Psychological Methods

How IRT Can Solve Problems of Ipsative Data in Forced-Choice Questionnaires

Anna Brown and Alberto Maydeu-Olivares

Online First Publication, November 12, 2012. doi: 10.1037/a0030641

CITATION

Brown, A., & Maydeu-Olivares, A. (2012, November 12). How IRT Can Solve Problems of Ipsative Data in Forced-Choice Questionnaires. *Psychological Methods*. Advance online publication. doi: 10.1037/a0030641

How IRT can Solve Problems of Ipsative Data in Forced-Choice Questionnaires

Anna Brown

University of Kent

Alberto Maydeu-Olivares

University of Barcelona

Author Note

Anna Brown, School of Psychology, University of Kent, UK.

Alberto Maydeu-Olivares, Faculty of Psychology, University of Barcelona, Spain.

This research was supported by the ICREA-Academia Award and grant SGR 2009 74 from the Catalan Government, and grants PSI2009-07726 and PR2010-0252 from the Spanish Ministry of Education awarded to Alberto Maydeu-Olivares.

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

Abstract

In multidimensional forced-choice (MFC) questionnaires, items measuring different attributes are presented in blocks, and participants have to rank-order the items within each block (fully or partially). Such comparative formats can reduce the impact of numerous response biases often affecting single-stimulus items (aka, rating or Likert scales). However, if scored with traditional methodology, MFC instruments produce *ipsative* data, whereby all individuals have a common total test score. Ipsative scoring distorts individual profiles (it is impossible to achieve all high or all low scale scores), construct validity (covariances between scales must sum to zero), criterion related validity (validity coefficients must sum to zero), and reliability estimates. We argue that these problems are caused by inadequate scoring of forced-choice items, and advocate the use of item response theory (IRT) models based on an appropriate response process for comparative data, such as Thurstone's Law of Comparative Judgment. We show that by applying Thurstonian IRT modeling (Brown & Maydeu-Olivares, 2011), even existing forced-choice questionnaires with challenging features can be scored adequately and that the IRT-estimated scores are free from the problems of ipsative data.

Keywords: forced-choice format, ipsative data, multidimensional IRT, comparative judgment, Thurstonian IRT model, Thurstonian factor model

How IRT can Solve Problems of Ipsative Data in Forced-Choice Questionnaires

Assessments of personality, social attitudes, interests, motivation, psychopathology and well-being largely rely on respondent-reported measures. Most such measures employ the so-called single-stimulus format, where respondents evaluate one question (or item) at a time, often in relation to a rating scale (i.e. Likert-type items). Because the respondents rate each item separately from other items, they make *absolute judgments* about the extent to which the item describes their personality, attitudes, etc. Simple to answer and score and therefore popular with test takers and test users, the single-stimulus format makes several assumptions about the respondents' rating behaviors that are often unrealistic. For instance, the use of rating scales relies on the assumption that respondents interpret category labels in the same way. This assumption is very rarely tested in practice, but research available on the issue suggests that interpretation and meaning of response categories vary from one respondent to another (Friedman & Amoo, 1999). Furthermore, individual response styles may vary (Van Herk, Poortinga & Verhallen, 2004) so that some respondents avoid extreme categories (*central tendency* responding), whereas others prefer them (*extreme* responding). Sometimes respondents tend to agree with both positive and negative statements as presented (*acquiescence* bias). Another common problem is getting respondents to differentiate between ratings they give to single-stimulus items. When rating another person's attributes or behavior (as in the 360-degree feedback), respondents commonly give either high or low ratings on all behaviors (*halo/horn* effect) depending on whether they judge the person to score high or low on a single important dimension.

Forced-choice response formats were designed to reduce such biases. Instead of evaluating each item separately, respondents have to make *comparative judgments*, choosing

between several items according to the extent to which the items describe their preferences or behavior. M*ultidimensional forced-choice* (MFC) format questionnaires consist of blocks of two or more items intended to measure different attributes, for example:

A. I manage to relax easily
B. I am careful over detail
C. I enjoy working with others
D. I set high personal standards

Typically, respondents are asked to rank-order the items. For blocks of four or more items, respondents are sometimes asked to select the statement that best describes their behavior or preferences (“most like me”) and the statement that worst describes them (“least like me”). Examples of established and popular questionnaires using the forced-choice format are the Occupational Personality Questionnaire (OPQ32i; SHL, 2006), the Customer Contact Styles Questionnaire (CCSQ 7.2; SHL, 1997), the Gordon's Personal Profile Inventory (GPP-I; Gordon, 1993), the Survey of Interpersonal Values (SIV; Gordon, 1976), the Kolb Learning Style Inventory (Kolb & Kolb, 2005), and others.

Direct item comparison overcomes the problems with interpretation of the rating scale and categories altogether. The fact that respondents cannot endorse all items eliminates acquiescence responding (Cheung & Chan, 2002), and will typically result in a greater differentiation of ratings given to other persons. Bartram (2007) shows that when forced-choice formats are employed in assessments of workplace behaviors by line managers, where the halo effects are notoriously high, correlations with related personal attributes increase by as much as 50% in comparison to single-stimulus behavior ratings. Moreover, recent evidence shows that putting equally desirable items together in blocks may reduce the effects of *socially desirable*

responding, making the forced-choice format useful in applicant assessment contexts (Jackson, Wroblewski & Ashton, 2000; Christiansen, Burns & Montgomery, 2005).

While comparative judgments can have advantages over absolute judgments in reducing some common response biases, the use of the forced-choice format in psychological assessment has been controversial. In particular, forced-choice questionnaires have been heavily criticized because their traditional scoring produced *ipsative* data, where all respondents receive the same total number of points on the questionnaire. Evidence for the substantial psychometric challenges posed by ipsative data – including threats to score interpretation and validity – has been plentiful and ranged from pure mathematical derivations (e.g. Clemans, 1966; Dunlap & Cornwell, 1994) to sample-based empirical illustrations (e.g. Hicks, 1970; Johnson, Wood, & Blinkhorn, 1988; Tenopyr, 1988; Closs, 1996; Meade, 2004). The classical test theory (CTT) approach, which works reasonably well with Likert items, performs poorly when applied to forced-choice items. This is hardly surprising given that the implicit model underlying the CTT scoring of forced-choice tests bears no relation to the **psychological process** used in comparative judgments (Meade, 2004).

Recent developments in Item Response Theory (IRT) have opened doors to modeling the psychological process driving comparative decisions. This modeling enabled the development of scoring protocols that are suitable for use with forced-choice data. Three such approaches have been developed (Stark, Chernyshenko & Drasgow, 2005; McCloy, Heggestad & Reeve, 2005; and Brown & Maydeu-Olivares, 2011), each having distinct objectives and starting points, assuming different forced-choice designs and different properties of items used. The main objective of work by Stark, Chernyshenko and Drasgow (2005) and McCloy, Heggestad and Reeve (2005) was building **new** forced-choice tests that could successfully recover absolute trait

standings by adopting IRT-based modeling of comparative judgments. Brown and Maydeu-Olivares (2011) pursued a different goal – to introduce a response model and methods to estimate its item parameters and score individuals that could be readily applied to data collected with **existing** forced-choice questionnaires. To do so, they embed latent traits within Thurstone's (1927, 1931) Law of Comparative Judgment, giving rise to the Thurstonian IRT model, and they use a structural equation modeling (SEM) approach to estimate it. Brown and Maydeu-Olivares presented results of extensive simulation studies revealing that item parameters and individuals' scores can be estimated accurately in forced-choice designs using full rankings, and constructed a short questionnaire to illustrate their approach. Later they extended their approach to handle incomplete rankings – such as "most like me"-"least like me" response formats (Brown & Maydeu-Olivares, 2012).

While the results of these studies show that the absolute trait standings can indeed be recovered from forced-choice responses, what has been lacking is unequivocal evidence that the Thustonian IRT model applied to data at hand **solves the problems of ipsative data**. In particular, and of an utmost importance to the test users and the psychological community, what are the benefits of using this IRT-based scoring as compared to ipsative data?

The objective of this article is to illustrate how the innovations in IRT modeling and scoring of forced-choice questionnaires can enhance the quality of psychological research through a concrete example. We consider workplace assessments with the Customer Contact Styles Questionnaire (CCSQ 7.2; SHL, 1997), a popular test that yields fully ipsative data when scored traditionally. By comparing the scores derived from our Thurstonian IRT modeling approach with the ipsative scores, we aim to convince developers and users of MFC questionnaires that they should model and score these data using IRT. The article also aims to

convince the psychological community that there is nothing wrong with reasonably designed forced-choice tests, and that they provide a viable alternative to ratings – if they are modeled and scored using an appropriate IRT model.

The article is structured into five sections as follows. In the first section, we outline how classical scoring methods for forced-choice formats result in ipsative data, discuss the psychometric properties of ipsative data, and their implications for psychological assessment. In the second section, we show that the problems of ipsative data are caused by inadequate scoring of comparative judgments, and discuss recent IRT approaches to modeling comparative response processes. In the third section, we advocate the use of Thurstone's Law of Comparative Judgment as a foundation for modeling comparative decisions, and describe the Thurstonian IRT model. In the fourth section, we apply this model to the Customer Contact Styles Questionnaire and illustrate how the advocated IRT scoring overcomes the problems of ipsative data, and how the IRT-based scores enable comparison between individuals, yield undistorted estimates of construct and criterion-related validity, and facilitate estimation of measurement error. Finally, we summarize the research findings, discuss their implications for psychological assessment, and make recommendations for future research.

Problems of Ipsative Data

Data is *ipsative* when the sum of all scores obtained on the questionnaire is constant for any individual. Forced-choice questionnaires represent only one of various ways in which ipsative data can be obtained. Variations in forced-choice questionnaire design produce *fully ipsative* or *partially ipsative* scores, with partially ipsative scores being closer in their properties to *normative* scores (scores typically derived from single-stimulus formats). For ease of

exposition, we will concentrate on the most extreme and therefore the most problematic type: fully ipsative scores.

