Doing Less but Getting More: Improving Forced-Choice Measures with IRT

Anna Brown & Dave Bartram

SHL Group Ltd, UK

Paper presented at the 24th Annual conference of the Society for Industrial and Organizational Psychology, New Orleans, April 2-4 2009.

Poster session #200-3. Friday 3 April, 2009, 3.30-4.20 pm, Napoleon ABC.

Correspondence concerning this article should be addressed to Anna Brown:
anna.brown@shlgroup.com

TITLE
Doing less but getting more: Improving forced-choice measures with IRT

ABSTRACT
Using IRT we show how more efficient use can be made of information in forced-choice questionnaires. The approach described reduces the length of the instrument, and provides information on people's absolute trait standing and the scales' relationships. Both of these are impossible to obtain from CTT-scored forced choice questionnaires.

PRESS PARAGRAPH
Multidimensional forced-choice (MFC) questionnaires typically show good validities and are resistant to impression management effects. However, they yield ipsative data, which distorts scale relationships and makes comparisons between people problematic. Depressed reliability estimates also led developers to create tests of potentially excessive length. We apply an IRT Preference Model to make more efficient use of information in existing MFC questionnaires. OPQ32i used for selection and assessment internationally is examined using this approach. The latent scores recovered from a much reduced number of MFC items are superior to the full test's ipsative scores, and comparable to unbiased normative scores.

Forced-choice measures were designed to reduce response biases by forcing respondents to choose between statements measuring different traits (multidimensional forced-choice or MFC) according to the extent the statements describe their preferences or behavior. Typical MFC format tests consist of blocks of two or more statements from different dimensions. Usually respondents are asked to rank-order the statements, or, typically where four or more statements are involved, to select one statement which is "most like me" and one which is "least like me".

The forced-choice format has been shown to successfully reduce uniform response biases (Cheung & Chan, 2002), and to produce greater operational validity coefficients (Bartram, 2007; Christiansen, Burns & Montgomery, 2005). It is commonly found that the MFC format substantially reduces score inflation due to "faking good" compared to the single-stimulus (SS) format (Jackson, Wroblewski & Ashton, 2000; Martin, Bowen & Hunt, 2002; Christiansen et al., 2005) and is resistant to distortion to its covariance structure (Brown, 2008). However, forced-choice tests have been heavily criticized because their traditional scoring methodology results in ipsative data, very special properties of which pose threats to construct validity and score interpretation as well as other substantial psychometric challenges (e.g. Dunlap & Cornwell, 1994; Johnson, Wood, & Blinkhorn, 1988; Meade, 2004; Tenopyr, 1988).

Data is ipsative when the sum of the raw scores obtained over all measured scales is a constant for any individual. Variations in questionnaire design produce fully ipsative or partially ipsative scores. Here we will consider the most extreme, and therefore the most problematic type – fully ipsative scores. They are typically derived from the MFC format where statements' inverted rank-orders in a block are added to their respective scales. Regardless of the choices made, item scores in the block always sum to the same number, and therefore the total test score (sum of all the blocks) is the same for each individual.

Below we outline psychometric properties of ipsative data and discuss their implications for psychological assessment.

1. Relative nature of scores

Because the test allocates the same number of total points for everyone, it is impossible to get high (or low) scores on all scales in a multi-trait questionnaire. Therefore, many have argued, ipsative scores make sense for comparison of relative strength of traits within one individual, but they do not provide information on absolute (normative) trait standing, so comparisons between individuals are meaningless (e.g. Closs, 1996). The fact often overlooked is that the number of measured traits can substantially influence the validity of this claim. It has been shown that with a large number (30 or more) of relatively independent scales, the ordering of people on each trait largely corresponds to their normative ordering (Baron, 1996; SHL, 2006), therefore norming of ipsative scores is appropriate and intra-individual comparisons can be performed meaningfully.

Nevertheless, particularly for MFC instruments with few measured scales, this property can have serious implications for interpretation of scores, and remains the most serious limitation in practice.

2. Distorted construct validity

With the total test score constrained to be a constant, the total test variance is zero. Consequently, the average off-diagonal scales' correlation is a negative value and approaches zero as the number of scales increases (Clemans, 1966). Again, how much of a problem this is, depends on the number of scales in the questionnaire. With 30 scales, for example, the average off-diagonal correlation is only $r = -0.03$, allowing for a wide range of both negative and positive correlations between scales (Bartram, 1996; Baron, 1996).

Though less problematic with a large number of scales, scale correlations are typically suppressed in MFC measures, which clearly compromises construct validity of forced-choice questionnaires.

3. Distorted reliability estimates

It is generally agreed that the forced-choice format distorts traditional estimates of reliability. With a large number of measured dimensions reliabilities as measured by Cronbach's alpha are depressed (Bartram, 1996). It is also argued that alpha is an inappropriate statistics for the forced-choice format, unless a questionnaire meets very specific conditions (Brown & Maydeu-Olivares, 2009). Relying on coefficient alpha as a valid indicator of reliability has led test developers to creating questionnaires of potentially excessive length. This has an implication on the time it takes to complete the test and on test-takers' experiences.

4. Higher cognitive load

It is cognitively challenging to complete MFC tests, particularly when more than three items are involved in one block. Processing several items at the same time requires good reading skills and comprehension, and is generally found not suitable for people with low educational level (SHL, 2006). Unsurprisingly, success in faking MFC was found to be related to cognitive ability (Vasilopoulos et al., 2006).

These problems are serious enough to raise concerns with use of the forced-choice format. The first three, however, are not inherent to the format itself, but originate from the current way of scoring. The traditional scoring methodology based on the Classical Test Theory (CTT) approach does not adequately describe the decision-making process behind responding to MFC items.

