Online testing: Mode of administration and the stability of OPQ 32i scores

Dave Bartram & Anna Brown

SHL Group, Research Division

The Pavilion
1 Atwell Place
Thames Ditton
Surrey, KT7 0NE

e-mail for correspondence: Dave.Bartram@shlgroup.com

Originally submitted to the International Journal of Selection and Assessment
October 2002
Review received April 2003

Revised version for Information Exchange
September 2003

Online testing: Mode of administration and the stability of OPQ 32i scores

Abstract

This study explores the equivalence of web-based administration with no local supervision and traditional paper-and-pencil supervised versions of OPQ32i (the ipsative format version of the Occupational Personality Questionnaire). Samples of data were collected from a range of client projects and matched in terms of industry sector, assessment purpose (selection or development) and candidate category (graduate or managerial/professional).

The analysis indicates that lack of local supervision in high stakes situations has little if any impact on scale scores. At worst, some scales appear to show shifts of less than quarter of an SD, with most scales showing little if any change. Analysis in terms of the Big 5 show differences of less than 0.2 of an SD. Scale reliabilities and scale covariances appear to be unaffected by the differences between the supervised and unsupervised administration conditions.

Purpose

The present research was carried out to see whether data obtained from online administration of the OPQ32i personality inventory under conditions where individuals responded without formal supervision was subject to any biases compared to the more traditional supervised paper-and-pencil mode of administration. The key issue for the present research was one of how people behave under real rather than laboratory conditions. The research is therefore based on data collected from 'in vivo' use of the instrument.

Traditionally, paper-and-pencil and computer-based versions of personality inventories have been administered either by a test administrator or in the presence of a 'proctor'. More recently, all of the personality inventories that are widely used in occupational assessment have become available in web-based versions. This move towards availability on the Internet has been accompanied by an abandonment of the traditional requirement for local supervision. A review of online personality instruments shows that all the major instruments have been made available for administration under conditions where a unique log on is required for access but without any requirement for local supervision.

Lievans and Harris (2003) note that "initial evidence seems to indicate that measurement equivalence between web-based and paper-and-pencil tests is generally established. In addition, no large differences are found between supervised and unsupervised testing. Again, these results should be interpreted with caution because of the small number of research studies involved." The present paper adds to the research on this issue by examining whether lack of local supervision, results in any changes in the quality of data obtained compared with traditional paper or computer-based presentation in the presence of a test administrator.

Method

Measures

The OPQ32i was specifically designed to be resistant to the effects of response distortion and 'faking good'. OPQ32i consists of 416 items measuring 32 personality scales. The items are arranged in 104 blocks of four (ipsative, forced-choice format). For each block the test-taker has to choose one item as being 'Most like me' and one as 'Least like me'. Previous research (Martin, Bowen, & Hunt, 2002) has shown this format to be more resistant to faking than the 'normative' likert-scaled version of the OPQ32 model: OPQ32n. The study focuses on OPQ32i as this is the version of OPQ32 used most widely throughout the world.

Both OPQ32n and OPQ32i are based on the same OPQ 'Concept Model' (SHL, 1999). These instruments have replaced the earlier 30-scale OPQ CM5.2 and OPQ CM4.2, respectively. Details of the instrument and technical data can be found in SHL (1999) and an independent review of the OPQ32 instruments is presented in Drakeley & Lindley (2001). Factor analyses of the 32 OPQ scales show a close relation to the 'Big Five' personality factors, though a rather better fit is usually found to a six-factor model, with 'Achievement Motivation' separated out from Conscientiousness (Matthews and Stanton, 1994; SHL, 1999). 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. Saville et al, 1996; Robertson & Kinder, 1993; SHL, 1989; 1995).

Design issues

The use of 'in vivo' methodology creates the need for introducing sampling controls if a clear answer is to be found to the main research question. This is not a trivial problem to solve. First, there will tend to be many reasons why different

samples of people will have different mean personality profiles. Mostly these arise from real differences between the groups of people. Thus, samples of applicants for Employer X may genuinely differ from samples for Employer Y: they may have different patterns of self-selection, or one sample may be graduates and another experienced managers.