It is easy to see how this type of data comes about if we consider how the forced-choice format is conventionally scored. When a respondent is asked to perform a full ranking of items in each block, the items' inverse rank-orders (or a value derived from it through a linear transformation) are added to the corresponding scale scores. For example, in a block of three items (*triplet*), the item with the highest ranking adds 2 points to its respective scale, the lowest-ranked item adds 0 points, and the remaining item adds 1 point to its respective scale. As another example, consider the following responses to our sample block of four items (*quad*), and its corresponding classical item scores:

	Most like me	Least like me	Classical score
A. I manage to relax easily	○	○	1
B. I am careful over detail	●	○	2
C. I enjoy working with others	○	●	0
D. I set high personal standards	○	○	1

In this example using the "most"-"least" response format, the most preferred item (ranked first) adds 2 points to its respective scale, the least preferred (ranked last) adds 0 points, and the remaining items add 1 point each to their respective scales. In both examples, a constant number of points (3 in the triplet example, and 4 in the quad example) are distributed in each block. Therefore, regardless of the choices made, item scores in the block always add up to the same number, and therefore the total test score (sum of all the blocks) is the same for every individual, i.e. ipsative.

Next, we summarize the psychometric properties of ipsative data and discuss their implications for psychological assessment.

1. Relative nature of scores

Because ipsative scoring allocates the same total number of points for everyone, it is impossible to achieve all high (or all low) scale scores in a questionnaire measuring multiple attributes. Achieving a high score on one scale will inevitably imply receiving lower scores on other scales. Therefore, many have argued that ipsative scores make sense when comparing relative strength of traits within one individual, but they do not provide information on absolute (normative) trait standing, thereby making comparisons between individuals meaningless (e.g. Closs, 1996; Hicks, 1970; Johnson, Wood & Blinkhorn, 1988). For instance, Hicks (1970) discusses how a group of scientist may show a low mean score on the Aesthetic scale simply because their mean score on the Theoretical scale has to be very high. This, however, does not mean that the scientists are less aesthetic than other people are – such conclusions simply cannot be made based on ipsative data. Because people with very different absolute scores can have the same relative ordering of traits (and consequently the same ipsative scores), ipsativity can have serious implications for the interpretation of scores and selection decisions in applied settings.

2. Distorted construct validity

Because in ipsative measures the total test score (i.e. the sum of all scale scores) is a constant, it has zero variance. Therefore, all elements of the scales' covariance matrix will sum to zero (Clemans, 1966). It is easy to see that with the covariances summing to zero, the average off-diagonal covariance is a negative value, and the same is true for correlations. In fact, for an ipsative test consisting of k scales with equal variances, the average correlation among the scales must be

$$\bar{\rho} = -1/(k-1). \quad (1)$$

That is, when ipsative scale variances are approximately equal, the average correlation among ipsative scores in a 3-scale test will tend to −0.5, regardless of the 'true' relationships among the attributes measured by the scales. Thus, with a few scales, the distortion to the relationships between constructs may be very substantial, particularly when the true trait scores are supposed to correlate positively.

Because any scale in an ipsative measure can be perfectly determined from the remaining scales, the covariance matrix of ipsative scores is of a reduced rank; its rank is k -1 where k is the number of measured scales (Clemans, 1966). Therefore, one of the eigenvalues must be zero and maximum likelihood factor analysis cannot be applied (Dunlap & Cornwell, 1994). Principal components analysis can be performed on ipsative scores; however, it may yield quite different results from the analysis of corresponding normative scores. Ipsativity produces artifactual bipolar components, where traits that would typically belong to different components in normative data are contrasted to each other (Cornwell & Dunlap, 1994; Baron, 1996). Ipsativity, therefore, clearly compromises construct validity of forced-choice questionnaires.

3. Distorted criterion-related validity

Because all ipsative scale scores have to sum to a constant, their covariances with any external measure will sum to zero (Clemans, 1966; Hicks, 1970). This means that criterion-related validity of an ipsative instrument will be distorted because any positive covariances with the external variable have to be compensated by some negative covariances, and vice versa. Thus, if the instrument’s scales are expected to co-vary with the criterion mostly positively (or negatively), the ipsative constraint will distort these relationships, creating spurious compensatory covariances (Johnson, Wood & Blinkhorn, 1988).

4. Distorted reliability estimates

It is generally agreed amongst researchers that the forced-choice format distorts conventional measures of reliability, but in which direction and to what degree appears to be highly dependent on specific conditions. For instance, Baron (1996) argued that coefficient alpha underestimates internal consistency reliabilities in MFC questionnaires measuring a large number of scales. On the contrary, Johnson, Wood and Blinkhorn (1988) suggested that reliabilities of ipsative tests overestimate the actual scale reliabilities.

The general problem with using conventional reliability statistics is that ipsative scoring violates most assumptions these statistics rely on. In particular, the assumption of consistent coding (i.e. for positively keyed items, higher item sores correspond to higher true scores on the traits) is violated in ipsative scoring. When giving the top rank to one item, the respondent does so not because he/she agrees with the item, but because he/she agrees with it *more* than with the other items in the block. Therefore, items measuring the same trait might receive the highest rank (maximum number of points) in one block, and the lowest in another. Thus, a response that might be consistent with the true scores will appear to be inconsistent from the item coding perspective.

Another basic assumption underlying reliability statistics – that of independent errors – is also violated in ipsative scoring because items within a block are **not assessed independently**. Rather, a forced-choice item is ranked according to the degree it is preferred to other items, creating mutual dependencies between **all** items in the block (and all traits measured by those items). Some authors have argued that due to local dependencies between items within blocks, the concept of random error is dubious in forced-choice data (Cornwell & Dunlap, 1994), therefore appropriateness of other reliability statistics is also doubtful (Meade, 2004). Due to

these limitations, the actual measurement precision of forced-choice questionnaires remains unknown.

Increasing the Number of Measured Scales as the Classical Solution to the Problems of Ipsativity

Within a CTT framework, the impact of ipsativity can be alleviated somewhat by using a large number of scales. Baron (1996) argued that with a large number of relatively independent scales, only a very low proportion of respondents will have most of their true scale scores all high or all low – and therefore only a few profiles will be badly distorted by the ipsative centering on the mean score. Also, several authors have shown that in carefully designed forced-choice questionnaires with 30 or more measured traits, the ordering of people on each trait largely corresponds to their normative ordering (e.g., Baron, 1996; Karpatschof & Elkjaer, 2000), and therefore the standardizing of ipsative scores is appropriate and inter-individual comparisons can be performed meaningfully.

Because the average correlation in Equation (1) approaches zero as the number of scales increases, with 30 scales the average correlation would be approximately −0.03, allowing for a wide range of both negative and positive correlations between scales (Baron, 1996). As a result, principal components analysis of ipsative scores may also be more interpretable, although it does not produce the same results as analysis of normative scores. In summary, increasing the number of scales somewhat alleviates the problems associated with ipsativity but does not solve them.

Item Response Modeling of Forced-Choice Questionnaires

From the above discussion, it is apparent that classical scoring is not well suited to forced-choice responses, because it treats relative rankings as if they were absolute ratings. When the wrong measurement model is adopted from the start, it is not surprising that the resulting

scores yield unexpected properties. To solve the problems of ipsative data, one needs to depart radically from classical scoring schemes, and adopt a measurement model that reflects the decision process that respondents use when answering forced-choice items (Meade, 2004). Several such models have been proposed recently.

Stark (2002) substantially advanced the field by explicitly providing an IRT model for multidimensional pairwise comparisons (i.e. forced-choice blocks of two items), and by using Bayes modal estimation for the latent traits. The Multi-Unidimensional Pairwise-Preference (MUPP) model, further described by Stark, Chernyshenko and Drasgow (2005), approximates the probability of preferring one item to another by the joint probabilities of accepting one item and rejecting the other. The acceptance and rejection of individual items are assumed independent events, and their probabilities are described as unidimensional IRT functions. Stark and colleagues allow a broad class of items to be used in pairwise comparisons; thus, they go beyond the usual *dominance* items described by s-shaped item response functions, and assume items with bell-shaped *ideal point* response functions (Stark et al., 2006). They use an ideal point model (namely, the generalized graded unfolding model or GGUM; Roberts, Donoghue & Laughlin, 2000) as a basis for IRT calibration of individual items. To date, the MUPP model has been used successfully to create new forced-choice questionnaires with items presented in pairs, and to recover individuals' absolute trait scores, after item parameters have been estimated from single stimulus trials (e.g. Chernyshenko et al., 2009).

McCloy, Heggestad and Reeve (2005) sketched a method for creating forced-choice questionnaires using an implicit IRT model. Their approach draws on specific item properties – namely, equally discriminating items with ideal-point response functions must be used. The model relates the likelihood of preferring one item to another to the relative distances between

the item locations and the respondent's trait scores. Thus, the respondent is more likely to prefer the item located closer to one's own standing on the respective trait, then the item located further from one's own standing on the trait. Assuming known item parameters, a questionnaire can be assembled from blocks of items with locations that vary across the traits continuum. Furthermore, McCloy and colleagues show that by combining items with different locations, the absolute trait levels for an individual can be recovered.

These innovations provide a way forward in creating forced-choice questionnaires yielding normative measurement. Unfortunately, they cannot be used to solve the problems of ipsative data in existing forced-choice questionnaires, because neither Stark and colleagues nor McCloy and colleagues described how to estimate the IRT item parameters; instead, they assume that these parameters are known.

A solution to the problem of modeling forced-choice data so that the model parameters can be estimated came from structural equation modeling. Chan and Bentler (1998; see also Chan, 2003) and Maydeu-Olivares (1999) proposed different SEM methods for estimating a factor analysis model embedded within Thurstone's (1927, 1931) Law of Comparative Judgement, and applied these methods to model responses to a single ranking task (a single forced-choice block). Maydeu-Olivares and Böckenholt (2005) provided a detailed overview of Thurstonian scaling methods –including factor analytic models– and their estimation using SEM methods. Maydeu-Olivares and Brown (2010) pointed out that Thurstonian factor analytic models underlying a single ranking task can be reformulated as IRT models so that respondents' scores on underlying dimensions can be estimated. Moving from a single block to multiple blocks, Brown and Maydeu-Olivares (2011) finally provided the first feasible IRT model suitable for modeling data gathered using existing MFC questionnaires. Their modeling framework can

be used with any number of blocks, any number of items per blocks (items presented in pairs, triplets, quads, etc.), correlated latent traits, etc.

