Running Head: Psychometric properties of forced choice and Likert rating scales

A comparison of the psychometric properties of the forced choice and Likert scale versions of a personality instrument

Tina Joubert*[1], Ilke Inceoglu[2], Dave Bartram[3], Kim Dowdeswell[1] and Yin Lin[4]

[1]CEB, Pretoria, South Africa

[2]Surrey Business School, University of Surrey, Guildford, UK

[3]CEB, Surrey, UK & University of Pretoria, South Africa

[4]CEB, Surrey, UK

* Correspondence should be addressed to Tina Joubert, CEB SA, PostNet Suite# 45, Private Bag X32, Highveld Park, 0157, South Africa (e-mail: Tina.Joubert@shl.com).

The present research investigated if an IRT-scored forced-choice personality questionnaire has the same normative data structures as a similar version that uses a 5-point Likert scale instead. The study was conducted using a sample of 349 training delegates who completed both an IRT-scored forced-choice and a normative single stimulus version of the questionnaire. Results largely supported the scaling properties, measurement precision and equivalence of the data structures of the two scoring methods.

The use of personality assessments in an organisational setting, especially for high stakes selection, is mitigated by the substantial evidence that candidates can, and do, intentionally distort their scores on personality measures that rely on Likert rating scales (Heggestad, Morrison, Reeve & McCloy, 2006). Unintentional distortion as a result of response biases relating to acquiescence, central tendency and halo effects also moderate the use of personality assessments (Brown & Maydeu-Olivares, 2013). Christiansen, Burns and Montgomery (2005, p. 268) explain that there is considerable "scepticism in industry regarding the use of self-report measures to facilitate hiring decisions." The most prevalent concern of decision makers regarding personality measures is that applicants can easily distort normative single-stimulus scores (Christiansen et al., 2005). A social desirability scale has traditionally been included in normative questionnaires to attempt to identify cases where large distortion has taken place. In certain questionnaires, a correction based on a social desirability scale score has been incorporated. Ellington, Sackett and Hough (1999) investigated whether using a correction based on the social desirability scale would remove the distortion and concluded that "a social desirability correction is ineffective and fails to produce a corrected score that approximates an honest score (p.155)."

The multidimensional forced-choice item format was introduced in an attempt to minimize response biases as found with Likert type rating scales. To study the effects of distortion Christiansen et al. (2005) conducted a study where they asked candidates to complete both a forced-choice item format as well as a normative single-stimulus personality questionnaire as if they were applying for a sales position. Scores from both formats shifted, but the change was markedly smaller in the forced-choice measure. In such measures, the test taker typically chooses a statement from a block of two or more statements from different dimensions that best describes him or her and a statement that is least like him or her. Forced-choice item formats are more robust to uniform response biases, since it is impossible to endorse every option (Brown & Maydeu-Olivares, 2013). A limitation of forced-choice measures is that, if scored with classical scoring procedures, they produce ipsative data which leads to distorted scale relationships and problematic psychometric properties (Brown & Maydeu-Olivares, 2011; Meade, 2004). Horst (as cited in Clemans, 1966, p. 4) explains the term ipsative by stating "any score matrix, which has the property that the sum of the scores over the attributes for each of the entities is a constant, will be said to be ipsative." This limits comparisons between individuals to rankings and emphasises intra-individual rather than inter-individual differences. Ipsative scores are relative scores and it is impossible for an individual to obtain very high or very low scores on all scales of an ipsative measure (Clemans, 1966). There are a large number of negative values in an ipsative intercorrelation matrix, which result in the average correlation between the scales of an ipsative test having a negative value (Clemans, 1966). Although this average approaches zero with an increase in the number of scales involved, it makes it difficult to evaluate the construct validity of an ipsative instrument (Brown & Maydeu-Olivares, 2013). Conventional factor analysis will not provide meaningful results from ipsative scale intercorrelations (Clemans, 1966). However, the problematic psychometric properties of ipsative data are not a result of the forced-choice item format, but rather the classical test scoring methodology applied to these instruments (Brown & Maydeu-Olivares, 2011).

