The Occupational Personality Questionnaire Revolution:

Applying Item Response Theory to questionnaire design and scoring

Anna Brown, *Principal Research Statistician*
Professor Dave Bartram, *Research Director*

Summary

The British Psychological Society, in its 2007 review, places the Occupational Personality Questionnaire (OPQ32) "at the top of the first rank of personality tests, especially those used in occupational settings." The instrument has strong technical and statistical credentials documented in the *SHL OPQ32 Technical Manual*[1] to back this up.

The ipsative version (OPQ32i) is more resistant to the effects of response distortion and 'faking good' than the normative version and is the most frequently used, particularly for selection. While the strength of OPQ32i as an instrument is well-established and documented in the *Technical Manual* and several published papers, and has been independently supported, the use of Classical Test Theory (CTT) for scoring OPQ32i has had some unwanted side effects. While some claims about the problematic properties of ipsative data originate from a fundamental misunderstanding of how such instruments work when they have a large number of scales, these technical limitations have unjustly detracted from the proven qualities of the OPQ32i and its predecessor, OPQ Concept Model 4.2.

After extensive research with the most up-to-date modelling techniques, we concluded that the Classical Test Theory approach simply does not make the most of the information individuals provide in their responses to the forced-choice items. Following recent advances in Item Response Theory (IRT), we have been looking for ways to model forced-choice responding that will provide all the benefits of forced-choice methods without the disadvantages.

SHL has found out how to achieve the benefits of the forced-choice response format without the disadvantages. The breakthrough made by researchers at SHL has been to understand the decision process people go through when responding to forced-choice items and to then model that process using IRT. This paper explains our approach to designing and scoring forced-choice questionnaires using IRT that has enabled a revolutionary improvement in efficiency, accuracy and scaling properties of OPQ32 trait scores, leading to the new OPQ32r. The latent scores recovered from a much reduced number of forced-choice items are superior to the full OPQ32i's ipsative scores and comparable to unbiased normative scores. These advantages are in addition to bias and fake-resistance for which OPQ32i has always been known.

[1] SHL OPQ32 Technical Manual (2006)

PEOPLE PERFORMANCE

OPQ32 Overview

The Occupational Personality Questionnaire was specifically designed to be reliable, valid and fair for the world of work in the 21st century. It is "amongst the best broad spectrum personality tests available – especially for use in occupational settings where a 'surface view' of an individual is needed"[2] . The OPQ32 can be used in a wide variety of occupational situations such as selection, promotion, counselling, development, team-building, organisational change and audits, training-needs analysis and research.

OPQ32 is an occupational model of personality, which describes 32 dimensions of people's preferred style of behaviour at work. It breaks personality down into three domains: Relationships with People, Thinking Styles and Feelings & Emotions. The three domains are joined by a potential fourth, the Dynamism domain, which is related to sources of energy. The OPQ model of personality provides users with a clear framework for interpreting complex patterns of personality.

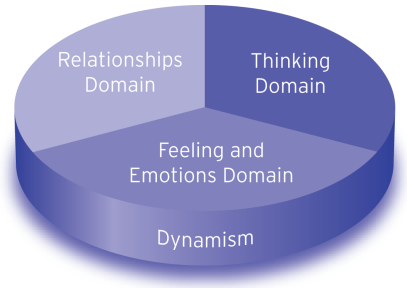


Figure 1: The OPQ model of personality

One of the clear advantages of OPQ32 is that it provides a fine-grained analysis of occupationally relevant personality traits. The 32 narrowband scales also map onto the well-established broadband factors of personality, the 'Big Five'. 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[3]. The comprehensive Big Five mapping allows validity generalisation and comparison with other personality instruments.

The OPQ32 is available in many languages and may therefore be used globally, which is important to multinational organisations and those servicing their HR needs. Detailed investigation into equivalence of language versions, as well as the effects of unsupervised online administration and group differences have been fully analysed and described in the *Technical Manual*[4] .

Advantages of forced-choice format

A major advantage of the OPQ32 is its acceptability to users. The items are clear and transparent, which makes the instrument uncontroversial. At the same time, clear connections between items and the traits they intend to measure has led to the downside that the questionnaire is easier to fake when used as part of an assessment process in which a lot is at stake.