New IRT models have been proposed to deal with some specific types of MFC measures (e.g. Stark, Chernyshenko & Drasgow, 2005; McCloy, Heggestad & Reeve, 2005).

A two-dimensional IRT Preference Model (Brown & Maydeu-Olivares, 2009) was introduced specifically for widely used MFC questionnaires with dominance items, such as the Occupational Personality Questionnaire (OPQ32i; SHL, 2006), the Customer Contact Styles Questionnaire (CCSQ7.2; SHL, 1997), the Survey of Interpersonal Values (SIV; Gordon, 1976), the Survey of Personal Values (SPV; Gordon, 1967) and others. It has been shown that embedding this IRT model in a confirmatory factor analytic framework allows estimating and scoring large tests like ones mentioned above (Brown & Maydeu-Olivares, 2009). Crucially, this approach deals with the limitations of existing MFC questionnaires, namely overcomes problems of ipsative data, and also provides the means of estimating the tests' reliability. For example, for the CCSQ7.2, measuring 16 work-related traits, reliability was found to be much higher than previously thought (Brown & Bartram, 2008).

Based on these findings, we would like to see if we can reduce the number of items in MFS questionnaires, while obtaining trait scores that are no longer ipsative. Instead of simply reducing the number of blocks, we will attempt to reduce the number of questions in each block, thus making completion less cognitively challenging. Why do we need to do this? First, we want test takers **to do less** – spend less time completing the questionnaire, without compromising its reliability and validity. Also, we want to make the format more appropriate for people with lower education level or reading skills. And finally, we want test users **to get more** – including information on absolute trait standing and true scales' relationships.

APPLICATION

Instrument

The Occupational Personality Questionnaire (OPQ32) is an occupational model of personality, which describes 32 dimensions of people's preferred style of behavior at work (SHL, 2006). It is a popular test used for selection and assessment internationally. Evidence supporting the job-related validity of the OPQ instruments has been reported in a number of

studies across a range of industry sectors and job types (e.g. Robertson & Kinder, 1993; SHL, 2006). Short scale descriptions for OPQ32 are given in Table 1.

There are two questionnaires using the above model, namely the OPQ32n (normative, using SS format) and OPQ32i (ipsative, using MFC format). The ipsative version of the OPQ32 was designed to be resistant to the effects of response distortion and 'faking good', and is used most frequently, particularly for selection. The instrument consists of 104 blocks of four statements measuring different dimensions. Each scale is measured by 13 items. For each block respondents have to choose one item that is 'most like me' and one 'least like me'. Here is an example of a block:

A. I like to do things my own way

B. I recognize weak arguments

C. I take care to follow procedures

D. I like to spend time with others

METHOD

Our approach relies on several assumptions. First, when rank-ordering statements, respondents perform mental pair-wise comparisons of all available options, that is, each statement is compared with every other one (Maydeu-Olivares, 1999). For instance, for an item to qualify to be "most like me" it has to be compared with all remaining items and "win" (or be preferred in) every comparison. Responses given to a block of four statements can be recoded into $4\times(4-1)/2=6$ directional paired comparisons as described in Maydeu-Olivares & Böckenholt (2005). If one statement is taken out of the block of four, making it a block of three, only $3\times(3-1)/2=3$ paired comparisons have to be performed by the respondent. Because one comparison is assumed not to influence outcomes of other comparisons, we can take existing data with four response alternatives, recode the block of four into six comparisons, and remove three comparisons related to the item to be removed from the block. It is easy to

see that by removing one item, the number of paired comparisons to be performed is actually halved, and theoretically making choices within a block of three should only take half of the time that a block of four takes (not including the time it takes to read the statements). Thus, removal of 25% of the items should almost halve the instrument completion time.

Second, according to Thurstonian theory of comparative judgment (Thurstone, 1927, 1931), one statement is preferred to another if its utility is larger for the respondent. In case of personality questionnaires, utilities of statements for the respondent are assumed to be caused by strengths of underlying personality traits. When a respondent chooses between two items, their standing on the two underlying traits will influence the utilities of the choice alternatives, and therefore, the outcome of the comparison. The two-dimensional IRT Preference Model for paired comparisons (Brown & Maydeu-Olivares, 2009) is applied to recoded responses to link them to latent traits measured by the questionnaire, and is given by

$$P_{ij}(\boldsymbol{\theta}) \equiv P(y_{ij} = 1 | \theta_q, \theta_r) = \Phi(\alpha_{ij} + \beta_i \theta_q + \beta_j \theta_r) \qquad (1)$$

where β_i and β_j are the factor loadings describing the strength of the relationship between the factors θ_q and θ_r and the underlying response, and α_{ij} is the threshold.

Third, the model assumes that the items fit a dominance model, that is, when the true score on the underlying trait increases, probability of agreeing with the item is non-decreasing. All items of OPQ32i without exception are very strong positive statements created under CTT approach. They were first trialed in the single-stimulus format and only statements correlating strongly with the total scale score were retained. Re-examining the statements with IRT confirmed their good fit to a dominance model. In addition, we assume that each item in the block measures only one trait. We also assume unidimensionality of the 32 measured traits. These assumptions are necessary to guarantee a good fit to a confirmatory factor model, where each paired comparison serves as a dichotomous indicator for two latent traits (first-order factors), the 32 latent traits are allowed to correlate freely, their variances

are set to 1, and several additional constraints are imposed on the parameters for identification and substantive theoretical reasons (for details, see Brown & Maydeu-Olivares, 2009).

Selecting best items

Two samples were used to inform selection of items for the shortened version.