To counter these problems, the approach adopted was to identify a number of pairs of matched samples, with one sample in each pair having had online unsupervised administration and the other offline supervised. In this way, it is possible to isolate effects that are consistently related to the issue in question: traditional supervised administration versus online unsupervised administration.

Data samples

Samples of online (unsupervised) OPQ32i administrations were collected from the 'SHL Solutions' web server, and followed up to identify the type of test taker (graduate, manager or professional), the purpose of testing (external selection, promotion or development) and whether the session was supervised or not.

Comparison samples were obtained from SHL UK Consultancy and SHL UK Bureau databases, which contain paper-and-pencil supervised session data for a wide range of groups. In addition, comparison groups were also sought from other countries using English versions of OPQ32i.

Most of the samples (see Table 1) were private sector, with four of them being global financial organisations (A, B, F, J). C was an independent not-for-profit centre.

Insert Table 1 about here

Procedure

The analysis involved comparing unsupervised web-based administration samples with paper-and-pencil supervised samples. The first analysis examines "Effect Size" of unsupervised web administration samples in comparison with matching supervised samples. Cohen (1977) has suggested that an effect size of 0.20 can be considered small, 0.50 considered medium, and 0.80 considered large. The second analysis explores test reliability in supervised and unsupervised conditions. Finally the effect of web-based remote administration on the pattern of scale inter-correlations is examined.

Results

Managerial and Professional samples

There were three pairs of matching samples:

1. Sample A (Hong Kong, financial sector, supervised, paper-and-pencil, N=610) compared with Sample F (the same client organisation, controlled web-based, N=154). Managerial and professional.
2. Sample B (supervised, paper-and-pencil, N=116) compared with Sample G (controlled web-based, N=146). These are managerial and professional level, UK, financial sector but different organisations.
3. Sample C (supervised, paper-and-pencil, N=100) compared with Sample H (controlled web-based N=281), managerial and professional, UK, different organisations providing management career solutions.

Of the above three pairs, the best-matched is the first. Samples A and F are comparable in all respects apart from the paper-and-pencil sample having been assessed the year prior to the web-based sample.

Table 2 shows differences (d values) between the three pairs of samples. Note that a negative value means that online unsupervised scores are lower than supervised; a positive value means that unsupervised scores are higher than supervised.

Insert Table 2 about here

The largest negative difference for Hong Kong samples (A and F) is -0.23 (scale Conceptual) and the largest positive difference is +0.24 (scale Tough Minded). These differences are statistically significant due to large samples sizes (N=610, N=154), but are small in Cohen's classification. They would not have any material effect on the interpretation of candidates' score profiles.

The two pairs of UK managerial and professional samples are not as well matched as the Hong Kong pair because they do not represent candidates drawn from the same candidate pool, which we would expect to be the case if we have two samples from the same organisation. So mean differences for those samples will be affected by 'real' sample differences as well as any potential systematic biases. It can be seen that mean differences are generally larger than for the Hong Kong pair and for some scales reach d=0.5. However, scales that show the biggest differences in one pair of samples do not show any difference or show opposite sign difference in others (see, for example, the scale Relaxed).

Finally, it can be seen that the weighted averages of the difference scores (unsupervised - supervised) for the three pairs of managerial and professional samples are no bigger than d=0.27, with most scales showing little if any difference. When

scored in terms of the Big 5 scales, the differences range from d=0.16 for Conscientiousness to d=-.15 for Openness to new experience.

Graduate samples

There are two UK pairs of matching samples for graduates:

1. Sample D (supervised, paper-and-pencil, N=99) compared with Sample I (controlled web-based, N=96). These represent graduate recruitment in different organisations for sales, marketing and client relationship management positions.
2. Sample E (supervised, paper-and-pencil, N=202) compared with Sample J (controlled web-based, N=91). These are both graduate external selection, different organisations with similar selection criteria.