There are two questionnaires using the OPQ model, namely the OPQ32n (*normative*, using single-stimulus format) and OPQ32i (*ipsative*, using forced-choice format). Normative scales have been favoured by traditional research practices and are widely used in personality assessment. However, they are subject to numerous response biases such as acquiescence, leniency or central tendency, and to socially desirable responding. These biases can be a serious threat to validity, particularly in high-stakes situations, where the motivation for impression management is the highest.

SHL pioneered the multi-dimensional forced-choice format in 1981 to create tests that were free from uniform response bias, more robust to impression management distortion or 'faking good' and consequently were more valid in high-stakes situations. OPQ32i reduces response bias by forcing respondents to choose between statements measuring different traits according to the extent to which the statements describe their preferences or behaviour. The forced-choice format has been shown to successfully reduce uniform response bias, and to produce greater operational validity coefficients[5]. It is commonly found that the forced-choice format substantially reduces score inflation compared to the single-stimulus format[6] and is resistant to distortion to its covariance structure[7].

The OPQ32i is one of the best examples of forced-choice tests. The OPQ32i consists of 104 blocks of four statements measuring different traits. 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 recognise weak arguments

C I take care to follow procedures

D I like to spend time with others

[2] SHL OPQ32 Technical Manual (2006)

[3] e.g. Robertson & Kinder (1993); Bartram (2005)

[4] SHL OPQ32 Technical Manual (2006)

[5] Christiansen, Burns & Montgomery (2005); Bartram (2007)

[6] Jackson, Wroblewski & Ashton (2000); Martin, Bowen & Hunt (2002); Christiansen et al. (2005)

[7] Brown (2008a)

SHL Group Limited 2009. All rights reserved. 'SHL' and 'OPQ' are registered trademarks of SHL Group Limited.

Limitations of Classical Test Theory for scoring forced-choice instruments

Despite their clear advantages in reducing bias, forced-choice tests have been criticised 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[8].

In ipsative questionnaires, item scores in the block always add up to the same number regardless of the choices made, and therefore the total test score – the sum of all the blocks – is the same for each individual. Of course, OPQ32i allows for a great variability of scores on the measured scales within each individual profile.

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) raw scores on all scales in a multi-trait questionnaire. Therefore, some 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.

The fact nearly always overlooked by such critics 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, only a very low proportion of respondents will have most of their true scale scores on the same side of the profile, that is, all high or all low[9]. With 30 or more measured scales, norming of ipsative scores is appropriate and intra-individual comparisons can be performed meaningfully. Most importantly, the ordering of people on each trait largely corresponds to their normative ordering. A large study comparing results from OPQ32i and OPQ32n showed that the ordering of respondents on scales derived from the two formats is indeed very similar, and is approaching reliability values. Thus, selection decisions made using either version of OPQ32 would be similar.

Nevertheless, while allowing for a great variability of scale scores within each profile, ipsative OPQ does not have the same variability of average profile locations as the normative version. Put simply, it is impossible to have very high or very low scores on all 32 scales. Despite very low empirical probability of such profiles, this remains a theoretical limitation of ipsative data.

2. Distorted construct validity

The averaged correlation between scales is a negative value in ipsative tests, and approaches zero as the number of scales increases. Again, how much of a problem this is depends on the number of scales in the questionnaire. With 32 scales, the average off-diagonal correlation is only -0.03, allowing for a wide range of both negative and positive correlations between scales[10]. However, scale correlations are depressed in OPQ32i as compared to OPQ32n, which makes it difficult to directly evaluate construct validity of the ipsative version. Moreover, conventional factor-analytic procedures are inappropriate with ipsative data.

3. Lower internal consistency

It is generally agreed that the forced-choice format distorts the internal consistency of instruments. With a large number of measured dimensions, reliabilities as measured by Cronbach's Alpha are depressed. Relying on coefficient Alpha as a valid indicator of internal consistency in the past has led test developers to create questionnaires of potentially excessive length. While six to eight 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 forced-choice OPQ32i. This has an implication on the time it takes to complete the test and on the experience of test-takers.

Having discussed the psychometric properties of ipsative data, it is very important to point out that these properties are not inherent to the forced-choice format itself, but originate from the current way of scoring. The Classical Test Theory (CTT) scoring methodology simply cannot adequately describe the decision-making process behind responding to multidimensional forced-choice items. Modelling this decision process correctly is the key to making the most of this response format.