In the next section, we describe the Thurstonian IRT approach to modeling forced-choice questionnaire data, concentrating on the psychological process involved in comparative judgments and the model's value in applied settings – the aspects of the model that have not been discussed before. The technical detail of the model is described elsewhere – for the full description of model features and simulation studies, we refer the reader to Brown and Maydeu-Olivares (2011), and for a step-by-step guide on how to fit the model using Mplus to Brown and Maydeu-Olivares (2012).

Modeling Forced-Choice Decisions Using Thurstone's (1927) Law of Comparative Judgment

Psychological models suitable for comparative responses have existed for a long time, and they are well known. Simply, they have not been applied to the problem of modeling responses to forced-choice items. One of the oldest, and probably the most influential, psychological framework for modeling comparative data was proposed by Louis Thurstone. His model relies on two key notions. First is the notion of an unobserved *psychological value* or *utility* (Thurstone, 1929) underlying each item, also referred to in his earlier work as *discriminal process* (Thurstone, 1927). Second is the notion of utility maximization, which implies that when confronted with a choice between two items, respondents will choose the item with the highest psychological value (utility).

Universally applicable to any value judgments made about objects or ideas, the psychological value describes "the affect that the object calls forth". The psychological value for an object varies across individuals, and similarly it varies across objects within an individual. As

such, it can be placed on a *psychological continuum* (Thurstone, 1929). For any given object, Thurstone further assumed that its psychological value is normally distributed across individuals.

Without loss of generality, we can focus on personality items, but the same reasoning readily applies to attitudes, motives, interests, patient-reported outcomes, etc. In the case of personality items such as "I am careful over detail", we suggest that the psychological value can be taken as one of *likeness* (similarity or representation) between the behavior described in the item and the respondent's own behavior, as perceived by the respondent. Those behavioral statements that the respondent judged to be very much representative of their own behavior (or to be like self) are placed towards the positive end of the likeness continuum, and consequently are given a positive value; those statements that were judged unrepresentative of their own behavior (or **un**like self) are placed towards the negative end.

Because the likeness values for different statements are placed on the same continuum, they can be compared to each other. For instance, "I am careful over detail" may be judged to be very much like the respondent and "I manage to relax easily" to be not very much like him/her. If we now ask the respondent directly which statement out of the two is *more like* him/her, the answer will require a *comparison* of the psychological values by the respondent. In our example, the respondent should prefer "I am careful over detail" to "I manage to relax easily" because his/her likeness value for the former is greater than for the latter.

This process is described by Thurstone's Law of Comparative Judgment (1927), which explains preference decisions by the relative psychological values of the objects or ideas under comparison. According to Thurstone, item *i* is preferred to item *k* if the psychological value of *i* (t_i) is greater than the value of *k* (t_k). Let y_l^* denote the difference in psychological values of the two items

$$y_l^* = t_i - t_k, \qquad l = \{i, k\}. \tag{2}$$

We can now code the outcome of comparison $\{i, k\}$ as 1 if item i is preferred to k, and 0 otherwise, and relate it to the latent difference of values using Thurstone's simple law:

$$y_l = \begin{cases} 1 & \text{if} \quad y_l^* \geq 0 \\ 0 & \text{if} \quad y_l^* < 0 \end{cases}. \tag{3}$$

Now, what happens when three, four or more items are compared? Thurstone (1931) suggested coding any ranking of n objects using $n(n\text{-}1)/2$ binary outcome variables, i.e. directional pairwise comparisons. For instance, to rank order four items A, B, C and D, one needs to make six pairwise comparisons: item A has to be compared with B, C and D; item B compared with C and D; and item C compared with D. The item ranked first has to be preferred in every comparison involving that item, and the item ranked last not preferred in any comparison. For example, the ordering {B, A, D, C} is equivalent to the six pairwise judgments:

Ranking				Binary Outcomes					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
2	1	4	3	0	1	1	1	1	0

In the case when the block consists of more than three items, and only the "most" and "least" preferred items are requested, some binary outcomes will be unknown. For instance, with four items per block, as in our previous example, the outcome of the comparison between items A and D is unknown:

Partial ranking				Binary Outcomes					
A	B	C	D	{A,B}	{A,C}	{A,D}	{B,C}	{B,D}	{C,D}
	most	least		0	1	.	1	1	0

In forced-choice questionnaires using ranking blocks with no repeated items, no inconsistencies in pairwise judgments may arise and binary outcomes are completely determined by the differences in item utilities. When inconsistencies in the pairwise judgments are possible, for instance when sets of paired comparisons with repeated items are presented to the respondent instead of ranking blocks, then an error term needs to be added to (2). See Maydeu-Olivares and Böckenholt (2005) for further details.

Thurstonian Factor Models

Thurstone (1927) described several special cases of his model. The best known of such models is Case V, in which the psychological values of objects under comparison are assumed independent with equal variances. If the objects under comparison are personality items, and if they share common variance (the psychological traits they measure), a model that assumes no common variance, such as the Case V model, is inappropriate. Rather, a factor-analytic model underlying the items' psychological values is called for. In such a model, underlying common factors such as personality dimensions (e.g Emotional Stability, Conscientiousness) may be measured.

Thurstonian factor analytic models (or *Thurstonian factor models*) are less well known than classical models such as Case V because their estimation has only recently become feasible (Maydeu-Olivares & Böckenholt, 2005; Tsai & Böckenholt, 2001). Thurstonian factor models are similar to second-order factor analysis models with binary outcomes. Every binary outcome y_l of comparison $\{i, k\}$ is determined by the unobserved difference of two utilities t_i and t_k, as per Equation (2). The latent utilities (first-order factors), in turn, depend on psychological attributes (second-order factors). Maydeu-Olivares and Böckenholt (2005) show how to embed these models within a familiar SEM framework so that they can be easily estimated and tested.

The Thurstonian IRT Model for Forced-Choice Questionnaires

The second-order Thurstonian factor models are popular in marketing applications, where psychological values associated with different services and goods at the population level are the focus of research. In personality and other person-centric research, however, the psychological values of items (first-order factors) are not of interest. Rather, interest lies in estimating second-order factors (e.g. personality traits, motivation factors, interests, etc.). Unfortunately, when applied to ranking data, the Thurstonian factor model does not allow scoring the respondents on the latent traits due to the zero error variance of the outcome comparison variables – notice that there is no error term in Equation (2). To overcome this problem, and to bypass estimation of psychological values when these are not of interest, Brown and Maydeu-Olivares (2011; Maydeu-Olivares & Brown, 2010) introduced the Thurstonian IRT model, which is a reparameterization of the Thurstonian factor model as a first-order factor model (i.e., both models are mathematically equivalent). Next, we provide a brief account of the Thurstonian IRT model. Further technical details on these models and their relationship with the Thurstonian factor models can be found in Brown and Maydeu-Olivares (2011).

The Thurstonian IRT model is a first-order model that links the binary outcomes to the traits directly, by substituting the latent utilities with linear functions describing their relationships with the underlying traits $\boldsymbol{\eta}$. We assume an independent clusters structure (McDonald, 1999), so that each item is underlain by only one factor. Thus, the utility of item *i* is

$$t_i = \mu_i + \lambda_i \eta_a + \varepsilon_i \,, \tag{4}$$

where μ_i is the utility mean, λ_i is the loading on the measured attribute η_a, and ε_i is the unique factor (uniqueness). Inserting Equation (4) into Equation (2) we find that for items *i* and *k* measuring attributes η_a and η_b respectively, the latent difference y_l^* is

$$y_l^* = t_i - t_k = -\gamma_l + \left(\lambda_i \eta_a - \lambda_k \eta_b\right) + \left(\varepsilon_i - \varepsilon_k\right), \quad (5)$$

where $\gamma_l = -\left(\mu_i - \mu_k\right)$ is a threshold parameter replacing the difference of utility means. It follows from (5) and the threshold process (3) that the conditional probability of preferring item *i* to item *k* depends on the interplay between the **two factors** underlying the items' utilities:

$$P_l\left(y_l = 1 \middle| \eta_a, \eta_b\right) = \Phi\left(\frac{-\gamma_l + \lambda_i \eta_a - \lambda_k \eta_b}{\sqrt{\psi_i^2 + \psi_k^2}}\right), \quad (6)$$

where $\Phi\left(x\right)$ denotes the cumulative standard normal distribution function evaluated at *x*, and ψ_i^2 and ψ_k^2 are unique variances of the two utilities, i.e., $\psi_i^2 = \mathrm{var}\left(\varepsilon_i\right)$. As we would expect, for positively keyed items, the probability of preferring item *i* to item *k* increases when the score on the trait underlying item *i* increases and the score on the trait underlying item *k* decreases.

Equation (6) describes the item response function of the binary outcome y_l using a *threshold/loading* parameterization. As this equation reveals, the Thurstonian IRT Model is a two-dimensional normal ogive IRT model with some special features. First, the uniqueness of every binary outcome is structured so that its variance equals to the sum of unique variances of the two utilities. This means that whenever the same item is involved in more than one pairwise comparison (as in ranking blocks consisting of three or more items ($n \geq 3$)), error variances of binary outcomes involving item *i* will have a shared part, with variance ψ_i^2, so that local dependencies exist among them. Second, within blocks of three or more items ($n \geq 3$), all binary outcomes involving item *i* will share the same factor loading λ_i. To summarize, the Thurstonian IRT model is the extension of the normal ogive model to items presented in blocks. As such, when blocks consist of three or more items, it involves within-block patterned dependencies.

Estimation of Model and Person Parameters

Because Thurstonian IRT models are first-order factor models with binary outcomes and some special features, they can be estimated with general-purpose software. We use Mplus (Muthén & Muthén, 1998-2010), which can handle dichotomous observed variables and can easily incorporate the necessary parameter constraints. Brown and Maydeu-Olivares (2012) provide a step-by-step tutorial on model specification, identification and estimation using Mplus; they also supply a macro writing Mplus syntax for any forced-choice design.

Limited information methods (i.e. estimation based on tetrachoric correlations among the binary outcomes) are recommended, as they are computationally efficient and unaffected by the number of latent traits –which can be large in forced-choice questionnaires. The estimation is fast; however, current computing capabilities prevent from computing goodness of fit indices and standard errors in questionnaires with more than 100 or so items.