Sample 1. Single-stimulus trial of OPQ32i. In this trial OPQ32i items were administered using a 5-point Likert scale. Participants volunteered and completed the questionnaires online to receive a comprehensive feedback report. Among N=632 participants 51% were female and 49% male. The age ranged from 18 to 64 with the largest group being between 22 and 34 years of age.

Sample 2. OPQ32i Standardization sample. The OPQ32i standardization sample consisted of 807 respondents. About two-thirds were adults working in industry and commerce, and the remaining third were students. Some respondents completed the questionnaire for self-development purposes, the others solely for the purposes of the standardization study. 43% of the sample were male, 57% female. Age ranged from 16 to 68, with a mean of 31 and a standard deviation of 11.

First, each scale in the questionnaire had to be examined in relation to its dimensionality. This was done by fitting 1, 2 and 3 dimensional IRT models to the Likert responses on each scale separately (Sample 1). Exploratory factor analysis (ML with oblique rotation) was used to extract 1, 2 and 3 factors and produce fit indices to each of those models. Most scales were one-dimensional, and for those items with lowest factor loadings were highlighted for possible deletion. In several scales there was a second dimension that could not be ignored despite being highly correlated with the first. The second dimension typically consisted of 3 to 5 items with similar content. In those cases items from the second dimension that did not load on the first dimension were highlighted as potential candidates

for deletion. The common-factor model fitted to these scales after deletion of the highlighted items showed satisfactory fit. For two scales, the second dimension was largely independent. This resulted in almost zero discrimination some items showed on the common factor. It was important that these items were removed.

Next, items from the MFC completion (Sample 2) were considered. This step was very important for two reasons. First, when put in blocks, items can interact with each other in ways that cannot be envisaged from the SS presentation. Second, only actual trialing of items in blocks can establish their true "desirability" for respondents. If almost everybody (or nobody) in the sample selects an item in a block, that item provides very little information for most of the trait range. Examination of the MFC responses from Sample 2 carried out by fitting the IRT Preference Model (Brown & Maydeu-Olivares, 2009) generally confirmed the same items as in the SS trial to be problematic, and revealed few additional items that were highlighted for deletion.

Finally, judgmental reviews of all blocks were performed in order to remove one item from each block based on the criteria outlined above. One additional constraint was imposed: we were looking to remove equal number of items from each scale (3 or 4, retaining 9 or 10 items per scale). This step required not only statistical information obtained from samples 1 and 2, but also detailed expert knowledge of the questionnaire's scales in order to retain items important for the construct's meaning and breadth. If two items highlighted for deletion happened to be in the same block, the most problematic one was removed. If a block did not have any highlighted items, it was used to remove items from scales that were generally very good and balance the number of removed items.

The final version was assembled that had 104 blocks of 3 items (312 items), with 9 or 10 (and one scale with 11) items per scale.

Estimating IRT parameters and individual's trait level

The structural model for this test contained 32 freely correlated latent traits (corresponding to the 32 OPQ scales), and 312 observed binary outcomes of paired comparisons. The model was estimated in Mplus (Muthén & Muthén, 2006) using unweighted least squares (ULS) estimation with the OPQ32i Standardization Sample (Sample 2). After the model parameters are estimated, Mplus conveniently provides factor scores as the mode of the posterior distribution of the latent traits. We scored several samples to evaluate properties of latent theta scores recovered from the forced-choice ratings on the shortened version of OPQ32i.

RESULTS

Reliability and standard error of measurement

While 6 to 8 items per scale are enough to reach acceptable reliability with OPQ32n, as many as 13 items per scale were required to reach the same levels with the forced-choice OPQ32i (SHL, 2006). However, this is where reliability estimation is based on use of alpha.

As in multidimensional IRT models generally, directional test information can be computed for each theta value in the 32-dimensional space (Ackerman, 2005). Details are beyond the scope of this paper and can be found in Brown & Maydeu-Olivares (2009). Average standard errors for the 32 scales can be computed for a sample of respondents, and a composite reliability then can be computed by comparing the average squared standard errors, σ_θ^2, to the trait score variance, σ^2 , which is in this model set to 1 (Embretson & Reise, 2000):

$$r'_{tt} = 1 - \frac{\sigma_\theta^{\,2}}{\sigma^2} \qquad (2)$$

Table 2 shows composite reliabilities estimated from the IRT information for the full version of OPQ32i and the shortened version, and also full version's alphas for comparison. The composite reliabilities for the short version are not much lower than for the full version

(median reliability 0.85 as compared to 0.92). Reliabilities estimated from the IRT information, as expected, are much higher than alphas, even for the reduced number of items.

Construct validity

For the first time it became possible to recover true correlations between OPQ32i scales. Exploratory factor analysis (ML with oblique rotation) was performed on the estimated theta scores for the Standardization sample (Sample 2), which extracted 5 factors explaining 54.2% variance (see Table 3). This solution clearly represents the Big Five factors (McCrae and Costa, 1987). For comparison, five or six factors are typically extracted from the normative OPQ32n, five of which represent typical "Big Five" descriptions. The sixth dimension, if extracted, is not consistent across samples (SHL, 2006).

Scaling properties

The most interesting and much debated question is whether scores based on MFC responses can resemble normative trait standing. To evaluate individual scores' properties, we will consider Sample 3, where respondents took both the normative and the ipsative versions of OPQ32.

Sample 3. Training delegates sample (OPQ32n and OPQ32i). This sample consisted of 551 individuals that participated in OPQ training courses and completed both the ipsative and the normative instruments within a few days interval. The participants were primarily Human Resources professionals, consultants or people working in related fields. 21.3% were male, 75.4% female and 3.3% did not provide gender data.