Item Response Theory as a basis to model forced-choice responding

While some still argue about controversies of ipsative data, the focus of the debate has moved on during the last few years. Nobody who has done serious research with the forced-choice format is in any doubt that it can deliver significant advantages. In addition the fact that the format does not have to be associated with the CTT scoring totally changes the outlook.

[8] e.g.Closs (1996); Meade (2004)
[9] Baron (1996)
[10] Bartram (1996); Baron (1996)

SHL Group Limited 2009. All rights reserved. 'SHL' and 'OPQ' are registered trademarks of SHL Group Limited.

Advances in IRT, specifically in multidimensional IRT, have made it possible to introduce models that deal with some specific types of multidimensional forced-choice measures. We introduced a two-dimensional IRT Preference Model, specifically to work with large forced-choice questionnaires like OPQ32i.

Here we discuss how this model works when applied to OPQ32i. For those interested in the technical details, we refer you to *Development and Psychometric Properties of OPQ32r*[11].

To enable use of this model with forced-choice items, we need to recode responses given to a block of statements. Instead of working with inverted rank orders of the statements, we present them as paired comparisons. This is the standard coding used in the Thurstonian modelling literature[12]. When rank-ordering statements, respondents perform mental pair-wise comparisons of all available options, that is, every statement is compared with every other one. In effect, respondents are asking themselves: "Is statement A more, or less, true of me than statement B?" If you ask yourself that question for item A, comparing it with items B, C and D, and then repeat the same for each of the remaining items in a quad, then you have **six** pairs of comparisons to make: {A,B}, {A,C}, {A,D}, {B,C}, {B,D} and {C,D}. 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.

Having recoded the choices made in a block into paired comparisons with outcomes {A,B}=1 (when A is preferred to B) or {A,B}=0 (when B is preferred to A), we then link those item responses to the underlying personality traits. According to Thurstonian theory of comparative judgment, one statement is preferred to another if its utility is larger for the respondent. In case of personality questionnaires, utilities of statements are caused by strengths of underlying personality traits. When respondents choose between two items, their standing on the two underlying traits will influence the outcome of the comparison. The two-dimensional IRT Preference Model for paired comparisons links them to two latent traits measured by the two items involved in the comparison through a likelihood function. For example, this function assigns high probability to the outcome {A,B}=1 (A is preferred to B) if an individual has a high score on the scale underlying item A, and a low score on the scale underlying item B. How much higher one scale score should be in relation to the other is determined through so-called 'item parameters', established through large sample-based item calibration.

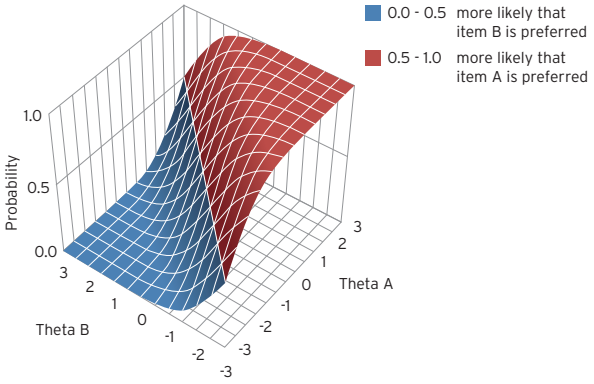


Figure 2: Item characteristic surface for a paired comparison

Reducing the number of items in a block

Research with our IRT model reported at international conferences in last two years shows that CTT scoring substantially underestimates the true reliability of forced-choice instruments. This is true in relation to OPQ32i. Based on these results, it appears that we can cut down the number of items in OPQ32i and still retain good levels of reliability. We could have just cut down the number of quads in the instrument, but there was another consideration.

It is well known that a multidimensional forced-choice format can be cognitively challenging, 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. Unsurprisingly, success in faking multidimensional forced-choice instruments was found to be related to cognitive ability[13]. Better understanding of the decision process behind forced-choice responding offers an explanation of why it is so much more challenging to make choices in a block of four statements: this is because the number of mental comparisons to be performed is 6 for a block of 4 statements, but only 3 for a block of 3 statements.