Brown and Maydeu-Olivares (2011) conducted ground breaking research by applying the Thurstonian Law of Comparative Judgement to provide a modelling framework for forced-choice questionnaires measuring multiple dimensions using Item Response Theory (IRT), which was termed Thurstonian IRT modelling. This framework describes the decision-making process underlying the responses to multidimensional forced-choice items more adequately than the classical scoring approach and “enables straightforward estimation of individual trait scores and test information functions” (Brown & Maydeu-Olivares, 2011, p. 493). Therefore this scoring methodology allows using a forced-choice item format in personality measures. This controls for uniform response biases, while generating normative scale scores (Brown & Maydeu-Olivares, 2013). The Occupational Personality Questionnaire (OPQ32) is based on an occupational model of personality that describes people’s preferred style of behaviour. OPQ32 has 32 primary scales and was specifically developed for use in the world of work. A study conducted by SHL (2013) examined the scaling properties of the OPQ32 by analysing the data of a group of UK based training delegates who completed both the normative single stimulus OPQ32n and the forced-choice OPQ32i, the precursor of the OPQ32r. The OPQ32i is an ipsative version of the OPQ32 utilising a forced-choice item format and is scored using a classical scoring methodology. For the SHL (2013) study, the responses to the OPQ32i were rescored using IRT to produce normative scores similar to those produced by the OPQ32r (for details regarding the IRT rescoring of the OPQ32i refer to SHL, 2013). The IRT scored forced choice scores showed similar psychometric properties to the normative scores.

The aim of the current research is to extend the study conducted by SHL (2013) using South African data. The present study uses the OPQ32r, the actual IRT-scored instrument as used for assessments and not a rescored proxy as in the SHL (2013) study. Normative scores from the classically-scored single-stimulus OPQ32n and the Thurstonian IRT-scored forced-choice OPQ32r were compared. Data structures and scaling properties were analysed to determine whether the two versions measure constructs equivalently.

Construct equivalence, or structural invariance, exists when two instruments measure similar psychological constructs (Van de Vijver & Poortinga, 1997). Van de Vijver and Poortinga (1997) state that the same psychological constructs are obtained between two instruments if the patterns of correlations between the variables are the same across the instruments. In this instance, the study will utilize the OPQ32 for this research as the instrument is available as a single-stimulus normative version (OPQ32n) and an IRT scored forced-choice normative version (OPQ32r). The study will therefore look at how closely the IRT based OPQ32r scores resemble the normative scores of the OPQ32n in a group of South African training delegates who completed both the OPQ32n and the OPQ32r.

Method

Measure

The Occupational Personality Questionnaire (OPQ32) provides information on 32 preferences at work which fall into three domains: Relationships with People, Thinking Style and Feelings and Emotions (SHL, 2013). Originally two versions were developed for the OPQ32, a normative single stimulus version (OPQ32n) and a forced-choice, ipsative version (OPQ32i) (SHL, 2013). The item content and format employed in the two versions is completely different. The normative version requires respondents to rate each statement on a 5-point Likert scale ranging from Strongly Disagree (1) to Strongly Agree (5). The OPQ32n consists of 230 items with an average of seven per scale. The ipsative version requires respondents to choose from sets of four statements (i.e., ‘quads’) one that is ‘Most like me’ and one that is ‘Least like me’. All the statements are positively worded and written in such a way that they are of equal attractiveness. The OPQ32i consists of 104 sets of these quads. As a result of the work by Brown and Bartram (2009) the Thurstonian item response model was applied to OPQ32i forced choice data and normative scale scores were produced. From this an IRT scored version, the OPQ32r, was developed. OPQ32r consists of a subset of OPQ32i items which are presented as triplets (104 blocks of three items). The development of the OPQ32r, evidence for its normative scale properties, and for its construct and criterion validity are described in detail in the OPQ32 technical manual (SHL, 2013). The OPQ32i has been retired.

Sample

The sample (Sample 1) comprised 349 South African OPQ training course delegates who completed the OPQ32n and OPQ32r between 2010 and 2013 online (UK English versions) before attending the OPQ training course. The participants were primarily Human Resources professionals, consultants and individuals working in related fields. The age of the sample ranged from 21 to 63 years with a mean of 32.15 (SD=7.79) with 291 females (83.4%) and 58 males (16.6%)[1].

