Citation
Brown, A. (2010). Doing less but getting more: Improving forced-choice measures with Item Response Theory. *Assessment and Development Matters, 2(1)*, 21-25.

Doing Less but Getting More: Improving Forced-Choice Measures with Item Response Theory (IRT)

Anna Brown, SHL Group.

Type: Research article

Measure
The Occupational Personality Questionnaire (OPQ32) describes 32 dimensions of behavioural style at work (SHL, 2006). The ipsative version (OPQ32i) consists of 104 blocks of 4 statements. For each block, respondents have to choose one statement that is 'Most Like Me' and one 'Least Like Me'.

Participants
Multiple international samples were considered for different parts of this research.

Digested Message
Forced-choice tests, despite being resistant to response biases and showing good operational validities, have psychometric problems if scored traditionally. These questionnaires are generally longer than their normative counterparts, and more cognitively challenging.

The OPQ32i was shortened and re-scored using the latest advances in IRT. One item was removed out of each block, making the completion quicker and less cognitively complex. The shortened version (OPQ32r) shows good reliability, equivalent or better validity than the full ipsative version, and produces scale scores with normative properties.

Results suggest that the IRT methodology can significantly improve efficiency of existing forced-choice measures so that test takers can *do less* (complete shorter and easier questionnaire) and test users can *get more* (bias-resistant instrument of superior psychometric quality).

Introduction
Despite the popularity of Likert scales in personality assessment, responses are subject to biases, for example acquiescence and halo effects. Forced-choice formats are designed to reduce such biases by forcing respondents to choose between equally desirable statements measuring different traits. Respondents cannot endorse all items, which typically results in reduction of impression management effects (Jackson et al, 2000) and greater operational validity coefficients (Bartram, 2007).

Despite these advantages, forced-choice tests have been heavily criticised because their traditional scoring methodology results in *ipsative* data, which pose threats to construct validity and score interpretation (Baron, 1996). These problems, however, do not arise from the format itself, but from overly simplistic scoring methodology, which ignores obvious violations of most statistical assumptions (Meade, 2004).

A new IRT model, based on Thurstone's (1927) theory of comparative judgement, was introduced to describe the decision process behind responding to forced-choice items (Maydeu-Olivares & Brown, 2009). This model provides the means for correct estimation of item parameters, scale relationships and test reliability. Brown (2008) has shown that this approach can be successfully applied to existing questionnaires, producing trait scores that are no longer ipsative. Among other findings, it emerged that reliability of forced-choice tests is in fact higher than previously thought.

Research Objectives

While the IRT methodology can be used to estimate and score existing forced-choice tests, this research takes it further and looks at how IRT might guide test development. Given the potentially excessive number of items in the widely used OPQ32i, we attempted to shorten the instrument while improving its psychometric properties.

Methodology

The idea behind Thurstone's model is quite simple. When rank-ordering n statements, respondents perform $n*(n-1)/2$ mental pair-wise comparisons, i.e. every statement is compared with every other one. Thus choices made in a block of 4 statements can be equivalently presented as 6 paired comparisons.

Each comparison can have one of two outcomes: the first item is preferred to the second, or otherwise. These pair-wise preferences are determined by items' "utilities" for the respondent (Maydeu-Olivares & Böckenholt, 2005), which are in turn influenced by underlying personality traits. The IRT approach links the binary outcomes of paired comparisons to personality traits through probability functions – put simply, we know how likely the observed response is for different trait levels.

When it comes to shortening, the Thurstone's model can also inform item and format selection. Rather than simply reducing the number of blocks, it justifies the removal of one item in each block, making it a block of three. Making only 3 mental paired comparisons instead of 6 should almost halve the completion time and reduce the test's cognitive complexity.

Analysis

To inform selection of items, several large OPQ32i samples including the standardization sample (N=807) were examined. Items providing least information in the confirmatory IRT model were highlighted for deletion. This

selection was cross-validated by administering all items to 632 volunteers using a 5-point Likert scale.

Finally, a judgemental review was performed to remove one item from each block, balancing the number of items per scale. This step required detailed expert knowledge of the questionnaire in order to retain items important for the scale's construct. A final version (named *OPQ32r*) was assembled with 104 blocks of 3 items (312 items in total, 9 or 10 items per scale).

The shortened test was administered to 518 volunteers from several English-speaking countries. The IRT model parameters were estimated and factor scores produced using M*plus* software.

Discussion

Reliability and standard error of measurement

Table 1 compares reliability estimates and standard errors for OPQ32i and OPQ32r. The most striking result is that IRT-based estimates for the short version are often higher than alphas for the full version. This is because classical statistics such as alpha are inappropriate with forced-choice data.