If one statement is taken out of the block of four, making it a block of three, only **three** paired comparisons have to be performed by the respondent. This makes the completion task less cognitively challenging, and therefore can be completed by people with more diverse educational background. Crucially, this offers a

[11] SHL Supplement to OPQ32 Technical Manual (2009)
[12] Maydeu-Olivares & Bockenholt (2005)
[13] Vasilopoulos et al. (2006)

SHL Group Limited 2009. All rights reserved. 'SHL' and 'OPQ' are registered trademarks of SHL Group Limited.

significant improvement to test-taker experiences. Of course, another added bonus is significant reduction in completion time.

Selecting items providing most information

To select the best, or most informative, items we carefully examined each of the OPQ32 scales. This was first done based on large trials using single-stimulus (normative) format. Each measured scale was examined in relation to its dimensionality and by fitting several IRT models to the data. Items that provided least information (had low discriminations) for their one-dimensional scale were highlighted for possible deletion. The crucial point was that after such a deletion, the scale should be no worse than it was before. The items should reliably measure a coherent one-dimensional construct. Also, the meaning of the scale should remain the same, so we were careful not to reduce the domain measured to a very narrow set of items.

Next, we considered several real samples from the forced-choice completion of OPQ32i. This step was very important. When put in blocks, items can interact with each other in ways that cannot be envisaged from the normative presentation. If almost everybody (or almost nobody) in the sample selects an item in a block, that item provides very little information for all but very extreme trait scores. Examination of the forced-choice responses carried out by fitting the IRT Preference Model generally found the same items as in the normative trial to be less informative, but also revealed a number of additional items that were highlighted for removal.

Then, a judgmental review was performed in order to remove one item from each block, based on the criteria of least information. We were looking to remove an equal number of items from each scale, retaining 9 or 10 items per scale. This step not only involved statistical considerations described previously, but also required detailed expert knowledge of the questionnaire's scales in order to retain items that are important for the scale's construct. This is how we assembled the final version with 104 blocks of 3 items, 312 items in total, with 9 or 10 items per scale.

Estimating IRT parameters and producing individual scale scores

To establish the IRT parameters needed to score individual responses, a very large structural model was fitted, linking responses to all paired comparisons from a large sample to the underlying 32 traits.

Using the established IRT parameters, the IRT Preference Model computes individual scores by working not on a scale-by-scale basis, as the old scoring method did, but on an item-pair-by-item-pair basis, for all scales simultaneously. The likelihood of observing the given outcome of a paired comparison is expressed in terms of the strength of the relevant underlying traits that influence the choice made by the respondent. When you consider we are looking at 32 traits and hundreds of pairs of choices, finding the correct combination of underlying trait scores is a highly complex computer modelling problem. It is the development in the technology for finding the optimal solution that has led to the breakthrough for the new OPQ32r.

Key features of OPQ32r scores recovered from forced-choice ratings

By finding the most probable combination of scale scores to explain the individual choices made in blocks of statements, we produce scores that are no longer ipsative. This is because the new scoring algorithm takes into account the multidimensionality underlying the choices made between items, which the Classical Test Theory approach ignored.

Reliability and standard error of measurement

In Classical Test Theory, a single estimate of reliability is obtained for a scale. Item Response Theory offers a much more comprehensive approach to reliability, assessing it in terms of the amount of information provided by all items on the scale. The crucial difference is that the information actually varies depending on the IRT scale score (called a **theta** score), so we can see how measurement error varies along the whole of the measurement scale. As in all multidimensional IRT models, standard errors for OPQ32r are computed through directional test information for particular theta values in the 32-dimensional space. A composite indicator of reliability can be computed by comparing the average squared standard errors for a sample to the trait score variance. This composite coefficient allows comparison between reliability of the IRT-based scores and the traditional scores.

Comparing reliabilities of traditionally scored OPQ32i, and IRT-scored OPQ32r, it is surprising at first that the new instrument, with its much reduced number of items, provides higher reliability for almost every scale (median 0.84). The explanation lies in the fact that for the first time the true estimates of internal consistency of a forced-choice OPQ have been produced. For years the

SHL Group Limited 2009. All rights reserved. 'SHL' and 'OPQ' are registered trademarks of SHL Group Limited.

test developers and researchers relied on Cronbach's Alpha as a valid indicator of reliability of forced-choice tests, but it is an inappropriate measure because the basic assumptions made are violated in forced-choice tests. It now becomes very clear that the true internal consistency was grossly underestimated for the ipsative OPQ32i. In fact, it was much higher than previously thought and so high that it was possible to remove 25% of items while still retaining great reliability levels!