Figure 1 illustrates the distribution of the average profile scores for this sample. The classical ipsative profiles for this sample, as expected, were centered on zero (the average of the standardized ipsative scores ranged from z = -0.07 to z = 0.06 with mean 0.00 and standard deviation 0.02). The IRT-based score profiles, however, were distributed normally with the average profile score ranging from θ = -0.77 to θ = 0.71, with mean 0.0 and standard

deviation 0.26. For comparison, the average normative profile scores of OPQ32n ranged from z = -0.86 to z = 0.84 with mean 0.00 and standard deviation 0.29.

Ordering of participants based on the single-stimulus and forced-choice responses is similar. While correlations between normative and traditionally computed ipsative scores ranged from 0.49 to 0.80 with median 0.71, correlations between normative and IRT scores are higher, ranging from 0.56 to 0.80 with median 0.70 (see Table 4). Moreover, the average profile scores based on the IRT forced-choice scale scores correlated with the average normative profile scores (r = 0.56), demonstrating that forced-choice ratings can provide information on absolute trait standing.

Individual test profiles

Next we consider the 32-scale profiles based on CTT normative and IRT forced-choice scores, looking at their shape and absolute position. We measured similarity of shapes by correlating 32 scale scores (normative and IRT recovered, k=32) for the same individual in the sample of OPQ training delegates (Sample 3). These profile similarity coefficients were distributed as shown in Figure 2. Most people (56%) had profiles with similarity 0.7 or higher and only 10% of respondents had profiles with similarity less than 0.5. Clearly, self-referenced relative ordering of scales was similar based on SS and FC responses.

We measured the distance between the average of standardized normative scores and average of IRT forced-choice scores for the 32 scales. Figure 3 shows the distribution of the profile distance scores. It can be seen that the distance or "shift" between the forced-choice and the normative profile is distributed almost normally. Most people's (97%) profiles lie within 0.5 from each other, and 80% have their profiles within 0.2 or closer. Thus absolute positions of scales were also similar based on SS and FC responses.

Criterion-related validity

Sample 4. Validation sample (OPQ32i and competency ratings). This validation study was conducted in an organization in the food manufacturing industry. 835 Directors and Senior Managers located across Europe, Asia Pacific, North, Central and South America completed OPQ32i for development purposes. Ages ranged between 35 and 60 years, almost all educated to university level. The appropriate language version of the OPQ32i was used in different countries. The SHL Inventory of Management Competencies (IMC) was used as the 360-degree tool to obtain performance ratings. The IMC was completed by self and manager/s in the appropriate language version.

Composite Big Five scores were produced from OPQ32 scales, following the mappings given in OPQ32 Technical Manual (SHL, 2006). Tables 5 and 6 compare correlations between the performance ratings and the Big Five (based on both OPQ32i and shortened version IRT scores). It can be seen that for most competencies there are only insignificant differences between validity coefficients for the full ipsative version and the short IRT version, and that for some competencies the IRT scoring introduced improvement.

CONCLUSIONS

Multidimensional forced-choice measures, despite being resistant to impression management distortion and showing operational validities equal to or better than normative measures, have psychometric problems if scored with classical scoring procedures. Ipsative scores derived from these instruments make it difficult to establish construct validity, absolute location of profiles and reliability estimates. They are also generally longer than their SS counterparts, and more cognitively challenging.

We examined the forced-choice version of OPQ32 to see if ratings provided to blocks of items can be used more efficiently with IRT. Specifically, we wanted to see if the questionnaire can be significantly reduced in length, without compromising its reliability, and provide information on true relationships between scales and normative trait standings.

We examined each measured scale, and removed a quarter of the items that provided least information. We applied the IRT Preference model (Brown & Maydeu-Olivares, 2009) to estimate latent traits from the shortened version. The recovered scores show properties similar to the normative scores: they extract the same second-order factors, provide very similar ordering of respondents on all measured scales and even indicate respondents' absolute trait standing. This suggests that IRT can significantly improve the efficiency of existing MFC measures, without compromising their reliability and validity. Most importantly, for tests with sufficient number of largely independent dimensions, like OPQ32, it can also provide normative information, which, it has been argued for a long time, could not be recovered from forced-choice measures.

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.

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

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. Journal of Occupational Psychology, 69, 25-39.

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

Brown, A. (2008). The Impact of Questionnaire Item Format on Ability to "Fake Good". In Brown, A. (chair): Exploring the use of ipsative measures in personnel selection. Symposium presented at the 6th Conference of the International Test Commission, Liverpool.

Brown, A. & Bartram, D. (2008). IRT model for recovering latent traits from forced-choice personality tests. Paper presented at the 23th annual conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Brown, A. & Maydeu-Olivares, A. (2009). How IRT can solve problems of ipsative data. Paper submitted for publication.

Cheung, M.W.L, & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. Structural Equation Modeling: A Multidisciplinary Journal, Volume 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.

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.

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

Gordon, L.V. (1984). *Survey of personal values. Examiner's Manual*. Chicago, IL: Science Research Associates.

Haaland, D., & Christiansen, N. (1998). Departures from linearity in the relationship between applicant personality test scores and performance as evidence of response distortion. Paper presented at the 22nd Annual International Personnel Management Association Assessment Council Conference, Chicago, IL.

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.

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.

Martin, B. A., Bowen C.C., & Hunt, S. T. (2001). How effective are people at faking on personality questionnaires? Personality and Individual Differences, 32, 247-256.

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

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

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.

McCrae, R. & Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. Journal of Personality and Social Psychology, 52, 81-90.

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. (1996-2006). *Mplus User's guide. Fourth edition*. Los Angeles, CA: Muthén & Muthén.

Robertson, I.T. & Kinder, A. (1993). Personality and job competencies: An examination of the criterion-related validity of some personality variables. Journal of Occupational and Organizational Psychology, 66, 225-244.

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

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

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.