Interpretation of scores

To investigate scaling properties of the IRT-scored OPQ32r, a sample of 551 OPQ training delegates was considered, who also took the normative OPQ32n. The most problematic property of ipsative data is that the average profile score is the same for every individual (zero for z-standardised scores). The IRT scores demonstrate a remarkably different pattern: average profile scores ranged from -0.59 to 0.57 (SD = 0.23). The IRT methodology clearly produces scores that are no longer ipsative. Figure 1 compares distributions of the average profile score for IRT-scored short OPQ32r, full ipsative OPQ32i, and normative OPQ32n.

Construct validity

Unlike ipsative OPQ32i, the IRT-scored OPQ32r can be factor analysed. The training delegates sample (N=551) extracted six factors with the Big Five embedded in the solution, which was almost identical to the one extracted from the normative OPQ32n.

Criterion-related validity

Directors and senior managers from a multinational food manufacturing firm (N=835) completed OPQ32i for development purposes. The Inventory of Management Competencies (IMC) was used to obtain performance ratings from manager/s and others.

There were only insignificant differences between scales' validity coefficients based on standard OPQ32i scoring, and IRT scoring for the short OPQ32r version. Furthermore, IRT scoring introduced significant improvements in validity coefficients at the Big Five level. Because relationships between OPQ scales are no longer affected by ipsative constraints, aggregated Big Five scores capture more variance.

References/Sources

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(3),* 263-272.

Brown, A. (2008). How to get the best of both worlds: recovering normative scores from ipsative ratings. *Paper presented at DOP annual conference, Stratford-upon-Avon.*

Maydeu-Olivares, A. & Brown, A. (2009). IRT modeling of paired comparison and ranking data. *Submitted for publication.*

Jackson, D., Wroblewski, V., & Ashton, M. (2000). The Impact of Faking on Employment Tests: Does Forced Choice Offer a Solution? *Human Performance, 13(4),* 371–388.

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

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.

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

The Author

Anna Brown is a Principal Research Statistician at SHL Group.

Table1: Scale reliability estimates for the full OPQ32i and shortened OPQ32r

	Short OPQ32r *Calibration sample (N=518)*			*Full OPQ32i (13 items per scale)* *Stand. sample (N=807)*		
OPQ32 measured trait	**Number of items**	**IRT composite Reliability**	**Standard Error (theta=0 for all scales)**	**Alpha**	**IRT composite Reliability**	**Standard Error (theta=0 for all scales)**
Persuasive	10	0.83	0.36	0.81	0.94	0.18
Controlling	9	0.91	0.22	0.87	0.95	0.15
Outspoken	10	0.86	0.31	0.76	0.92	0.24
Independent minded	9	0.77	0.41	0.72	0.89	0.25
Outgoing	9	0.89	0.25	0.85	0.95	0.15
Affiliative	10	0.84	0.33	0.82	0.93	0.20
Socially Confident	9	0.87	0.29	0.83	0.94	0.18
Modest	10	0.81	0.34	0.81	0.88	0.29
Democratic	9	0.74	0.43	0.68	0.84	0.33
Caring	10	0.81	0.37	0.78	0.88	0.28
Data Rational	10	0.88	0.26	0.88	0.93	0.15
Evaluative	9	0.80	0.39	0.67	0.87	0.30
Behavioural	10	0.79	0.39	0.82	0.93	0.18
Conventional	10	0.68	0.49	0.74	0.84	0.33
Conceptual	10	0.78	0.40	0.79	0.94	0.18
Innovative	10	0.89	0.27	0.88	0.95	0.14
Variety Seeking	9	0.77	0.40	0.72	0.89	0.27
Adaptable	10	0.87	0.28	0.82	0.92	0.22
Forward thinking	11	0.87	0.30	0.75	0.90	0.25
Detail Conscious	10	0.89	0.24	0.8	0.93	0.16
Conscientious	10	0.84	0.35	0.82	0.92	0.20
Rule Following	10	0.89	0.26	0.84	0.90	0.22
Relaxed	10	0.87	0.28	0.85	0.94	0.16
Worrying	9	0.78	0.37	0.88	0.92	0.21
Tough Minded	9	0.80	0.39	0.82	0.92	0.22
Optimistic	10	0.81	0.37	0.8	0.93	0.21
Trusting	10	0.88	0.28	0.81	0.91	0.24
Emotionally Controlled	10	0.86	0.29	0.85	0.90	0.24
Vigorous	10	0.88	0.27	0.75	0.91	0.23
Competitive	10	0.87	0.30	0.86	0.93	0.20
Achieving	10	0.79	0.41	0.79	0.93	0.21
Decisive	10	0.83	0.35	0.8	0.93	0.21
Median		***0.84***	***0.34***	***0.81***	***0.92***	***0.21***

Figure 1: Distribution of average profile scores in the full ipsative OPQ32i, IRT-scored shortened OPQ32 and normative OPQ32n