Comparing individuals

The most controversial and much-debated question is whether scores based on forced-choice responses can resemble normative trait standing. They certainly can with our new approach.

We examined a large sample of people who took both the normative and forced-choice versions of OPQ32, comparing their normative scores with the latent scores recovered from the forced-choice ratings. IRT scores from the new OPQ32r approximate the normative scale scores even better than the ipsative OPQ32i did. Ordering of people based on their normative OPQ32n scale scores and the IRT-scored new OPQ32r is very similar (median correlation 0.70). Most people have profiles with very similar shapes based on the normative and the new forced-choice versions (correlated at 0.7 or higher). The profiles of most people (98%) lie within one sten of each other, and the profiles of 80% are within 0.4 sten or less.

The IRT-based forced-choice scores also show great variability in profile locations, just like the normative scores. It is now possible to get all high, or all low, scores in one profile.

Clearly, the IRT scoring methodology produces scores that are close to normative, in both relative position and absolute location. The forced-choice ratings can provide an accurate indication of absolute trait standing.

Construct validity

For the first time it is possible to recover true correlations between OPQ32 scales using the forced-choice format. We can also apply conventional factor-analytical procedures to the recovered scores, just as we would to normative data. However, there is an added bonus. Unlike normative data, as provided by OPQ32n, no overall response bias is present in the data. This means that the factor structure is much cleaner, and we can now recover a very clear factor structure for OPQ32: the Big Five factors[14] and sample specific factors such as Achievement.

Criterion-related validity

A range of validation studies have been examined in order to evaluate the predictive validity of the new way of scoring responses to the shorter forced-choice OPQ32r. This was done across multiple language versions of OPQ. Correlations between the OPQ scale scores (based on both OPQ32i and the new OPQ32r IRT scores) and performance ratings by multiple raters were computed. The evidence is overwhelming that the shorter OPQ32r preserved the validity of the full OPQ32i, and the validity coefficients for the composite Big Five scores improved significantly.

Benefits of using the IRT-scored OPQ32r

In summary, the new IRT approach to design and scoring of the forced-choice OPQ32 achieves major benefits for both test-takers and test users. The test takers now need **to do less**, spending less time completing the questionnaire without compromising its reliability and validity. At the same time, the test users **get more** – greater precision of measurement, accurate information on absolute trait standing and the true relationships between scales. All the great features of OPQ32i are still there: resistance to bias and impression-management effects, work-relevant dimensions that predict workplace competence, and all the great reports and other materials that are available for the test users.

The benefits are many:

- The triplet format is less cognitively challenging than the quad format, and therefore more appropriate for people with lower education level or reading skills.
- The completion time is reduced by up to 50%.
- The IRT scores show none of the psychometric problems associated with ipsative data, which means that test results can be analysed with all the standard techniques, just like the normative scores.
- The IRT scores provide a good indication of the absolute trait standing, but with none of the uniform response bias often present in normative scores.
- The new OPQ32r is highly reliable and standard errors are now computed for each individual set of scores, giving a precise indication of the error margin for each of the 32 scores reported.
- The famous fake-resistance of OPQ32i is preserved.
- The criterion-related validity of OPQ32i is preserved and, in some areas, is even enhanced.

[14] McCrae and Costa (1987)

SHL Group Limited 2009. All rights reserved. 'SHL' and 'OPQ' are registered trademarks of SHL Group Limited.

References

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.)

Bartram, D. 2005. T*he great eight competencies: A criterion-centric approach to validation.* (Journal Of Applied Psychology, 90(6): 1185-1203.)

Bartram, D. (2007). *Increasing validity with forced-choice criterion measurement formats.* (International Journal of Selection and Assessment, Vol. 15, Issue 3, pp. 263-272.)

BPS. (2007). *Occupational Personality Questionnaire (OPQ32).* (British Psychological Society, Psychological Testing Centre, Test Reviews.)

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

Brown, A. (2008, b). *How to get the best of both worlds: recovering normative scores from ipsative ratings.* (Presented at the Division of Occupational Psychology annual conference, Stratford, UK, January 2008.)