After the item and model parameters have been estimated, respondents can be scored on the latent traits. Maximum likelihood (ML) estimation, the mode of the posterior distribution (MAP), or the mean of the posterior distribution (EAP) may be used (Embretson & Reise, 2000) to find the most likely combination of scores $\hat{\boldsymbol{\eta}} = \left(\hat{\eta}_1, \hat{\eta}_2, ...\hat{\eta}_d\right)'$. Here, we estimate the respondents' traits levels by the MAP method, which is well suited to multidimensional IRT applications, as its computational burden does not depend on the number of latent traits. The MAP method is conveniently implemented in M*plus* as an option within the estimation process. For further technical detail related to model and person parameter estimation, we refer the reader to Brown and Maydeu-Olivares (2012).

Precision of Measurement and Test Reliability

In IRT, unlike in classical scoring, the precision of measurement is not the same for all respondents but depends on their trait scores. Brown and Maydeu-Olivares (2011) apply general multidimensional information theory (Reckase, 2009; Ackerman, 2005) to provide item information functions (IIF) needed to compute the standard errors of estimated scores for the Thurstonian IRT model.

For computing the item information, it is convenient to write the item characteristic function given by Equation (6) in *intercept/slope* parameterization

$$P_l\left(y_l = 1 \middle| \eta_a, \eta_b\right) = \Phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right), \tag{7}$$

by letting

$$\alpha_l = \frac{-\gamma_l}{\sqrt{\psi_i^2 + \psi_k^2}}, \quad \beta_i = \frac{\lambda_i}{\sqrt{\psi_i^2 + \psi_k^2}}, \quad \beta_k = \frac{\lambda_k}{\sqrt{\psi_i^2 + \psi_k^2}}. \tag{8}$$

Then the information provided by one binary outcome about traits η_a and η_b is, respectively:

$$\mathcal{I}_l^a\left(\eta_a, \eta_b\right) = \frac{\left[\beta_i - \beta_k \operatorname{corr}\left(\eta_a, \eta_b\right)\right]^2 \left[\phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right)\right]^2}{P_l\left(\eta_a, \eta_b\right)\left[1 - P_l\left(\eta_a, \eta_b\right)\right]}, \tag{9}$$

$$\mathcal{I}_l^b\left(\eta_a, \eta_b\right) = \frac{\left[-\beta_k + \beta_i \operatorname{corr}\left(\eta_a, \eta_b\right)\right]^2 \left[\phi\left(\alpha_l + \beta_i \eta_a - \beta_k \eta_b\right)\right]^2}{P_l\left(\eta_a, \eta_b\right)\left[1 - P_l\left(\eta_a, \eta_b\right)\right]}, \tag{10}$$

where $\phi(x)$ denotes the standard normal density function evaluated at *x*.

The total information about trait η_a is a sum of all information functions from binary outcomes independently contributing to the measurement of the trait. However, we know that within ranking blocks of three or more items, structured dependencies exist between the error terms of the binary outcomes. It has been shown that the test reliability is overestimated only very slightly when these within-block local dependences are ignored[1] (Maydeu-Olivares &

Brown, 2010). We therefore recommend that in applications researchers make a simplifying assumption of local independence, to enable straightforward computation of the total trait information.

In addition to the information provided by item responses, Bayesian scoring methods such as MAP contribute information from a prior distribution of the latent traits (multivariate normal distribution with covariance matrix $\mathbf{\Phi}$). Thus, the *posterior information about trait* η_a is given by

$$\mathcal{I}_P^a(\boldsymbol{\eta}) = \sum_l \mathcal{I}_l^a(\boldsymbol{\eta}) + \varpi_a^a, \tag{11}$$

where ϖ_a^a is the diagonal element of the inverted trait covariance matrix $\mathbf{\Phi}^{-1}$ related to the dimension of interest, η_a. Finally, the standard error of the MAP-estimated score $\hat{\eta}_a$ is given by

$$SE(\hat{\eta}_a) = \frac{1}{\sqrt{\mathcal{I}_P^a(\hat{\boldsymbol{\eta}})}}. \tag{12}$$

While the score precision varies for each respondent, providing a summary index can be useful for comparison with classical test reliability statistics, and for estimating expected levels of recovery of the true latent trait. However, the main challenge is to summarize multidimensional information for a questionnaire with high dimensionality. Since in forced-choice questionnaires items from the focus trait are compared with items from many (often all) other traits, the information in the direction of the target trait is conditional on different traits. To summarize such contributions for all values of all traits, it is necessary to consider a grid with the number of dimensions corresponding to the number of measured scales. Such a grid for a large number of dimensions would consist of millions of points, making computation of population

summary indices such as *marginal reliability* (Green, Bock, Humphreys, Linn & Reckase, 1984) infeasible.

We recommend a sample-based approach to compute reliability for forced-choice tests whereby the trait estimates for each person, for example MAP point estimates $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1, \hat{\eta}_2, ...\hat{\eta}_d)'$, provide a sample of points on the multidimensional grid. The reliability index based on the estimated scores for a sample is referred to as *empirical reliability* (Du Toit, 2003). It is the ratio of the true score variance (observed variance minus error variance) to the observed score variance for the sample:

$$\rho = \frac{\sigma_P^2 - \bar{\sigma}_{error}^2}{\sigma_P^2}, \tag{13}$$

where the observed score variance σ_P^2 is simply the variance of the estimated MAP score. Estimation of empirical reliability proceeds as follows. First, for each binary outcome, information values in two relevant directions (η_a and η_b) are computed for a particular set of MAP estimates, $\hat{\boldsymbol{\eta}}$. To obtain the test information for one trait, scalar values from all pairs contributing to the measurement of that trait are summed, and the posterior information $\mathcal{I}_p^a(\boldsymbol{\eta})$ is computed by adding the prior information, as given by Equation (11). The squared standard error of measurement is computed as the reciprocal of the posterior test information (or the square of the MAP standard error as per Equation (12)), for each estimated MAP score in the sample. Finally, the sample error variance $\bar{\sigma}_{error}^2$ for the trait is computed by averaging the squared standard errors across all respondents.

It is important to emphasize that because the classical concept of test reliability has no direct correspondence in IRT, any estimate of reliability obtained from the test information is

only an approximation. Since the reliability will vary for different levels of the latent trait, single indices would be more descriptive of the sample as a whole when the test information function is relatively uniform.

An Illustration Using the Customer Contact Styles Questionnaire

In this section, we illustrate how the Thurstonian IRT model may be applied to a forced-choice questionnaire in order to solve the ipsativity problems caused by the use of classical scoring procedures. To this end, we examine the Customer Contact Styles Questionnaire (CCSQ version 7.2, published by SHL), a popular questionnaire used for workplace assessments for customer service and sales roles. The CCSQ was chosen for this illustration because it is a classic MFC questionnaire producing fully ipsative data.

The Instrument

The CCSQ measures 16 job-related dimensions covering a wide range of behavioral styles, with a strong emphasis on achievement motivation (SHL, 1997). The number and nature of constructs assessed by the CCSQ is typical for a workplace questionnaire, and is comparable to another IRT-based forced-choice application, the MDPP CAT (Stark, Chernyshenko, Drasgow & White, in press), which measures 15 traits. Short scale descriptions for the CCSQ can be found in the Appendix. The number of items measuring each dimension varies from seven to ten, with 128 items in total. Items are presented in 32 blocks of four statements. All statements are positively worded and keyed. Here is a sample block:

I am the sort of person who…

A. generates imaginative solutions
B. easily forgets unfair criticism
C. needs to beat the opposition

D. is eager to help others out

For each block the respondents have to rate all four statements on a 5-point Likert scale (from ‘strongly disagree’ to ‘strongly agree’), and then select one item that is ‘most like’ them and one ‘least like’ them. Thus, the questionnaire combines both single-stimulus and forced-choice formats in one. As such, this questionnaire is ideally suited for this illustration as it enables comparison of results for rating and ranking formats using both classical and IRT procedures. On the other hand, the questionnaire has challenging features – it employs large forced-choice blocks, and the ‘most’-‘least’ response format that yields incomplete rankings.

Data Description

Two datasets were used: a calibration sample and a validation sample. For calibration, the CCSQ UK standardization sample was used. These data were collected in 2001 using a paper-and-pencil supervised administration. The sample was drawn from nine different organizations in industry, commerce and the public sector. Out of $N = 610$ respondents, 255 were job applicants, and 355 respondents completed the questionnaire in return for feedback. Sixty-one percent were male. Most respondents were currently working in sales (61%) and customer service roles (26%). The average age was 33 years. All investigations that follow, except studies of criterion-related validity, are performed with the standardization sample.

Because we wished to provide an illustration with validity coefficients involving performance-related criteria and this was not available in the calibration data, we used a second data set collected from a telecommunications company in 2006. The sample consisted of $N = 219$ call center operators, 46% were male. Age ranged from 19 to 40 years, with the mean 27.7 and standard deviation 5 years. The criterion was an incentive bonus the job incumbents received within the same year that the CCSQ data was collected. The bonus had been awarded based on

various indicators of the operators' job performance and represented a continuous variable (monetary value) distributed close to normal but positively skewed (skewness = 0.70, se = 0.16). Because this dataset is small, we used it as validation sample. That is, item parameters estimated with the standardization sample were used to compute IRT scores for individual respondents in this validation sample.

Scoring Procedures

Four sets of scores were computed for each sample:

a) *classical single-stimulus* (SS-CTT), *normative*: sum scores for the single-stimulus ratings;
b) *classical forced-choice* (FC-CTT), *ipsative*: sum scores for the forced-choice rankings;
c) *single-stimulus IRT* (SS-IRT): IRT scores for the single-stimulus items;
d) *forced-choice* IRT (FC-IRT): IRT scores for the forced-choice items.

The classical normative scores were obtained by summing the single-stimulus ratings (coded from 1 to 5), and the classical *ipsative* scores by summing the forced-choice rankings (coded 2 for 'most', 0 for 'least' and 1 for intermediate choices). To enable meaningful comparisons of these classical sets with each other and with the IRT-based sets of scores, we standardized each scale score by its mean and standard deviation across respondents, as is routinely done in assessment applications.