[1] As the sample in this study (Sample 1) is very specific and may not be reflective of a general applicant sample, it was compared with a large South African sample (Sample 2) of applicants who completed the OPQ32r (N=122333) mostly in high stake settings for selection purposes. Sample 2 was drawn from an online system and included data from January 2010 to April 2014. The OPQ32r mean scores, intercorrelations and internal consistencies were compared with the sample in this study (Sample 1) as well as a candidate sample (Sample 3) as reported in the OPQ32r Technical Manual (2013). The intercorrelation matrices and IRT based reliabilities of Sample 1 in this study are very similar to the applicant Samples 2 and 3. Differences were found in the mean scores which relates to the nature of the samples, where the training sample (Sample 1) obtained higher scores on the Behavioural scale, but lower scores on Data Rational than the applicant sample (Sample 2). The training sample (Sample 1) was also compared with scores obtained by randomly drawing a sub sample (Sample 4) from Sample 2 that is the same size as Sample 1 (Affourtit & Inceoglu, 2014) and that reflects the demographic details (age, gender, ethnicity and education) of Sample 1 (Sample 4, N=349). The intercorrelations and IRT based reliabilities were once again very similar to Sample 1. Although smaller mean score differences were found, there were still large differences on Behavioural and Data Rational. Construct equivalence between the training sample (Sample 1) and the high-stakes sample (Sample 4) were tested using structural equation modelling with EQS6.1. The fit indices indicated excellent fit for the tested model between Samples 1 and 4: CFI = 0.993, RMSEA = 0.018 and SRMR = 0.044.

Unfortunately, item level data for the OPQ32n was only available for 160 individuals in this sample. This is as a result of technical issues experienced during a short period between 2010 and 2011 as the scoring system did not capture the item data after scoring the OPQ32n. This was resolved, but the item level data for some of the cases could not be retrieved.

Results

Reliability

Cronbach alpha coefficients were calculated for the classically scored OPQ32n for the subset with item level data (N=160). For the IRT-scored OPQ32r empirical reliability was computed through the use of IRT test information functions (Brown & Maydeu-Olivares, 2011). The alpha coefficients for the OPQ32n ranged from 0.71 to 0.91 (mean: 0.83, median: 0.85). The empirical reliability for the OPQ32r ranged from 0.67 to 0.92 (mean: 0.83, median: 0.83), indicating that mean and median reliability coefficients for the scales are very similar for the classically scored OPQ32n and the IRT scored OPQ32r (Table 1).

- Insert Table 1 about here –

Individual profiles

The OPQ32n and the OPQ32r profiles for each person were compared. Profile similarities were computed as a correlation between the 32 scores of OPQ32n and OPQ32r for each individual (k=32). These correlations yielded a median profile similarity of 0.73 (Figure 1). For 63% of the OPQ course delegates the two profiles correlated 0.70 or above and for 86% the two profiles correlated 0.60 or above.

- Insert Figure 1 about here -

The distance between profiles was calculated to determine the absolute location of the OPQ32r in relation to the normative OPQ32n. This was done by computing the difference between the average of the standardised normative scores and the average of the forced-choice IRT scores for the 32 scales. The distribution of the profile distance scores is almost normal. Most people's (96%) OPQ32r profiles lie within 0.5 z scores (or standardised scores) from their OPQ32n profiles (equates to one sten), and 60% have their profiles within 0.2 z scores (equates to 0.4 sten) or closer. It is therefore evident that the IRT-scored forced-choice responses not only closely resemble the shape of the normative profiles, but also the absolute location.

Correlations and covariance structures between scales

It is possible to recover scale correlations of the forced-choice OPQ32r without the ipsative distortion as there is no constraint on the overall score recovered from the forced-choice ratings using IRT scoring. The intercorrelations between the OPQ32n scales ranged from -0.62 to 0.61, with 72% falling between -0.20 to 0.20. For the OPQ32r the scale intercorrelations ranged from -0.67 to 0.66, with 69% falling between -0.20 and 0.20. The pattern of scale intercorrelations for each of the two versions is generally similar, with the highest negative correlation obtained for both versions between Conventional and Variety Seeking (r=-0.62 and r=-0.67 respectively). The highest positive correlation for the OPQ32n is between Outgoing and Affiliative and although this is not the highest correlation for the OPQ32r, the correlation coefficient for both is 0.61. The highest positive correlation for the OPQ32r is between Detail Consciousness and Conscientious (r=0.66). For the OPQ32n, the correlation coefficient between Detail Consciousness and Conscientious is 0.49. These two 32 × 32 tables are too large to reproduce here, but are available to interested readers upon request. Correlations between the same OPQ32n and OPQ32r scales range from 0.50 to 0.84 (median: 0.73).