Brown A. (2007). *Structural Equation Modelling of Latent Traits with Forced-Choice Ipsative Data: Theory and Applications.* (Presented in P. Converse (Chair) symposium "Forced-choice measures in selection", Society for Industrial and Organizational Psychology, 22nd Annual Conference, April 27-29, 2007. New York.)

Brown, A. and Bartram, D. (2008, a). *IRT model for recovering latent traits from forced-choice personality tests.* (Presented at SIOP annual conference, San Francisco, 11 April 2008.)

Brown, A. and Bartram, D. (2008, b). *Doing less but getting more: Improving forced-choice measures with IRT.* (Presented at SIOP annual converence, New Orleans, 3 April 2009.)

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

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

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.)

Martin, B. A., Bowen C.C., & Hunt, S. T. (2001). *How effective are people at faking on personality questionnaires?* (Personality and Individual Differences, Volume 32, Issue 2, 19 January 2002, Pages 247-256.)

Maydeu-Olivares, A. and Böckenholt, U. (2005). *Structural equation Modeling of Paired-Comparison and Ranking Data.* (Psychological Methods, vol. 10, 3, 285-304.)

McCrae, R. and 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 (20040, 77, 531-552.)

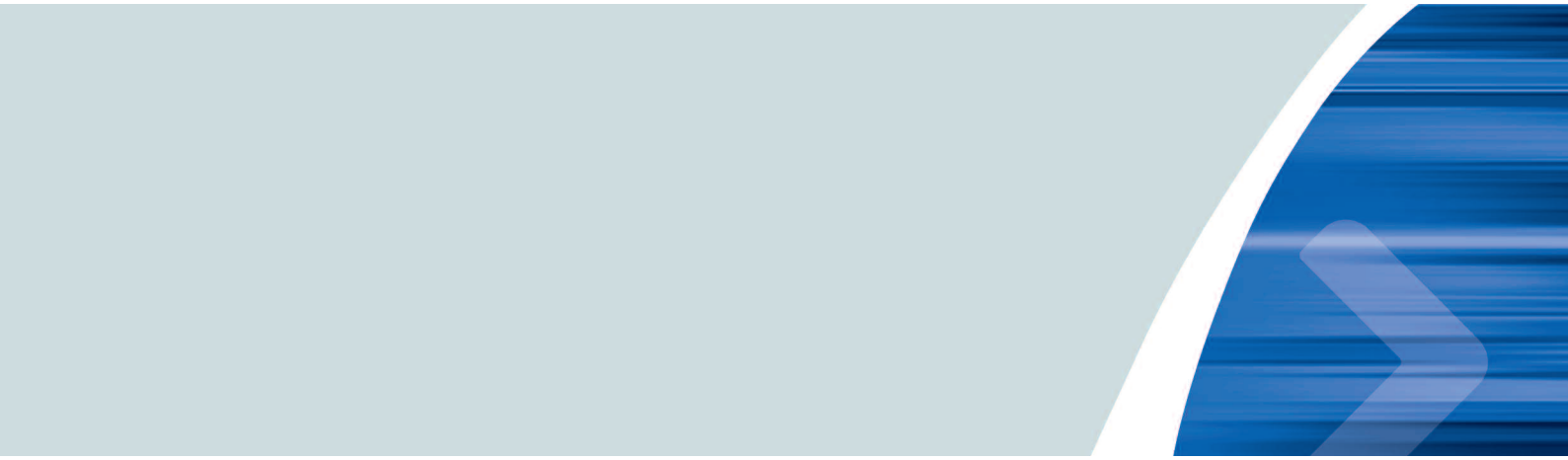
Robertson, I.T. & Kinder, A. (1993). *Personality and job competencies: An examination of the criterion-related validity of some personality variables.* (Journal of Occupational and Organizational Psychology, 66, 225-244.)

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

SHL (2009). *Development of OPQ32r using Item Response Theory. Supplement to OPQ32 Technical Manual.* (Surrey, UK. SHL Group Ltd.)

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.)

SHL Group Limited 2009. All rights reserved. 'SHL' and 'OPQ' are registered trademarks of SHL Group Limited.



www.shl.com

PEOPLE PERFORMANCE

4023