For IRT scoring of single-stimulus items in (c), we used the popular graded response model (Samejima, 1969) fitted to each scale separately. The Thurstonian IRT model was used in (d), where a structure with 16 correlated latent factors was fitted to the binary outcomes of pairwise comparisons. In total, 196 binary outcomes were generated from 32 blocks of four items. Given the large model size, neither standard errors nor goodness of fit indices could be computed for the Thurstonian IRT model given current computational capabilities.

The same parameter estimation method and scoring method were used in (c) and (d). Item parameters were estimated from tetrachoric correlations (therefore using limited information estimation) with the unweighted least squares (ULS) method as implemented in Mplus. The maximum a posteriori (MAP) method was used to obtain respondents' scores on the 16 measured traits, also using Mplus.

Note on Limited Information Estimation when only Partial Rankings are Available. Because the partial ranking format with 'most' – 'least' alternatives is used in our CCSQ application, one out of six binary outcomes per block is not known. The mechanism for missing data in this case is *missing at random* (MAR), but not missing completely at random (MCAR), because the pattern of missing outcomes (i.e. which outcome out of six will be missing) is determined by the observed outcomes. Missing outcomes present the stiffest challenge to estimating the Thurstonian IRT model, because MCAR data is required for correct estimation of item parameters by limited information methods (Asparouhov & Muthén, 2010). To overcome this problem, Brown and Maydeu-Olivares (2012) suggested resorting to multiple imputation (MI) methods, also implemented in Mplus. They ran a series of simulations examining the degree of agreement between true and estimated item parameters for a different number of multiple imputations. Their results suggested that as few as 10 multiple imputations provide sufficiently accurate estimation of item and other model parameters. Following this recommendation, we generated 10 datasets with imputed values, estimated the CCSQ Thurstonian IRT model parameters for each of the imputed datasets, and then averaged the obtained estimates to produce the final model parameters. Using these model parameters, persons' parameters (trait scores) were estimated using the original dataset with missing responses, not the imputed datasets.

Results

Individual Profiles. Since the four kinds of scores considered (normative, ipsative, and the IRT-based single-stimulus and forced-choice) are on the familiar *z*-scale, we can compare these sets of scores for the same individual. Figure 1 provides four profiles (one for each scoring procedure) for an individual from the Standardization sample, who completed the CCSQ in return for feedback. As can be seen, this profile was dominated by below average scores, which was clearly reflected by the single-stimulus scores (both classical and IRT). However, the classical ipsative scores failed to reflect the overall negative location of the profile, pulling most scales towards positive values to compensate for several very low scale scores. The IRT forced-choice profile was free from this distortion and was very similar to the single-stimulus profiles.

Insert Figure 1 about here

More generally, we can compute the *average profile* score across all 16 dimensions for each individual, one for each scoring method. Figure 2 depicts the distribution of average profile scores for each of the four scoring methods. As can be seen in this figure, ipsative scoring prevented from observing all-high or all-low profiles. Rather, the average *z*-score ranged only between $z = -0.13$ and $z = 0.12$ (SD = 0.04). In contrast both classical and IRT scoring of single-stimulus responses yielded some all-high or all-low profiles. The average for classical normative scores ranged from $z = -1.56$ to $z = 1.44$ (SD = 0.51), for SS-IRT it ranged from $z = -1.26$ to $z = 1.49$ (SD = 0.47). IRT scoring of forced-choice data yielded profiles with similarly varied average *z*-scores, ranging from -1.22 to 1.03 (SD=0.40). Clearly, the IRT forced-choice scores allowed variability in *absolute* locations of the personality profiles, and it was possible to obtain high/low scores on all scales.

Insert Figure 2 about here

To investigate the similarity of profiles across scoring methods further, we computed Mahalanobis distances between profiles based on single-stimulus and forced-choice responses. The Mahalanobis distance is a measure of the distance between two points in a multidimensional space (here, between two sets of 16 scores for the same individual), taking into account the non-orthogonal nature of the axes (the 16 personality traits). Arbitrarily, estimates of latent correlations between the 16 traits based on the single-stimulus responses were used as measures of correlations between the axes. The distances between SS-IRT and FC-IRT profiles ranged from 1.50 to 6.56 (median 3.02, mean 3.15 and standard deviation 0.83) and were distributed as shown in Figure 3. It can be seen that these distances were smaller than the distances between classical ipsative and normative scores, which ranged from 1.80 to 8.02 (median 3.91, mean 4.02 and standard deviation 1.07). We conclude that the IRT scoring brought the individuals' forced-choice profiles closer to their single-stimulus profiles.

Insert Figure 3 about here

Ordering of Respondents: Correlations Across Scores. Despite the distortion to individual profiles, ordering of respondents based on their ipsative scores was quite similar to their normative ordering. Table 1 shows that correlations between the two CTT scale scores (normative and ipsative) ranged from 0.50 to 0.73, with a median of 0.68. Correlations between the two IRT-based scores ranged from 0.52 to 0.79, with a median of 0.70. Therefore, in this application the ordering of respondents across formats was only marginally more similar when IRT scoring was used rather than classical scoring.

We do know from the previous section, however, that IRT scoring brought the individual forced-choice profiles closer to the respective single-stimulus profiles, and therefore IRT scoring must have changed the ordering of respondents compared to ipsative scoring. Indeed, Table 1 shows that cross-method correlations for the forced-choice scores (IRT versus CTT) ranged from 0.83 to 0.91, median 0.88. These far from trivial differences in ordering of individuals based on forced-choice responses were in contrast with nearly perfect correlations between single-stimulus scores, derived by IRT and CTT scoring (ranging from 0.97 to 0.99, median 0.98). Clearly, the greatest change introduced by the IRT scoring of forced-choice data was concerned with **systematic re-ordering** of respondents based on whole profiles, not merely on individual scales.

Insert Table 1 about here

Reliability estimates. Table 2 provides estimates of reliability for the four sets of scores, computed as coefficient alpha for the classical scores, and as empirical reliability using Equation (13) for the IRT-based scores. Reliability estimates for the SS-CTT (normative) scores obtained using coefficient alpha ranged from 0.78 to 0.91, median 0.84. Estimates of empirical reliability for the IRT single-stimulus scores ranged from 0.76 to 0.91 (median 0.85), and were very close to alphas for the CTT counterparts.

The reliability estimates (coefficient alpha) for the ipsative scores ranged from 0.57 to 0.80 with a median of 0.72 (see Table 2). They were substantially lower than the alphas obtained for the classical normative scores, and similar in magnitude to the cross-format correlations between the ipsative and the normative scores. The empirical reliabilities for the IRT forced-choice scores ranged from 0.72 to 0.89 with a median 0.80, higher than estimates for their CTT counterparts[2] (ipsative scores). The median difference between the forced-choice IRT and CTT

reliability estimates was 0.10, with four scales showing virtually no difference, and the scale Results Orientated reaching a large difference of 0.27. Some increase in reliabilities was to be expected, since precision of scores must improve when an appropriate model is used for scoring. Furthermore, Bayesian MAP estimation with a multivariate normal prior increased precision by "borrowing strength" from related traits.

While the improvement in forced-choice reliability compared to ipsative scoring was impressive, reliabilities of the IRT single-stimulus scores were still higher (median difference 0.04). One reason is that the five-point rating scale provided four pieces of information for each single-stimulus item, while comparisons between each forced-choice item and three other items in the block (binary outcomes) provided only three pieces of information. Furthermore, for some items the number of binary outcomes may have been reduced by one when the item was not selected as 'most' or 'least' (see our discussion on missing data). Another reason was that the forced-choice CCSQ is not optimally designed from an IRT perspective to maximize measurement precision. As Brown and Maydeu-Olivares (2011) show, MFC questionnaires measuring positively correlated traits with positively keyed items have a reduced ability to recover absolute trait standings of individuals, which follows directly from analysis of IRT information functions. Thus, it is not surprising that in the CCSQ questionnaire with the average correlation between normative scales being 0.21, and all items being keyed positively, the reliabilities of the IRT forced-choice scores were not as high as they could have been if the questionnaire development had been informed by IRT modeling.

Insert Table 2 about here

Construct Validity. While the classical normative scores inter-correlated positively on average ($r = 0.22$), the average correlation among classical ipsative scores was $r = -0.07$, as dictated by Equation (1). In contrast, the average correlation among IRT forced-choice scores was $r = 0.12$, much more similar to the normative average (average correlation for IRT single-stimulus scores was $r = 0.21$).

Principal component analysis (PCA) was performed to explore sources of common variance in the 16 CCSQ traits produced by the four scoring methods. Since ipsative data produces a non-invertible covariance matrix, factor analysis cannot be applied to it. Moreover, factor analytic methods introduce the concept of error, and many authors have argued that it is not clear what error means when the data is ipsative (Hicks, 1970; Meade, 2004). Despite being a simple data reduction technique, PCA is still useful for direct comparison of principal sources of variance and illustrating similarities and differences between the four sets of CCSQ scores.

In all cases, a solution with four principal components was deemed most appropriate. An oblique solution was sought every time, using the direct oblimin rotation. Table 3 provides the solution for the CCSQ classical single-stimulus scores. As can be seen in this table, the four components were labeled as ‘Conscientiousness’, ‘Dominance’, ‘Agreeableness’, and ‘Adaptability & Dynamism’, explaining 58.3% of the total variance. A virtually identical solution was obtained for IRT single-stimulus scores, with four components accounting for 58.4% of the variance (also given in Table 3).

Table 4 provides the solution for the classical ipsative scores. The retained four components accounted for just over 50% of the variance. They were “contrast” components as typically found in ipsative data: ‘Conscientiousness versus Creativity’, ‘Drive versus Agreeableness’, ‘Social Adjustment versus Analysis’, and ‘Adaptability versus Influence’. Thus,

each component had several strong positive loadings, and several strong negative loadings. For instance, the second component illustrates that selecting items related to 'Drive' means rejecting items related to 'Agreeableness'. Though somewhat interpretable, these "contrast" components present a problem for understanding the relationships between personality traits in this questionnaire.