Table 3 shows differences between the two graduate pairs of samples, unsupervised compared with supervised. Weighted average differences between unsupervised and supervised graduate samples are generally no bigger than d=0.25, except scale Conceptual, where the weighted average is equal d=-0.43. In terms of Big 5 scales, as for the managerial samples, the differences range over no more than one sixth of an SD: from d=0.15 for Neuroticism to d=-.18 for Openness to new experience.

Insert Table 3 about here

Reliability of controlled-mode (unsupervised) administration

It was not possible to obtain reliability estimates for the Paper-and-pencil samples, as no item response data was available. However, ALPHA coefficients from the OPQ32i standardisation sample (SHL, 1999) can be used as the basis for checking the reliability of the web-based data sets. This standardisation sample was assessed under traditional supervised paper-and-pencil conditions.

Table 4 shows the ALPHA coefficients and Standard Errors of Measurement (SEm). ALPHA values for the web-based samples have been computed for each sample and then averaged across the seven samples. The SEm values are all based on sample raw scores. While there is some indication of slightly lower reliability for the web-based data (overall average alpha =0.77 for web-based and 0.80 for the standardisation sample), the SEms are about the same. In fact the average SEm is slightly better (i.e. smaller) for web-based (2.07) than it is for supervised paper-and-pencil (2.12). This is due to differences in the raw scale score SDs for the two conditions.

Insert Table 4 about here

This analysis suggests that web-based controlled-mode (unsupervised) administration does not compromise scale reliability or measurement accuracy.

Similarity of scale inter-correlations

A simple comparison can be made between inter-scale correlations in paper-and-pencil supervised samples and web-based samples. When the average inter-correlation of the 8 paper-and-pencil sample correlation matrices was compared with the average of the 7 web-based sample correlation matrices (excluding the main diagonal from the averaging), the average absolute difference was 0.044, with an SD of 0.033.

More formal comparison of covariance structures was carried out using Structural Equation Modelling with AMOS. As the ipsative version of OPQ32 was being used, it does not make sense to compare factor structures as the constraints present in the ipsative model can introduce artefactual instability in such solutions. However, it is possible to directly compare scale covariance structures of samples with the prior removal of one scale from the correlation matrix, so that the degrees of

freedom (i.e. 31 in this instance) become equal to the number of scales. For the purpose of testing the similarity of the covariances it does not matter which scale is deleted, so long as the same one is deleted from all samples. The model tested was that all the correlation matrices were samples from the same population. To achieve this, the model was constraining so that each scale pair inter-correlation was equal in each pair of samples tested. We tested 3 groups of samples:

1. The five matched pairs of samples (A-B, B-G, C-H, D-I, and E-J),
3. Three pairings of the three supervised managerial samples (A-B, A-C, B-C), and the pairing of the two supervised graduate samples (D-E).
4. Three pairings of the three unsupervised managerial samples (F-G, F-H, G-H), and the pairing of the unsupervised online graduate samples (I-J).

There are a number of statistics that can be used to measure how adequately this hypothesized model describes the sample data. The comparative fit index (CFI; Bentler, 1990) ranges from zero to 1.00 and provides a measure of complete covariation in the data. Although a value >0.90 was originally considered representative of a well-fitting model, a revised cut-off value close to 0.95 has recently been advised (Hu & Bentler, 1999). Another fit measure is the root mean square error of approximation (RMSEA). This index has recently been recognized as one of the most informative criteria in covariance structure modelling; values less than 0.05 indicate good fit (Byrne, 2001).

Table 5 shows goodness-of-fit statistics for the model constraining all intercorrelations to be equal in each sample. Average CFI was 0.949 for pairs of matched samples, 0.951 for pairs of supervised samples and 0.934 for pairs of unsupervised samples; with RMSEA 0.031, 0.030 and 0.034 respectively.

Insert Table 5 about here

This analysis suggests that relationships between scales are not affected by mode of supervision as the model fits equally well for all pairs of samples, across and within different modes.