Covariance structures of the 32 scales of the OPQ32n and OPQ32r were compared using structural equation modelling with EQS6.1 (Bentler, 2006)[2]. The model tested was that the covariance matrices for the OPQ32n (classically scored) and the IRT scored OPQ32r were identical[3]. Commonly used fit indices indicated excellent fit for the tested model: CFI = 0.967 (robust method: 0.969), RMSEA = 0.039 (robust method: 0.035) and SRMR = 0.054. A significant Chi-square of 753.866 (robust method: 710.172) with df = 496 was obtained, which is influenced by sample size and a less useful indication of fit (e.g. Byrne, 2006). It is evident that the scaling properties and construct relationships of the OPQ32n are preserved in the IRT-scored OPQ32r. The results strongly confirm the construct equivalence of the classically scored OPQ32n and IRT scored OPQ32r.[4]

Discussion

This study examined the psychometric properties and equivalence of data collected using an IRT-scored forced-choice item format yielding normative data and a normative single stimulus item format. For this purpose, two versions of the OPQ32 in a sample of South African training delegates were utilized: a normative single stimulus format (OPQ32n) and a forced-choice item format (OPQ32r) that is scored using an IRT scoring methodology

[2] The covariance structures of the two OPQ32 instruments were compared using SEM rather than the factor structures. The reason is that the OPQ was not designed on the basis of a factorial structure and hence a good-fitting structure cannot be achieved (Bartram, 2012).

[3] The OPQ32n data deviated from normality as the skewness and kurtosis deviated significantly from zero and the normalised estimate of Mardia’s coefficient was equal to 20.77, which is larger than the cut-off of 5 indicating normality (Bentler, 2005). Although the OPQ32r data was normally distributed, the robust method for analysing the structures was also investigated and reported.

[4] To explore the amount of score variations in OPQ32n due to different scoring methods, Item Response Theory (IRT) scoring was applied to the subsample of N=160 where item response data was available. The graded response model (Samejima, 1969) was fitted to each of the 32 scales respectively. Despite the fact that the instrument was developed completely under the classical scoring methodology, the resulting IRT model fit was acceptable for most scales – 27 out of 32 scales reached a CFI of 0.90 or higher. The resulting IRT-based scores were very similar to the classical scores (correlating to 0.94 or higher), with intercorrelations ranging from -0.62 to 0.60 (69% falling between -0.20 and 0.20) and correlations with OPQ32r ranging from 0.42 to 0.85 (median: 0.74).

yielding normative scale data. Results showed comparable reliabilities, intercorrelations and covariance structures. Furthermore, individual profiles based on the IRT-scored forced-choice responses also resemble the shape of the normative profiles, and were similar in absolute location.

In summary, the results demonstrated that, in low-stakes assessments, the normative single stimulus OPQ32n and the forced-choice IRT scored OPQ32r are equivalent in terms of their correlations and relationships between scales and yield similar data structures. This confirms previous studies that used a different personality instrument (Brown & Maydeu Olivares, 2011) and the precursor version of the OPQ32r (SHL, 2013). Brown and Maydeu Olivares (2011) report simulation studies as well as an empirical study where they used a Big Five questionnaire drawn from International Personality Item Pool (IPIP) items to determine if scores estimated from forced-choice questionnaires using an IRT methodology would reproduce normative latent traits well. They conclude that “the proposed IRT approach allows using the forced-choice format, which reduces certain response biases, while getting the benefits of standard data analysis techniques that users of single-stimulus questionnaires have enjoyed” (Brown and Maydeu Olivares, 2011, p. 498). The simulation studies explore the range of conditions under which the Thurstonian IRT scoring methodology can be applied.

The OPQ32r scores in this study are similar to the normative scores of OPQ32n and do not only resemble the shape of a normative personality profile but also the variability in its location (ipsative scale profiles all have the same location). There is a strong relationship between scores on the two OPQ32 versions, however, the correlations between scales of the two versions are lower than internal consistency reliabilities, suggesting that the constructs measured by the two versions are similar but the instruments are not parallel versions. This may arise for two reasons. Not only do the two instruments employ different response formats, but the content of the items employed in the two versions is different. This may lead to some conceptual differences between constructs.