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, 79, 281-299.

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

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

Table 1

Short descriptions of the 32 traits measured in OPQ32

	Low scorers	High Scorers
RELATIONSHIPS WITH PEOPLE		
Persuasive	rarely pressures others to change their views, dislikes selling, less comfortable using negotiation	enjoys selling, comfortable using negotiation, likes to change other people's view
Controlling	happy to let others take charge, dislikes telling people what to do, unlikely to take the lead	likes to be in charge, takes the lead, tells others what to do, takes control
Outspoken	holds back from criticising others, may not express own views, unprepared to put forward own opinions	freely expresses opinions, makes disagreement clear, prepared to criticise others
Independent minded	accepts majority decisions, prepared to follow the consensus	prefers to follow own approach, prepared to disregard majority decisions
Outgoing	quiet and reserved in groups, dislikes being centre of attention	lively and animated in groups, talkative, enjoys attention
Affiliative	comfortable spending time away from people, values time spent alone, seldom misses the company of others	enjoys others' company, likes to be around people, can miss the company of others
Socially Confident	feels more comfortable in less formal situations, can feel awkward when first meeting people	feels comfortable when first meeting people, at ease in formal situations
Modest	makes strengths and achievements known, talks about personal success	dislikes discussing achievements, keeps quiet about personal success
Democratic	prepared to make decisions without consultation, prefers to make decisions alone	consults widely, involves others in decision making, less likely to make decisions alone
Caring	selective with sympathy and support, remains detached from others' personal problems	sympathetic and considerate towards others, helpful and supportive, gets involved in others' problems
THINKING STYLE		
Data Rational	prefers dealing with opinions and feelings rather than facts and figures, likely to avoid using statistics	likes working with numbers, enjoys analysing statistical information, bases decisions on facts and figures
Evaluative	does not focus on potential limitations, dislikes critically analysing information, rarely looks for errors or mistakes	critically evaluates information, looks for potential limitations, focuses upon errors
Behavioural	does not question the reasons for people's behavior, tends not to analyze people	tries to understand motives and behaviours, enjoys analysing people
Conventional	favours changes to work methods, prefers new approaches, less conventional	prefers well established methods, favours a more conventional approach
Conceptual	prefers to deal with practical rather than theoretical issues, dislikes dealing with abstract concepts	interested in theories, enjoys discussing abstract concepts
Innovative	more likely to build on than generate ideas, less inclined to be creative and inventive	generates new ideas, enjoys being creative, thinks of original solutions

Variety Seeking	prefers routine, is prepared to do repetitive work, does not seek variety	prefers variety, tries out new things, likes changes to regular routine, can become bored by repetitive work
Adaptable	behaves consistently across situations, unlikely to behave differently with different people	changes behavior to suit the situation, adapts approach to different people
Forward thinking	more likely to focus upon immediate than long-term issues, less likely to take a strategic perspective	takes a long-term view, sets goals for the future, more likely to take a strategic perspective
Detail Conscious	unlikely to become preoccupied with detail, less organised and systematic, dislikes tasks involving detail	focuses on detail, likes to be methodical, organised and systematic, may become preoccupied with detail
Conscientious	sees deadlines as flexible, prepared to leave some tasks unfinished	focuses on getting things finished, persists until the job is done
Rule Following	not restricted by rules and procedures, prepared to break rules, tends to dislike bureaucracy	follows rules and regulations, prefers clear guidelines, finds it difficult to break rules
	FEELINGS AND EMOTIONS	
Relaxed	tends to feel tense, finds it difficult to relax, can find it hard to unwind after work	finds it easy to relax, rarely feels tense, generally calm and untroubled
Worrying	feels calm before important occasions, less affected by key events, free from worry	feels nervous before important occasions, worries about things going wrong
Tough Minded	sensitive, easily hurt by criticism, upset by unfair comments or insults	not easily offended, can ignore insults, may be insensitive to personal criticism
Optimistic	concerned about the future, expects things to go wrong, focuses on negative aspects of a situation	expects things will turn out well, looks to the positive aspects of a situation, has optimistic view of the future
Trusting	wary of others' intentions, finds it difficult to trust others, unlikely to be fooled by people	trusts people, sees others as reliable and honest, believes what others say
Emotionally Controlled	openly expresses feelings, finds it difficult to conceal feelings, displays emotion clearly	can conceal feelings from others, rarely displays emotion
Vigorous	likes to take things at a steady pace, dislikes excessive work demands	thrives on activity, likes to keep busy, enjoys having a lot to do
Competitive	dislikes competing with others, feels that taking part is more important than winning	has a need to win, enjoys competitive activities, dislikes losing
Achieving	sees career progression as less important, looks for achievable rather than highly ambitious targets	ambitious and career-centred, likes to work to demanding goals and targets
Decisive	tends to be cautious when making decisions, likes to take time to reach conclusions	makes fast decisions, reaches conclusions quickly, less cautious

Table 2

Scale reliability estimates for the CTT scored OPQ32i, OPQ32i short version, and IRT estimated forced-choice scales of OPQ32i short version (Sample 2, N=807)

