both items, and will be forced to make a random choice between the two. Low scorers will disagree with both items, and will again make a random choice. The same will happen for any given standing on the trait – the trait score will have no effect on the preference decision, with the random error playing the largest part. No IRT model, whether utilizing dominance items or linear factor analysis items, will be able to recover the absolute position on the trait in this situation. This is the fundamental property of comparative judgments, not of any specific model.

More generally, items that elicit similar utilities, using either the dominance process or the ideal-point process, are ineffective in forced-choice blocks. This is why comparing items measuring highly correlated traits is not effective, and combining items with factor loadings of opposite sign (measuring the positive and the negative extremes of positively correlated traits) may be necessary (Brown and Maydeu-Olivares, 2011). The same conclusions follow when ideal-point items are utilized in multidimensional forced choice – items with very different locations on traits are required (McCloy et al., 2005).

There has been much debate about relative merits of dominance and ideal-point items as utilized in forced-choice measurement (e.g. Drasgow et al., 2010; Brown and Maydeu-Olivares, 2010). However, it is becoming clear that all models for forced-choice are essentially similar; they model the same process, which has its own fundamental properties. The successful applications of both types of models in personality assessment show that nether dominance or ideal point family model is inherently superior for forced-choice measurement. Rather, the choice of model must be dictated by the nature of items used. If items that elicit the ideal-point response process are used (because they deemed more appropriate on conceptual grounds), an ideal-point forced-choice model is appropriate. If items that elicit dominance responses are used, a linear factor analysis forced-choice model is appropriate.

Finally, it is worth noting that all the existing IRT models are choice models; that is, they are suitable for modelling binary preference data (i.e. paired comparisons, rankings and partial rankings). Although extensions to ordinal or proportional preference data are possible within the generalized SEM framework adopted by Thurstonian modelling, these are yet to establish themselves in forced-choice personality assessment. Such developments would open doors for proper scaling of questionnaires employing graded preferences and "proportion-of-total-amount" tasks.

Conclusions and Directions for Future Research

The main objective of personality assessment – differentiating people in relation to traits of interest – can be achieved by collecting either absolute judgments or comparative judgments. In respondent-reported assessments, absolute judgments can be subject to numerous response biases, such as idiosyncratic uses of rating scales, unconditional agreement with statements as presented (acquiescence), lack of differentiation (halo effects) and others. Forcing choices between questionnaire items can help eliminate any effects acting uniformly across items. Reducing non-uniforms effects, such as differential inflation of scores on traits deemed important in specific contexts ("faking good"), with forced-choice formats is proving more challenging. More research is needed to understand the motivated distortions to forced-choice questionnaires in high stakes personality assessments. With rapid development of item response modelling approaches, we are well placed to tackle this challenge.

Even before the advent of item response modelling in the forced-choice assessment space, ipsative and partially ipsative scores demonstrated similar or even enhanced criterion-related validities compared to single-stimulus measures (Bartram, 2007; Salgado and Táuriz, 2014). This improvement is presumably due to reduction in response biases detrimental to validity. The psychometric problems of ipsative scores, however, should not be underestimated as they can result in spuriously significant validity coefficients (Brown and Maydeu-Olivares, 2013). It is my belief that conclusive validity evidence for forced-choice assessments can only be gained by using model-based measurement. With increasing repertoire of properly scaled forced-choice measures, there is much to look forward to in this area of research.

Recent developments in item response modeling have enabled the use of forced-choice personality assessments without the downsides of ipsative data. The potential reduction in response biases that forced-choice measures can provide, while maintaining interpersonal comparability of trait scores, is an exciting prospect for many applications in personality assessment. For instance, cross-cultural personality research, where culturally specific response sets present a challenge for score comparability, could benefit from the use of direct comparative judgments. Assessments by external raters, whether in workplace, health or education, could also benefit from the use of carefully designed forced-choice questionnaires to enhance validity by reducing rater effects such as halo and leniency/severity. Provided that appropriate methods are used to design and score such assessments, the forced-choice formats can be a viable alternative to the single-stimulus formats.