Considering the results within a measurement invariance framework, the results show a high level of measurement and structural invariance for a low stakes assessment condition. The forced-choice format (OPQ32r) controls for uniform response biases and is therefore recommended for use in contexts where uniform response biases are expected. These include high stakes selection settings and assessments for promotion, where the rank ordering of candidates can be affected, resulting in unfair decisions. Although it is still possible for candidates with higher cognitive ability to distort forced-choice responses, forced-choice item formats provide better control over some response biases that threaten the validity of Likert type normative scales (Christiansen, et al, 2005).

According to Schmit and Ryan (1993) the factor structures of assessment instruments should not be assumed to be constant across testing situations, such as for development or high stakes selection settings. They investigated the factor structure of the NEO Five Factor Inventory and concluded that its factor structure differs across applicant and nonapplicant samples. The training sample used in this study completed the OPQ32n and OPQ32r for development purposes (low stakes) and it was, therefore, deemed necessary to compare the factor structure of this sample with a high stakes applicant sample using structural equation modelling with EQS6.1 (see footnote 1). Since it is recommended that the OPQ32r be used in contexts where uniform response biases are expected, the factor structure of the training sample in the current study was compared with a random sample of applicants drawn from an online system to reflect the sample size and the demographical information of the training sample. For both comparison between low stakes samples and comparisons between low and high stakes samples (see footnote 1) the fit indices indicated an excellent fit, showing that the IRT-scored forced-choice instrument retains its data structure and the constructs hold up in high stakes situations. Different types of uniform response biases are observed in different cultures and have been the focus of recent research (e.g. extreme responding and acquiescence; He & van de Vijver, 2013). A forced-choice format is recommended for use in cross-cultural comparisons as it will eliminate culture-specific, uniform response biases (He, Bartram, Inceoglu & van de Vijver, 2014). Contexts where the Likert scale version might be preferable are where such response biases are not a major concern but where the cognitive complexity associated with use of forced-choice items would be an issue. Comparing three items is more cognitively demanding than rating them one at a time.

References

Affourtit, M. & Inceoglu, I. (2014, July). Construct equivalence using structural equation modelling. Paper presented at the 9th Conference of the International Test Commission, San Sebastian, Spain.

Bartram, D. (2012). Stability of OPQ32 personality constructs across languages, cultures, and countries. In: A. M. Ryan, F.T.L. Leong & F.L. Oswald (eds.), *Conducting multinational research: Applying Organizational Psychology in the workplace (pp.59-89)*. Washington, DC: American Psychological Association.

Bentler, P.M. (2005). *EQS 6 structural equations program manual*. Encino: Multivariate Software.

Brown, A., & Bartram, D. (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., & Maydeu-Olivares, A. (2011). Item Response Modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71(3),* 46-52. doi: 10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2013). How IRT Can Solve Problems of Ipsative Data in Forced-Choice Questionnaires. *Psychological Methods, 18(1),* 36-52. doi: 10.1037/a0030641

Byrne, B.M. (2006). *Structural equation modelling with EQS: Basic concepts, applications and programming.* (2nd edn.). New Jersey: Lawrence Erlbaum.

Christiansen, N.D., Burns, G.N., & Montgomery, G.E. (2005). Reconsidering Forced-Choice Item Formats for Applicant Personality Assessment. *Human Performance, 18(3),* 267-307. doi: 10.1207/s15327043hup1803_4.

Clemans, W.V. (1966). *An analytical and empirical examination of some properties of ipsative measures.* (Psychometric Monograph No.14). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN14.pdf

Ellington, J.E., Sackett, P.R., & Hough, L.M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84(2),* 155-166. doi: 10.1037/0021-9010.84.2.155

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. (2014). Response Styles and Personality Traits: A Multilevel Analysis. *Journal of Cross-Cultural Psychology*. DOI: 10.1177/0022022114534773. http://jcc.sagepub.com/content/early/2014/05/12/0022022114534773.refs.html

He, J., & van de Vijver, F. J. R. (2013). A general response style factor: Evidence from a multi-ethnic study in the Netherlands. *Personality and Individual Differences, 55,* 794-800. doi:dx.doi.org/10.1016/j.paid.2013.06.017

Heggestad, E.D., Morrison, M., Reeve, C.L., & McCloy, R.A. (2006).Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91(1),* 9-24. doi: 10.1037/0021-9010.91.1.9

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77(4)*, 531-551. doi: 10.1348/0963179042596504.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17).* Richmond, VA: Psychometric Society.