Table 5 provides the four IRT forced-choice principal components, which accounted for 68% of the variance. The components were labeled 'Conscientiousness', 'Dominance', 'Agreeableness' and 'Adaptability and Dynamism'. This solution is very similar to the one derived from the normative scores (see Table 3) and dissimilar to the ipsative solution (Table 4). Clearly, the IRT methodology overcomes the problem of the distortion to construct validity produced by the ipsative scoring.

Insert Tables 3, 4 and 5 about here

Criterion-Related Validity. Criterion-related validities of the CCSQ scores produced by the four scoring methods were explored using the validation sample. We computed product-moment correlations between the trait scores and the incentive bonus received by the call center operators. Operational validities (not corrected for any artifacts) are given in Table 6. All traits measured by CCSQ were expected to relate **positively** to performance indicators in sales and customer service settings. Correlations between the CTT normative scores and the criterion were in line with expectations – all significant relationships were positive, and the most predictive traits were Analytical, Detail Conscious and Conscientious. The same was true when IRT single-stimulus scoring was used.

In contrast, when classical forced-choice (ipsative) scoring was used, two traits related to the criterion negatively (Flexible at -0.21**, and Persuasive at -0.14*). As has been discussed before, the sum of all covariances between the ipsative score and the criterion must sum to zero. Hence, any positive correlations with the external variable have to be compensated by some negative correlations. However, IRT scoring of forced-choice responses overcame this problem. All five significant correlations between the CCSQ scales and the incentive bonus were positive (see Table 6). These relationships involved traits that would be expected the most predictive in a technical call center – Analytical, Structured, Detail Conscious, Conscientious and Results Orientated. The IRT forced-choice validity coefficients were higher in magnitude than respective ipsative validities, and approached values of the single-stimulus validities.

Insert Table 6 about here

Discussion

Users of existing forced-choice questionnaires are keenly interested in establishing the test's structure, reliability, carrying out item analysis, or producing individual profiles. Unfortunately, these simple objectives have been impossible to achieve without encountering the distortions and artifacts produced by ipsative scoring, as we have illustrated here with the CCSQ. Yet, the limitations of ipsative data resulting from the traditional scoring of forced-choice questionnaires can be overcome by the use of item response modeling.

The particular IRT model used here is a reparameterization of the Thurstonian factor model. In turn, the Thurstonian factor model is a factor-analytic model embedded within the Law of Comparative Judgment. Although the latter was introduced by Thurstone as early as 1927, embedding a factor-analytic structure within it (and hence developing an IRT counterpart) was

only possible very recently, for computational reasons (see Maydeu-Olivares, 1999, 2001; Maydeu-Olivares & Böckenholt, 2005; Tsai & Böckenholt, 2001).

The Thurstonian IRT model provides a unified framework for estimating item parameters and obtaining individual scores for *any* existing questionnaire employing ranking or paired comparison format. Thus, after coding forced-choice data using binary outcome variables, the model can be specified within a familiar SEM framework to be estimated and scored by general-purpose software Mplus, which also conveniently estimates trait scores for individuals. Questionnaires measuring any number of traits, using ranking blocks of any size can be modeled. Brown and Maydeu-Olivares (2011) present extensive simulation studies revealing that item parameters and individuals' scores can be estimated very accurately in forced-choice designs using full ranking. When the 'most'-'least' format is used in blocks of four or more items, partial rankings are obtained and a missing data problem arises. Brown and Maydeu-Olivares (2012) show that Bayesian multiple imputations can be used to overcome this problem, as we have done with the CCSQ.

As the CCSQ application illustrates, IRT trait scores estimated from forced-choice responses are superior to ipsative scores **in all respects**.

Firstly, the IRT forced-choice scores allow **variability in profile locations** and therefore are directly interpretable for comparison between individuals. The IRT scoring yields non-trivial differences with ipsative scores in ordering of respondents. The re-ordering is systematic and brings the whole profile (rather than individual scales) closer to the respective normative profile.

Secondly, the IRT forced-choice scores provide a better **measurement precision** than the ipsative scores do. This is not surprising given that the IRT methodology is model based, and extracts most information on each preference decision. In contrast, the ipsative scoring assumes a

model that is wrong a priori, leading to violation of basic assumptions made in CTT. In addition, the IRT model provides means of estimating conditional standard errors for each individual combination of scores.

Thirdly, the underlying structure (i.e. **construct validity**) of the IRT forced-choice scores is no different from that of normative scores[3]. Once the IRT scoring has been applied, this structure is easy to establish. The average scale inter-correlation no longer has to be negative, and the forced-choice questionnaires can be factor analyzed using the same standard data analysis techniques that users of normative data have enjoyed.

The last, and perhaps the most important result from practitioners' point of view, is that the **criterion-related validity** of the IRT forced-choice scores is superior to that of ipsative scores. The IRT scoring removes the ipsative constraint forcing all correlations with an external criterion to sum to zero, thus eliminating any spurious validity coefficients, as the CCSQ application clearly shows. At the same time, because the IRT scoring is based on forced-choice responses, it preserves any potential gains that the comparative format might bring to the test's validity by reducing response biases.

To summarize, IRT scoring of forced-choice responses overcomes the problems of ipsative data – the scores show variability of profile locations, unconstrained scale correlations, and undistorted correlations with external criteria. These results suggest that there is absolutely no reason to hold on to ipsative scoring in forced-choice questionnaires.

Directions for future research and concluding remarks

Recent advances in computing capabilities have made item response modeling of forced-choice data possible, and new approaches have emerged recently to creating and scoring MFC questionnaires (e.g. Stark et al., 2005; McCloy et al., 2005). The Thurstonian IRT model

described here is currently the only model that can be readily applied to data collected with **existing** forced-choice questionnaires, with the objectives of estimating item parameters, relationships between the latent traits, and persons' parameters. Embedding these models into a general SEM framework enables further advantages that modeling with latent variables brings – such as including additional predictors or outcome variables and establishing relationships with error-free latent constructs rather than estimated scores. On the other hand, equipped with the knowledge of model parameters, researchers may use the approach to develop **new** forced-choice questionnaires, as the development of a short version of the Occupational Personality Questionnaire (OPQ32r; Brown, 2009) shows. However, this is a growing area of research and we expect and look forward to new IRT models for these kinds of data.

Both newly developed and existing forced-choice questionnaires, once scored appropriately, can be an excellent choice for substantive research. Because comparative formats remove nuisance factors acting *uniformly* across items, such as response sets or common method variance (Cheung & Chan, 2002), investigations using forced-choice questionnaires might prove more fruitful than those using single-stimulus items. A potentially tremendous advantage could be gained in cross-cultural personality research, where culture-specific response sets are consistently found (Van Herk et al., 2004; Johnson, Kulesa, Cho & Shavitt, 2005) and present a challenge for comparability of scores. Furthermore, the use of forced-choice formats is likely to prove fruitful in contexts outside of self-report personality assessment, particularly in contexts where over-generalization (or lack of differentiation) presents a particular challenge for validity. Examples of likely applications include assessments of other individuals, as in 360-degree feedback, notoriously affected by rater biases such as leniency/severity, or 'halo/horn' effects (Bartram, 2007), or patient satisfaction surveys known to suffer from halo effects due to

'affective overtones' (Brown, Ford, Deighton & Wolpert, 2012). While more research is needed to investigate whether and to what extent criterion-related validity of forced-choice formats prove superior to single-stimulus formats, it is clear that ipsative scoring should not be employed in such studies because it distorts validity estimates.

An important line of future research is related to the optimal design of forced-choice questionnaires. For instance, we have seen that with all its items keyed in the same direction, and all scales correlating positively, the forced-choice CCSQ is not optimally designed to maximize measurement precision. Designing forced-choice questionnaires is certainly a more complex endeavor than designing single-stimulus questionnaires, with more factors to consider and more design decisions to make. Brown and Maydeu-Olivares (2011) provide general guidelines for constructing MFC questionnaires; Maydeu-Olivares and Brown (2010) give specific guidelines for optimal design of one-dimensional measures using paired comparison and ranking tasks, but it is clear that more research on forced-choice questionnaire design is needed.

This outlook for future research, however, should not distract us from the main results of this paper, namely that IRT modeling can be successfully applied to existing forced-choice questionnaires, and that the IRT-estimated scores are free from the problems of ipsative scores. Simply put, scores obtained from forced-choice questionnaires do not have to be, and should no longer be, ipsative. The problem of ipsative data arising from forced-choice questionnaires has been effectively solved.

References

Ackerman, T.A. (2005). Multidimensional Item Response Theory Modeling. In A. Maydeu-Olivares & J. J. McArdle. (Eds.). *Contemporary Psychometrics* (pp. 3-26). Mahwah, NJ: Lawrence Erlbaum.

Asparouhov, T. & Muthén, B. (2010). *Multiple imputation with Mplus*. *Version 2*. Retrieved from http://www.statmodel.com

Baron, H. (1996). Strengths and Limitations of Ipsative Measurement. *Journal of Occupational and Organizational Psychology*, *69*, 49-56.

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

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., Ford, T., Deighton, J. & Wolpert, M. (2012). Satisfaction in child and adolescent mental health services: Translating users' feedback into measurement. *Administration and Policy in Mental Health and Mental Health Services Research*. Advance online publication. DOI: 10.1007/s10488-012-0433-9

Brown, A. & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71(3),* 460-502.

Brown, A. & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods*. Advance online publication. DOI: 10.3758/s13428-012-0217-x

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.

Chernyshenko, O.S., Stark, S., Prewett, M.S., Gray, A.A., Stilson, F.R. & Tuttle, M.D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: empirical comparisons with other formats. *Human Performance, 22,* 105-127.

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.

Closs, S. J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational Psychology, 69,* 41-47.

Cornwell, J. M. & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson. *Journal of Occupational and Organizational Psychology, 67,* 89-100.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research, 29,* 115-126.

Embretson, S., & Reise, S. (2000). *Item Response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Friedman, H., & Amoo, T. (1999). Rating the rating scales. *Journal of Marketing Management*, *9*, 114-123.

Gordon, L.V. (1976). *Survey of interpersonal values. Revised manual*. Chicago, IL: Science Research Associates.

Gordon, L.V. (1993). *Manual: Gordon Personal Profile-Inventory.* The Psychological Corporation, San Antonio, TX.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin*, *74*, 167-184.