References

Andrich, D., 1989. A probabilistic IRT model for unfolding preference data. Appl. Psych. Meas. 13, 193-216.

Baron, H., 1996. Strength and limitations of ipsative measurement. J. Occup. Organ. Psych. 69, 49-56.

Bartram, D., 2007. Increasing validity with forced-choice criterion measurement formats. Int. J. Select. Assess. 15, 263-272.

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

Brown, A., Bartram, D., 2009. Doing less but getting more: Improving forced-choice measures with IRT. http://www.shl.com/assets/resources/Presentation-2009-Doing-less-but-getting-more-SIOP.pdf (accessed 28.02.14)

Brown, A., Maydeu-Olivares, A., 2010. Issues that should not be overlooked in the dominance versus ideal point controversy. Ind. Organ. Psych. 3, 489-493.

Brown, A., Maydeu-Olivares, A., 2011. Item response modeling of forced-choice questionnaires. Educ. Psychol. Meas. 71, 460-502.

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

Cheung, M.W.L, Chan, W., 2002. Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. Struct. Equ. Modeling. 9, 55-77.

Christiansen, N. D., Burns, G. N., Montgomery, G. E., 2005. Reconsidering forced-choice item formats for applicant personality assessment. Hum. Perform. 18, 267-307.

Clemans, W. V., 1966. An Analytical and Empirical Examination of Some Properties of Ipsative Measures. Psychometric Monograph 14. Psychometric Society, Richmond, VA.

Coombs, C. H., 1950. Psychological scaling without a unit of measurement. Psychol. Rev. 57, 145-158.
Cornwell, J. M., Dunlap, W. P., 1994. On the questionable soundness of factoring ipsative data: A response to Saville & Willson. J. Occup. Organ. Psych. 67, 89-100.
Drasgow, F., Chernyshenko, O. S., Stark, S., 2010. 75 years after Likert: Thurstone was right! Ind. Organ. Psych. 3, 465–476.
Feldman, M. J., Corah, N. L., 1960. Social desirability and the forced choice method. J. Consult. Psychol. 24, 480-482.
Heggestad, E. D., Morrison, M., Reeve, C. L., McCloy, R. A., 2006. Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. J. Appl. Psychol. 91, 9-24.
Hicks, L.E., 1970. Some properties of ipsative, normative, and forced-choice normative measures. Psychol. Bull. 74, 167-184.
Jackson, D. N., Wroblewski, V.R., Ashton, M. C., 2000. The impact of faking on employment tests: does forced choice offer a solution? Hum. Perform. 13, 371-388.
Luce, R. D. (1959). Individual choice behavior: A theoretical analysis. New York, NY: John Wiley.
Maydeu-Olivares, A., Böckenholt, U., 2008. Modeling subjective health outcomes: Top 10 reasons to use Thurstone's method. Med. Care. 46, 346-348.
McCloy, R. A., Heggestad, E.D, Reeve, C.L., 2005. A silk purse from the sow's ear: Retrieving normative information from multidimensional forced-choice items. Organ. Res. Methods. 8, 222-248.
Meade, A. W., 2004. Psychometric problems and issues involved with creating and using ipsative measures for selection. J. Occup. Organ. Psych. 77, 531-552.
Roberts, J. S., Donoghue, J.R., Laughlin, J. E., 2000. A general item response theory model for unfolding unidimensional polytomous responses. Appl. Psych. Meas. 24, 3-32.
Robie, C., Brown, D. J., Beaty, J. C., 2007. Do people fake on personality inventories? A verbal protocol analysis. J. Bus. Psychol. 21, 489-509.
Salgado, J.F. Táuriz, G., 2014. The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. Eur. J. Work. Organ. Psy. 23, 3-30.
Stark, S., Chernyshenko, O.S., Drasgow, F., 2005. An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The Multi-Unidimensional Pairwise-Preference Model. Appl. Psych. Meas. 29, 184-203.
Thurstone, L. L., 1927. A law of comparative judgment. Psychol. Rev. 34, 273-286.
Zinnes, J. L., Griggs, R. A., 1974. Probabilistic, multidimensional unfolding analysis. Psychometrika. 39, 327-350.