Scmit, M.J., & Ryan, A.M. (1993). The Big Five in Personnel Selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78(6)*, 966-974. doi: 10.1037/0021-9010.78.6.966.

SHL. (2013). *OPQ32r Technical Manual version 1.0.* Thames Ditton, UK: SHL Group.

Van de Vijver, F.J.R., & Poortinga, Y.H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*, 21−29.

Table 1 Reliability coefficients and correlations between OPQ32n and OPQ32r scores

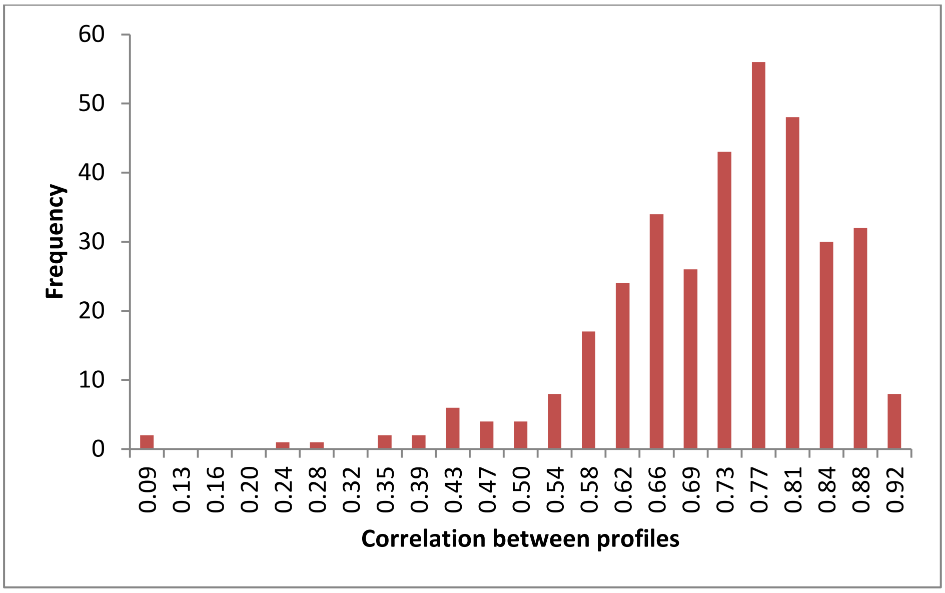
OPQ32 scales	OPQ32n (N=160)	OPQ32r (N=349)	Correlation between OPQ32n and OPQ32r (N=349)
Persuasive	0.88	0.77	0.75**
Controlling	0.87	0.92	0.78**
Outspoken	0.77	0.86	0.73**
Independent Minded	0.71	0.77	0.60**
Outgoing	0.91	0.87	0.84**
Affiliative	0.83	0.85	0.72**
Socially Confident	0.86	0.88	0.80**
Modest	0.86	0.82	0.69**
Democratic	0.74	0.73	0.50**
Caring	0.76	0.82	0.60**
Data Rational	0.81	0.87	0.78**
Evaluative	0.78	0.79	0.62**
Behavioural	0.86	0.77	0.66**
Conventional	0.84	0.66	0.77**
Conceptual	0.83	0.76	0.76**
Innovative	0.87	0.88	0.79**
Variety Seeking	0.80	0.76	0.72**
Adaptable	0.88	0.87	0.71**
Forward Thinking	0.79	0.86	0.64**
Detail Conscious	0.77	0.88	0.79**
Conscientious	0.73	0.81	0.63**

Rule Following	0.88	0.89	0.75^{**}
Relaxed	0.88	0.88	0.74^{**}
Worrying	0.88	0.78	0.78^{**}
Tough Minded	0.90	0.78	0.71^{**}
Optimistic	0.87	0.79	0.73^{**}
Trusting	0.86	0.89	0.68^{**}
Emotionally Controlled	0.87	0.87	0.77^{**}
Vigorous	0.88	0.89	0.63^{**}
Competitive	0.85	0.84	0.70^{**}
Achieving	0.74	0.78	0.66^{**}
Decisive	0.84	0.82	0.77^{**}

** $p \leq 0.01$

* $p \leq 0.05$

Figure 1 Profile similarities: distribution of profile correlations


Frequency
Correlation between profiles