	Short version		**Full version (13 items per scale)**	
OPQ32 scale	**Number of items in Short version**	**IRT composite reliability**	**IRT composite reliability**	**alpha**
Persuasive	10	0.88	0.94	0.81
Controlling	9	0.89	0.95	0.87
Outspoken	10	0.84	0.92	0.76
Independent minded	9	0.81	0.89	0.72
Outgoing	9	0.90	0.95	0.85
Affiliative	10	0.85	0.93	0.82
Socially Confident	9	0.88	0.94	0.83
Modest	10	0.77	0.88	0.81
Democratic	9	0.70	0.84	0.68
Caring	10	0.78	0.88	0.78
Data Rational	10	0.86	0.93	0.88
Evaluative	9	0.76	0.87	0.67
Behavioural	10	0.86	0.93	0.82
Conventional	10	0.72	0.84	0.74
Conceptual	10	0.89	0.94	0.79
Innovative	10	0.88	0.95	0.88
Variety Seeking	9	0.80	0.89	0.72
Adaptable	10	0.86	0.92	0.82
Forward thinking	11	0.83	0.90	0.75
Detail Conscious	10	0.87	0.93	0.80
Conscientious	10	0.83	0.92	0.82
Rule Following	10	0.81	0.90	0.84
Relaxed	10	0.88	0.94	0.85
Worrying	9	0.82	0.92	0.88
Tough Minded	9	0.83	0.92	0.82
Optimistic	10	0.85	0.93	0.80
Trusting	10	0.83	0.91	0.81
Emotionally Controlled	10	0.81	0.90	0.85
Vigorous	10	0.84	0.91	0.75
Competitive	10	0.87	0.93	0.86
Achieving	10	0.86	0.93	0.79
Decisive	10	0.86	0.93	0.80
Median		**0.85**	**0.92**	**0.81**

Table 3

Rotated factor loadings for the IRT scores estimated from forced-choice ratings of OPQ32i short version (Sample 2, N=807)

	1 Extraversion	2 Conscien-tiousness	3 Agreeableness	4 Neuroticism	5 Openness
Persuasive	**.49**	.04	.06	-.25	.17
Controlling	**.56**	.16	-.08	-.23	.15
Outspoken	**.48**	-.04	-.08	-.26	.12
Independent minded	.25	-.12	-.28	-.03	.38
Outgoing	**.62**	-.34	.28	-.20	-.13
Affiliative	.32	-.22	**.56**	-.01	-.15
Socially Confident	.30	-.08	.34	**-.62**	-.05
Modest	**-.62**	-.03	-.06	-.07	-.05
Democratic	.00	.11	**.61**	.08	.10
Caring	-.05	.05	**.74**	.04	.04
Data Rational	-.06	**.43**	-.15	-.07	.23
Evaluative	-.03	**.42**	-.03	-.12	**.63**
Behavioural	-.06	.01	**.57**	.10	**.48**
Conventional	-.21	.22	-.06	.01	**-.57**
Conceptual	-.11	-.04	.17	-.02	**.75**
Innovative	.20	.07	-.02	-.11	**.62**
Variety Seeking	.24	-.19	-.03	-.05	**.40**
Adaptable	.06	-.14	.06	.25	.17
Forward thinking	.17	**.54**	.10	.00	.27
Detail Conscious	-.10	**.78**	.05	-.03	-.15
Conscientious	.04	**.74**	.06	-.03	-.19
Rule Following	-.10	**.40**	.09	.06	**-.46**
Relaxed	-.12	-.04	.06	**-.63**	.04
Worrying	-.06	-.04	.02	**.82**	-.11
Tough Minded	-.02	-.02	-.06	**-.59**	.01
Optimistic	.18	.12	**.40**	-.33	.08
Trusting	-.05	.02	**.50**	-.15	-.04
Emotionally Controlled	**-.50**	-.05	-.35	-.11	.01
Vigorous	.22	**.46**	.11	.01	-.13
Competitive	**.58**	.08	-.38	.00	.01
Achieving	**.57**	**.42**	-.09	.09	.30
Decisive	.27	.02	-.28	-.30	.14
Factor correlations	1	2	3	4	5
1	1.00	.01	.09	-.26	.32
2		1.00	-.01	-.11	-.02
3			1.00	.02	-.03
4				1.00	-.12
5					1.00

Factor loadings above +/-0.4 are set in boldface

Table 4

Correlations of scores on the ipsative and normative versions of OPQ32 with the IRT recovered scores on the short forced-choice version (Sample 3, N= 551)

Correlations (N=551)	**OPQ32n with full OPQ32i (both CTT-scored)**	**OPQ32n with IRT-scored short version**
Persuasive	0.75	0.76
Controlling	0.73	0.74
Outspoken	0.68	0.69
Independent minded	0.49	0.56
Outgoing	0.78	0.80
Affiliative	0.68	0.70
Socially Confident	0.76	0.75
Modest	0.71	0.68
Democratic	0.59	0.58
Caring	0.60	0.60
Data Rational	0.80	0.80
Evaluative	0.61	0.63
Behavioural	0.63	0.62
Conventional	0.71	0.74
Conceptual	0.72	0.74
Innovative	0.80	0.80
Variety Seeking	0.60	0.66
Adaptable	0.63	0.65
Forward thinking	0.66	0.67
Detail Conscious	0.76	0.77
Conscientious	0.66	0.67
Rule Following	0.71	0.70
Relaxed	0.70	0.70
Worrying	0.75	0.73
Tough Minded	0.74	0.71
Optimistic	0.70	0.73
Trusting	0.65	0.66
Emotionally Controlled	0.75	0.76
Vigorous	0.61	0.61
Competitive	0.74	0.71
Achieving	0.77	0.74
Decisive	0.69	0.72
median	**0.71**	**0.70**

Table 5

Validity coefficients (correlations with manager ratings of performance) for composite Big 5 scores based on OPQ32i and IRT short forced-choice version (Sample 4, N=835)