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.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153-162.

Johnson, T., Kulesa, P., Cho, Y. I., & Shavitt, S. (2005). The relation between culture and response styles: evidence from 19 countries. *Journal of Cross-Cultural Psychology, 36(2),* 264-277.

Karpatschof, B., & Elkjaer, H. K. (2000). *Yet the bumblebee flies: The reliability of ipsative scores examined by empirical data and a simulation study*. Department of Psychology, University of Copenhagen: Research Report no. 1.

Kolb, A. & Kolb, D. (2005). *The Kolb Learning Style Inventory—Version 3.1. Technical Specifications.* Boston, MA: Hay Resource Direct.

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. & 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 approach*. Mahwah, NJ: Lawrence Erlbaum.

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. (1998-2010). *Mplus User's guide. Sixth edition*. Los Angeles, CA: Muthén & Muthén.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer.

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.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

SHL. (1997). Customer Contact: Manual and User's Guide. Surrey, UK. SHL Group.

SHL. (2006). *OPQ32 Technical Manual*. Surrey, UK. SHL Group.

Stark, S. (2002). *A new IRT approach to test construction and scoring designed to reduce the effects of faking in personality assessment*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

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., Chernyshenko, O., Drasgow, F. & Williams, B. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.

Stark, S., Chernyshenko, O.S., Drasgow, F., & White, L.A. (2012). Adaptive testing with multidimensional pairwise preference items: Improving the efficiency of personality and other noncognitive assessments. *Organizational Research Methods*. DOI: 10.1177/1094428112444611

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology, 73*, 749-751.

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

Thurstone, L.L. (1929). The measurement of psychological value. In Thomas Vernor Smith and William Kelley 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.

Van Herk, H., Poortinga, Y., & Verhallen, T. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology, 35*, 346-360.

Appendix

Short descriptions of the 16 traits measured by the Customer Contact Styles Questionnaire

1. **Persuasive** - enjoys selling, negotiating and gaining commitment.
2. **Self-control** - restrained in showing irritation or annoyance; rarely criticizes others openly; remains patient.
3. **Empathic** - sensitive and understanding towards others; prepared to go out of their way to help.
4. **Modest** - reserved about personal achievements and disinclined to talk about self.
5. **Participative** - enjoys team work and wants to develop constructive relationships.
6. **Sociable** - sociable, talkative and confident with different types of people; livens up group activities.
7. **Analytical** - enjoys analyzing information; working with data; probing the facts and solving problems.
8. **Innovative** - comes up with a wide range of ideas and offers imaginative or novel solutions.
9. **Flexible** - open to new approaches and readily adapts to different circumstances.
10. **Structured** - plans ahead; considers preparation, priority setting and structure to be important.
11. **Detail conscious** - ensures accuracy by checking details carefully and by being neat and tidy.
12. **Conscientious** - willing to persevere, to keep firmly to deadlines and to make sure that tasks are completed.
13. **Resilience** - copes with external stresses and pressures by being calm, thick skinned and looking on the bright side.
14. **Competitive** - needs to win at all costs, hates to lose and likes to be the best.
15. **Results orientated** - sets ambitious personal targets; stimulated by challenging targets; keen to improve own performance.
16. **Energetic** - enjoys being active; keeps busy; sustains a high level of energy over a long time.

Footnotes

[1] The simplifying assumption of local independence is only employed for latent trait estimation, not for model parameter estimation.

[2] The reliabilities of the forced-choice IRT scores are likely to be slightly overestimated, due to ignoring local dependencies existing in forced-choice blocks of four items. Previous simulation studies with 4-item blocks yielded an overestimation of empirical reliabilities by about 3% (Brown & Maydeu-Olivares, 2011).

[3] We assume that normative data arising from single-stimulus items are of good quality and is not badly affected by response sets such as acquiescence or extreme/central tendency responding. When strong biasing factors of this nature are present, differences in factor structure might be found between the normative and the forced-choice data, with the latter yielding more robust results.

Table 1

Correlations between CCSQ Classical and IRT Scores (N = 610)

	Cross-method		Cross-format	
	SS-CTT with SS-IRT	FC-CTT with FC-IRT	SS-CTT with FC-CTT	SS-IRT with FC-IRT
Persuasive	.98	.83	.69	.68
Self-Control	.98	.88	.63	.66
Empathic	.98	.83	.63	.66
Modest	.98	.91	.58	.64
Participative	.98	.88	.71	.73
Sociable	.97	.89	.72	.72
Analytical	.97	.88	.65	.68
Innovative	.99	.90	.69	.73
Flexible	.97	.83	.63	.66
Structured	.98	.88	.67	.73
Detail Conscious	.97	.91	.70	.73
Conscientious	.98	.89	.69	.72
Resilience	.98	.86	.50	.52
Competitive	.97	.91	.73	.79
Results Oriented	.98	.88	.69	.69
Energetic	.98	.86	.67	.71
Median	**.98**	**.88**	**.68**	**.70**

Note. SS-CTT = classical single-stimulus (normative); FC-CTT = classical forced-choice (ipsative); SS-IRT = IRT single-stimulus; FC-IRT = IRT forced-choice. All correlations are significant at the 0.01 level (1-tailed).

Table 2

Reliability estimates for CCSQ Classical and IRT Scores (N = 610)

CCSQ scale	Number of items	SS-CTT	SS-IRT	FC-CTT	FC-IRT
Persuasive	7	.80	.76	.68	.79
Self-Control	9	.89	.87	.72	.79
Empathic	9	.83	.82	.74	.76
Modest	9	.88	.87	.75	.75
Participative	10	.90	.91	.80	.80
Sociable	8	.78	.78	.68	.77
Analytical	8	.79	.78	.66	.85
Innovative	9	.91	.91	.78	.83
Flexible	7	.82	.84	.62	.74
Structured	8	.86	.86	.73	.85
Detail Conscious	7	.85	.84	.75	.89
Conscientious	7	.87	.86	.72	.84
Resilience	9	.83	.83	.64	.72
Competitive	7	.82	.87	.71	.85
Results Orientated	7	.82	.80	.57	.84
Energetic	7	.87	.88	.75	.74
Median		**.84**	**.85**	**.72**	**.80**

Note. SS-CTT = classical single-stimulus (normative); FC-CTT = classical forced-choice (ipsative); SS-IRT = IRT single-stimulus; FC-IRT = IRT forced-choice. Reliability estimates for classical scores were obtained using coefficient alpha, for IRT scores using the empirical reliability described by Equation (13).

Table 3

Rotated Pattern Matrix for CCSQ Single-Stimulus Classical and IRT Scores (N = 610)

	1	2	3	4
	Conscientiousness	Dominance	Agreeableness	Adaptability and Dynamism
Persuasive		**.55 / .63**		.34 / .21
Self-control		**-.52 / -.43**	**.44 / .51**	.38 / .36
Empathic		-.22 / -.16	**.76 / .81**	
Modest		**-.67 / -.67**		.25 / .30
Participative			**.69 / .66**	
Sociable		.38 / .40	**.48 / .50**	.28 / .25
Analytical	**.68 / .71**		-.22 /-.17	.21 / .15
Innovative	.22 / .22	.37 / .39		**.46 / .41**
Flexible	.24 / .20			**.47 / .47**
Structured	**.83 / .84**			
Detail conscious	**.89 / .89**			
Conscientious	**.80 / .79**		.23 / .22	
Resilience		-.23 / -.17		**.89 / .90**
Competitive		**.66 / .66**		
Results orientated	**.47 / .48**	.38 / **.43**		.22 / .20
Energetic		.26 / .29		**.56 / .54**
Correlations				
Component 1		.02 / .06	.18 / .20	.34 / .33
Component 2			.05 / .08	.16 / .17
Component 3				.25 / .26

Note. Loadings for the classical scores are given before the slash; loadings for the IRT scores after the slash. Only loadings >|0.2| are printed; loadings >|0.4| are set in boldface.

Table 4

Rotated Pattern Matrix for CCSQ Forced-Choice Classical (ipsative) Scores (N = 610)

	1	2	3	4
	Conscientiousness vs. Creativity	Drive vs. Agreeableness	Social Adjustment vs. Analysis	Adaptability vs. Influence
Persuasive	**-.40**			**-.69**
Self-control		**-.54**		.39
Empathic		**-.65**		
Modest		**-.56**		
Participative	-.33	**-.47**		
Sociable	-.28		**.49**	
Analytical		.21	**-.71**	
Innovative	**-.52**	.25	**-.60**	
Flexible	-.27	**.46**		**.44**
Structured	**.74**			
Detail conscious	**.72**			
Conscientious	**.66**			
Resilience			**.41**	**.49**
Competitive			.32	**-.64**
Results orientated		**.64**		
Energetic		**.42**	**.52**	
Correlations				
Component 1		-.04	-.19	-.02
Component 2			-.04	-.10
Component 3				.00

Note. Only loadings >|0.2| are printed; loadings >|0.4| are set in boldface.

Table 5

Rotated Pattern Matrix for CCSQ Forced-Choice IRT scores (N = 610)

	1	2	3	4
	Conscientiousness	Dominance	Agreeableness	Adaptability and Dynamism
Persuasive		**.88**		
Self-control		**-.84**		
Empathic		**-.55**	**.58**	-.29
Modest		**-.58**	-.37	
Participative			**.66**	
Sociable	-.26	.23	**.61**	.37
Analytical	**.82**			
Innovative	.32	.28		**.43**
Flexible	**.42**		.25	**.57**
Structured	**.90**			
Detail conscious	**.93**			
Conscientious	**.83**		.21	
Resilience		-.36		**.92**
Competitive		**.76**		
Results orientated	**.60**	.39	.34	.27
Energetic		.24		**.53**
Correlations				
Component 1		.04	.04	.15
Component 2			.03	.20
Component 3				.10

Note. Only loadings >|0.2| are printed; loadings >|0.4| are set in boldface.