Conclusions

Data sets obtained from web-based unsupervised controlled-mode administration appear to have comparable psychometric properties to paper-and-pencil supervised data in terms of reliability and relationships between scales. That is encouraging as it implies there is no distortion to the instrument itself. The result of the comparisons between means for different groups is also encouraging as it shows that there are only small differences between supervised and the unsupervised samples on some scales. This implies that the same norms can (and should) be used for both conditions.

Comparisons for the best-matched samples, suggest that local supervision could affect about half a dozen scales by at most plus or minus one quarter of an SD. In practical terms this is not excessive, and would have relatively little impact on interpretation if the same norms were used for both supervised and unsupervised administration. We also have to note the caveat that even these differences may be genuine sample difference effects rather than supervision effects.

Comparisons in terms of the Big 5 scales indicate effects of less than one sixth of an SD. Those who complete OPQ32i under web-based unsupervised conditions tend to produce scores that are slightly higher on Conscientiousness (d=0.14), lower on Extraversion (d=-0.09), Openness to new experience (d=-0.13) and Agreeableness (d=-0.06), with no difference on Neuroticism (d=-0.01). None of these differences would be sufficient to have any substantive impact on the interpretation of a profile.

References

Bentler P.M. (1990). Comparative fit index in structural models. *Psychological Bulletin, 107*, 238-246.

Byrne, B.M. (2001). *Structural Equation Modeling with AMOS. Basic concepts, applications and programming*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Cohen, J. (1977). *Statistical power analysis for the behavioural sciences (revised edition.)* New York: Academic Press

Drakeley, R., & Lindley, P. (2001). Occupational Personality Questionnaire – Concept Model Questionnaires (OPQ-C): Update review as OPQ32. In Lindley, P. (Ed.). *Review of Personality Assessment Instruments (Level B) for use in Occupational Settings.* Leicester, UK: BPS Books.

Hu, L.-T., & Bentler, P.M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling: A Multidisciplinary Journal, 6*, 1-55.

Lievans, F. & Harris, M.M. (2003). Research on Internet Recruitment and Testing: Current Status and Future Directions. In I. Robertson, & C. Cooper (Eds), *The International Review of Industrial and Organizational Psychology*, Chichester, England: Wiley.

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

Matthews, G. & Stanton, N. (1994). Item and scale factor analyses of the OPQ. *Personality and Individual Differences,* 16, 5, 733-743.

Robertson, I.T. & Kinder, A. (1993). Personality and job competences: An examination of the criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology*, 65, 225-244.

Saville, P., Sik, G., Nyfield, G., Hackston, J. & MacIver, R. (1996) - A demonstration of the validity of the Occupational Personality Questionnaire (OPQ) in the measurement of job competencies across time and in separate organisations. *Applied Psychology: An International Review*, 45, 243-262.

SHL (1989). *Validation Review*. Thames Ditton, UK: SHL Group plc

SHL (1995). *Second Validation Review*. Thames Ditton, UK: SHL Group plc

SHL (1999). *OPQ32 Manual and User's Guide*. Thames Ditton, UK: SHL Group plc

Table 1. Data sets examined in this analysis (Man&Prof = managerial and professional).

Paper-and -pencil samples (local invigilator present)

Sample	old	Country	N	Testing Purpose	Group
A	A	Hong Kong	610	Selection + Development	Man&Prof
B	E	UK	116	Development	Man&Prof
C	H	UK	100	Selection+Development	Man&Prof
D	C	UK	99	External selection	Graduates
E	B	UK	202	External selection	Graduates

Online samples (no local invigilator present)

Sample	old	Country	N	Testing Purpose	Group
F	K	Hong Kong	154	Selection + Development	Man&Prof
G	P	UK	146	Development	Man&Prof
H	N	UK	281	Selection+Development	Man&Prof
I	L	UK	96	External selection	Graduates
J	M	UK	91	External selection	Graduates

Table 2. Effect Sizes (d values) for three managerial/professional pairs of samples (unsupervised – supervised) for OPQ scales and Big 5.