Correlations in bold are hypothesised

	Extraversion		Openness		Emotional Stability		Agreeableness		Conscientiousness	
Manager ratings on:	Ipsative	IRT theta	Ipsative	IRT theta	Ipsative	IRT theta	Ipsative	IRT theta	Ipsative	IRT theta
leadership	**.16(**)**	**.17(**)**	.00	.06	.07(*)	.08(*)	.07(*)	.04	-.05	.03
planning organising	-.03	-.02	-.06	-.04	.04	.03	.04	.00	**.10(**)**	**.10(**)**
quality orientation	-.01	.00	-.02	-.02	.02	.03	.07	.04	**.07(*)**	**.08(*)**
persuasiveness	**.23(**)**	**.22(**)**	.03	.11(**)	.12(**)	.12(**)	.01	.01	-.13(**)	-.04
specialist knowledge	.12(**)	.11(**)	.03	.06	.05	.06	-.02	-.03	-.05	.00
problem solving	-.01	-.02	.01	.02	.04	.03	-.03	-.04	-.03	-.02
oral communication	**.20(**)**	**.20(**)**	.01	.10(**)	.11(**)	.12(**)	.00	.01	-.10(**)	.01
written communication	.01	.02	-.01	.03	.06	.06	-.03	.00	-.03	.03
commercial awareness	.11(**)	.12(**)	.00	.04	.05	.07(*)	-.09(**)	-.10(**)	-.01	.04
creativity innovation	.16(**)	.18(**)	**.24(**)**	**.25(**)**	.03	.06	-.03	-.02	-.06	-.01
decisiveness	.15(**)	.17(**)	.07(*)	.12(**)	.04	.07(*)	-.06	-.10(**)	.00	.06
strategic perspective	.08(*)	.09(*)	.08(*)	.12(**)	.04	.05	.00	.00	-.07(*)	-.01
interpersonal sensitivity	.03	.00	-.06	-.05	.09(**)	.06	**.23(**)**	**.20(**)**	-.12(**)	-.09(*)
flexibility	.03	.00	-.01	.00	.06	.04	.05	.03	-.03	-.01
resilience	-.04	-.05	-.04	-.02	**.13(**)**	**.10(**)**	.08(*)	.04	-.06	-.04
personal motivation	**.20(**)**	**.22(**)**	.08(*)	.14(**)	.01	.06	-.09(**)	-.08(*)	**.07**	**.15(**)**

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Table 6

Validity coefficients (correlations with self ratings of performance) for composite Big 5 scores based on OPQ32i and IRT short forced-choice version (Sample 4, N=835)

Correlations in bold are hypothesised

	Extraversion		Openness		Emotional Stability		Agreeableness		Conscientiousness	
Self ratings on:	Ipsative	IRT theta	Ipsative	IRT theta	Ipsative	IRT theta	Ipsative	IRT theta	Ipsative	IRT theta
leadership	**.17(**)**	**.24(**)**	.00	.07(*)	.04	.09(**)	.07(*)	.07(*)	.14(**)	.20(**)
planning organising	-.02	.02	-.13(**)	-.09(*)	.06	.06	.02	.01	**.42(**)**	**.39(**)**
quality orientation	-.03	.01	-.07(*)	-.05	-.02	-.01	-.03	-.02	**.33(**)**	**.32(**)**
persuasiveness	**.31(**)**	**.34(**)**	.04	.16(**)	.15(**)	.17(**)	-.01	.00	-.02	.09(**)
specialist knowledge	.10(**)	.15(**)	.04	.09(**)	.05	.08(*)	-.10(**)	-.06	.16(**)	.22(**)
problem solving	-.03	.02	.04	.08(*)	.08(*)	.08(*)	-.11(**)	-.09(*)	.14(**)	.18(**)
oral communication	**.30(**)**	**.32(**)**	.05	.16(**)	.21(**)	.23(**)	.05	.07(*)	-.02	.10(**)
written communication	.03	.07	-.01	.05	.15(**)	.14(**)	-.02	.01	.11(**)	.14(**)
commercial awareness	.13(**)	.19(**)	.00	.07	.08(*)	.12(**)	-.16(**)	-.14(**)	.17(**)	.24(**)
creativity innovation	.19(**)	.26(**)	**.48(**)**	**.48(**)**	.11(**)	.16(**)	-.04	-.01	-.04	.07(*)
decisiveness	.16(**)	.24(**)	.09(**)	.16(**)	.09(**)	.13(**)	-.06	-.09(**)	.14(**)	.21(**)
strategic perspective	.07	.14(**)	.15(**)	.19(**)	.06	.09(*)	-.02	.01	.11(**)	.17(**)
interpersonal sensitivity	.00	.02	-.06	-.04	.12(**)	.10(**)	**.35(**)**	**.32(**)**	-.04	-.02
flexibility	.01	.05	.10(**)	.10(**)	.18(**)	.17(**)	.15(**)	.13(**)	-.03	.01
resilience	-.05	.02	-.03	.02	**.34(**)**	**.31(**)**	.04	.03	.10(**)	.15(**)
personal motivation	**.19(**)**	**.28(**)**	.07(*)	.16(**)	.05	.13(**)	-.08(*)	-.03	**.26(**)**	**.35(**)**

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Figure 1

Distribution of the average profile score (average of standardized CTT normative and ipsative scores, and IRT-estimated forced-choice scores)