Table 6

Correlations between Incentive Bonus and CCSQ Classical and IRT Scores (N = 219)

	SS-CTT	SS-IRT	FC-CTT	FC-IRT
Persuasive	.02	.01	-.13*	-.03
Self-Control	.21**	.20**	-.04	.09
Empathic	.14*	.15*	-.03	.13
Modest	.14*	.14*	-.04	.07
Participative	.20**	.19**	.02	.11
Sociable	.09	.08	-.11	-.02
Analytical	.26**	.25**	.19**	.22**
Innovative	.04	.04	-.06	.02
Flexible	.09	.08	-.21**	-.05
Structured	.21**	.24**	.13	.20**
Detail Conscious	.28**	.31**	.20**	.26**
Conscientious	.31**	.32**	.23**	.26**
Resilience	.10	.08	-.04	.02
Competitive	.05	.04	-.10	-.01
Results Orientated	.19**	.20**	.14*	.19**
Energetic	.08	.08	-.12	.01
Average	**.15**	**.15**	**0**	**.09**

Note. SS-CTT = classical single-stimulus (normative); FC-CTT = classical forced-choice (ipsative); SS-IRT = IRT single-stimulus; FC-IRT = IRT forced-choice. ** Correlation is significant at the 0.01 level; * 0.05 level (2-tailed).

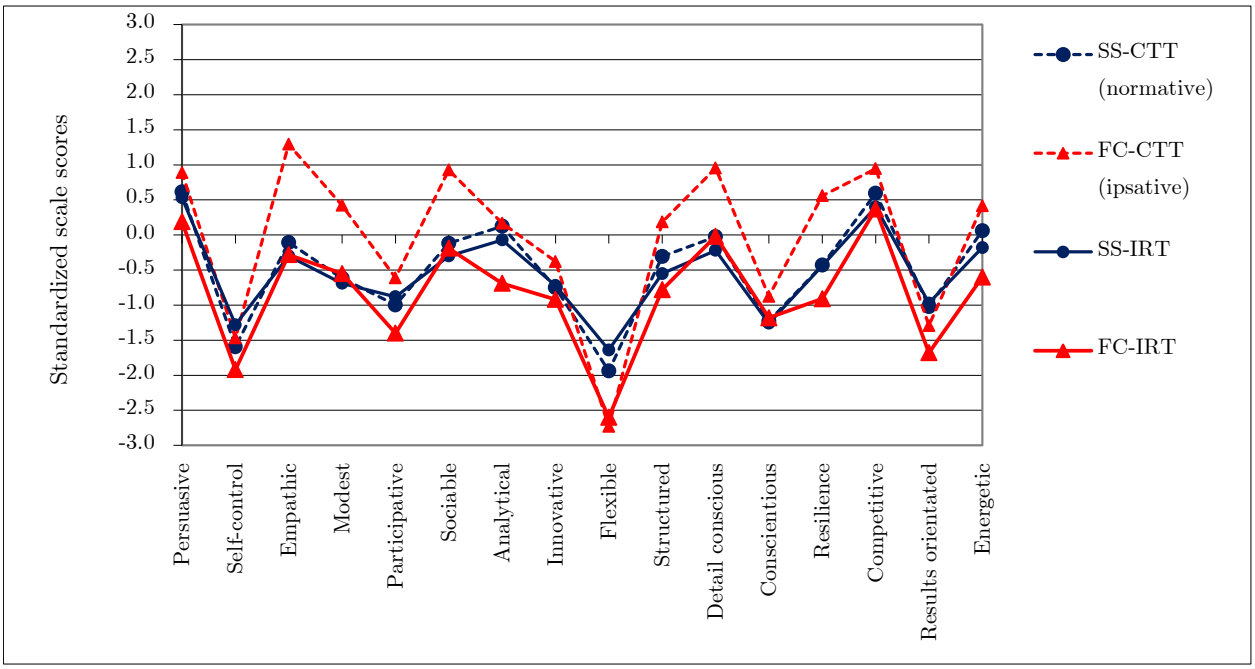
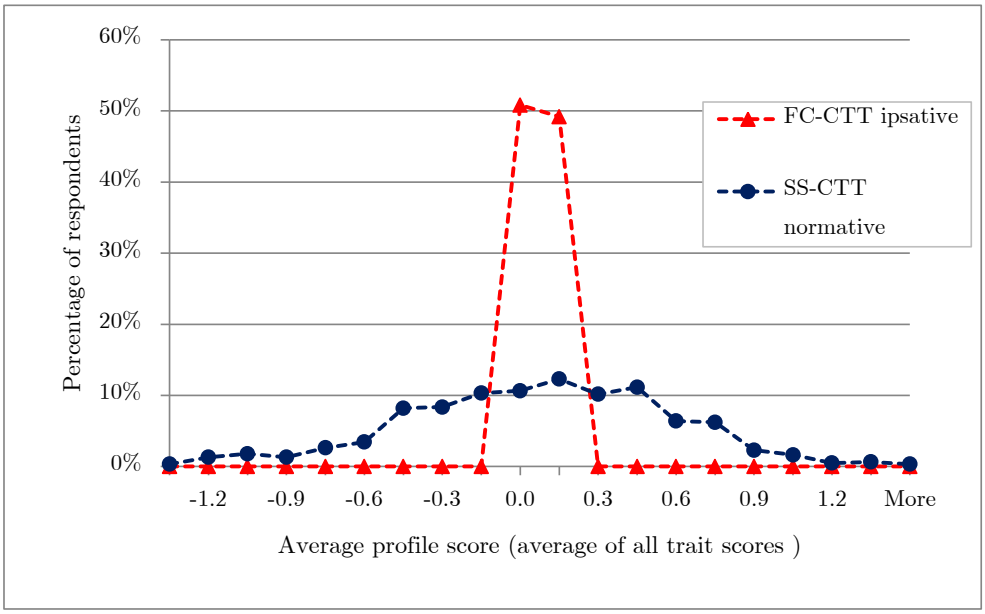
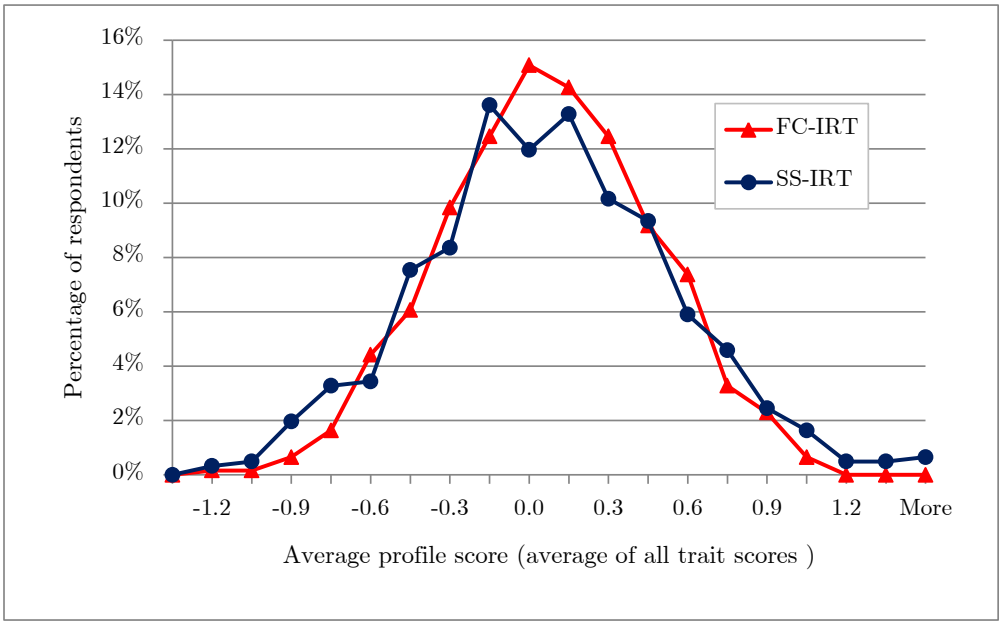


Figure 1. Sample CCSQ profile dominated by below average scores; ipsative scores fail to reflect the overall negative location of the profile.



(a) Classical normative versus ipsative scores



(b) IRT-estimated single-stimulus versus forced-choice scores

Figure 2. Distributions of average CCSQ profile scores.

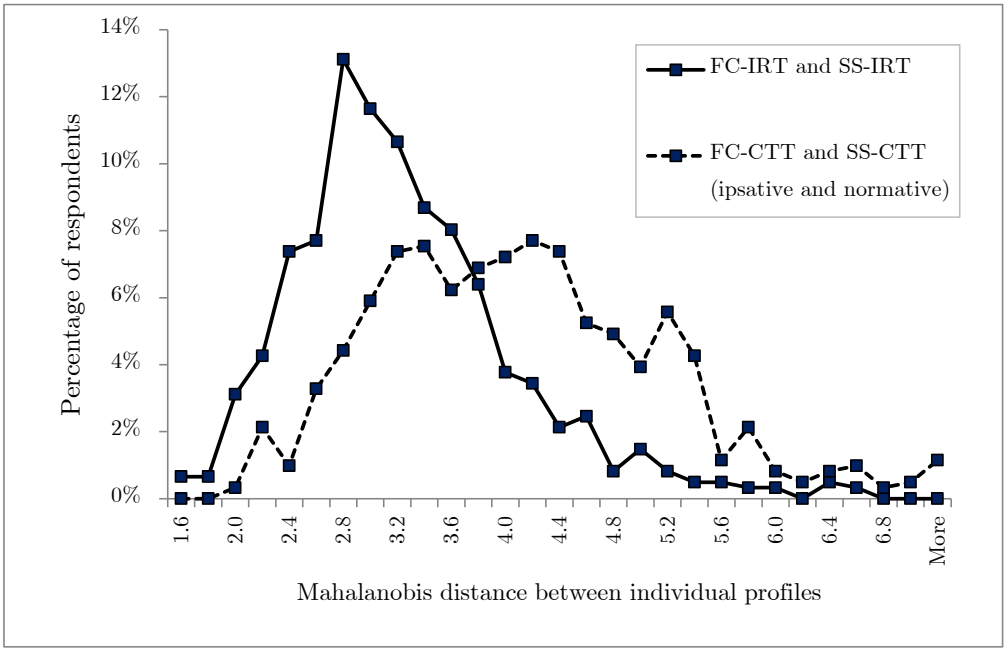


Figure 3. Distributions of Mahalanobis distances between single-stimulus and forced-choice CCSQ profiles.