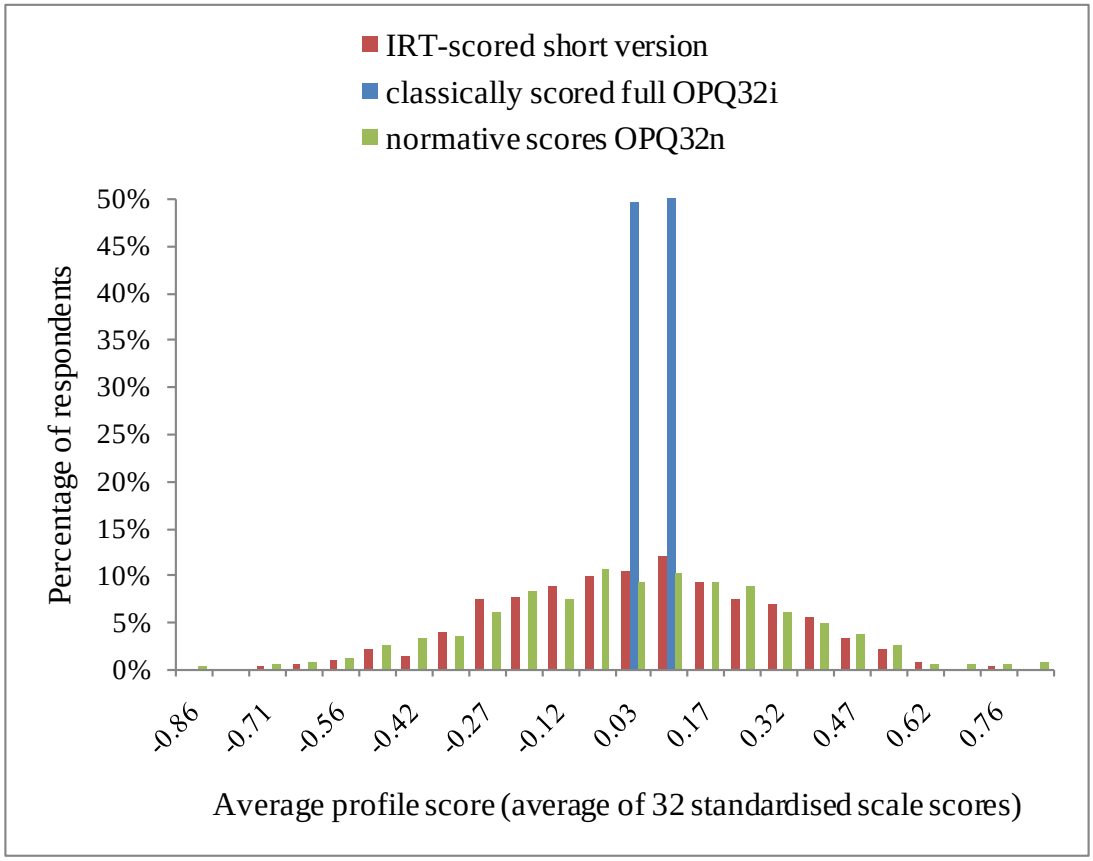

Figure 2

Distribution of profile similarity coefficients (similarity of CTT normative and IRT-estimated forced-choice profiles)

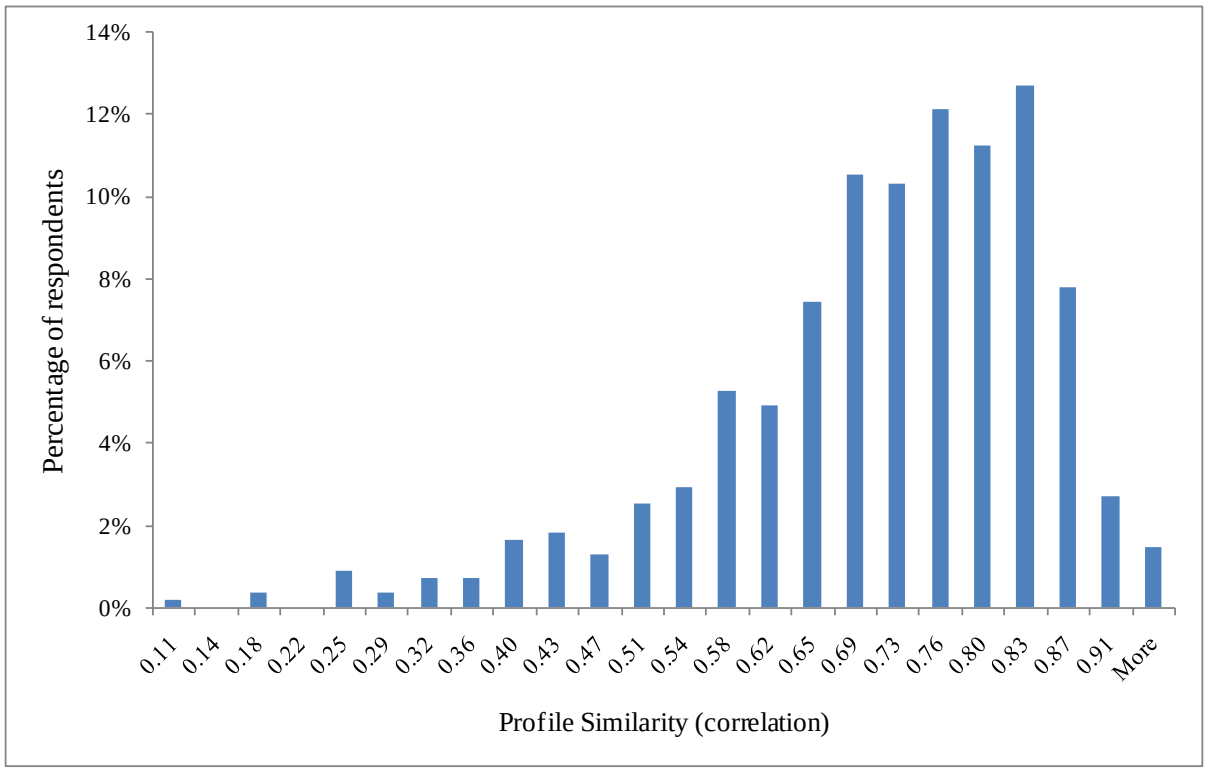

Figure 3

Distribution of profile distances (average distance between scales of CTT normative and IRT-estimated forced-choice profiles)