		Effect Sizes (d values)			
	Scales	Samples A-F	Samples B-G	Samples C-H	Weighted Average
1	Persuasive	-0.02	-0.03	0.21	0.03
2	Controlling	-0.05	-0.08	-0.27	-0.10
3	Outspoken	-0.22	-0.13	0.06	-0.14
4	Independent minded	0.05	-0.11	-0.11	-0.02
5	Outgoing	-0.06	-0.10	-0.21	-0.10
6	Affiliative	0.01	0.18	-0.10	0.02
7	Socially Confident	0.08	-0.25	0.04	0.00
8	Modest	0.08	0.21	-0.04	0.08
9	Democratic	-0.07	-0.02	-0.34	-0.12
10	Caring	-0.13	0.19	-0.32	-0.11
11	Data Rational	0.16	0.17	0.67	0.27
12	Evaluative	-0.01	0.07	0.15	0.05
13	Behavioural	-0.07	0.07	0.00	-0.03
14	Conventional	0.18	0.31	0.10	0.19
15	Conceptual	-0.23	0.25	-0.14	-0.12
16	Innovative	0.12	-0.20	-0.07	0.02
17	Variety Seeking	-0.08	-0.28	-0.17	-0.14
18	Adaptable	-0.09	0.04	-0.13	-0.07
19	Forward thinking	-0.03	-0.10	-0.11	-0.06
20	Detail Conscious	0.16	0.29	0.54	0.27
21	Conscientious	0.04	0.17	0.42	0.15
22	Rule Following	0.01	0.50	0.04	0.12
23	Relaxed	0.08	-0.52	0.02	-0.05
24	Worrying	-0.06	0.32	-0.22	-0.02
25	Tough Minded	0.24	-0.08	0.12	0.15
26	Optimistic	0.03	0.02	-0.42	-0.07
27	Trusting	0.14	-0.14	-0.55	-0.08
28	Emotionally Controlled	0.10	0.25	0.08	0.12
29	Vigorous	-0.01	-0.26	0.15	-0.03
30	Competitive	-0.10	-0.27	0.56	0.02
31	Achieving	-0.04	-0.48	0.20	-0.08
32	Decisive	-0.19	-0.02	-0.38	-0.20
	Neuroticism	-0.13	0.31	-0.12	-0.04
	Extraversion	-0.08	-0.16	-0.11	-0.10
	Openness	-0.16	-0.12	-0.14	-0.15
	Agreeableness	-0.05	0.06	-0.38	-0.10
	Conscientiousness	0.08	0.13	0.41	0.16

Table 3. Effect Sizes (d values) between two graduate pairs of samples (unsupervised – supervised) for OPQ scales and Big 5.

		Effect Sizes (d values)		
	Scales	Samples D-I	Samples E-J	Weighted Average
1	Persuasive	0.17	-0.14	-0.02
2	Controlling	0.26	0.29	0.28
3	Outspoken	-0.31	0.10	-0.07
4	Independent minded	-0.03	0.03	0.01
5	Outgoing	-0.18	-0.15	-0.16
6	Affiliative	-0.27	-0.17	-0.21
7	Socially Confident	-0.39	-0.15	-0.25
8	Modest	0.02	-0.08	-0.04
9	Democratic	-0.25	-0.01	-0.10
10	Caring	0.13	0.22	0.18
11	Data Rational	0.38	0.02	0.16
12	Evaluative	-0.18	-0.01	-0.08
13	Behavioural	0.03	0.21	0.14
14	Conventional	0.35	-0.08	0.09
15	Conceptual	-0.69	-0.26	-0.43
16	Innovative	0.08	-0.39	-0.20
17	Variety Seeking	-0.03	-0.01	-0.02
18	Adaptable	0.33	0.17	0.23
19	Forward thinking	0.56	-0.07	0.18
20	Detail Conscious	0.07	0.13	0.10
21	Conscientious	-0.01	-0.22	-0.14
22	Rule Following	0.02	0.03	0.02
23	Relaxed	-0.41	0.04	-0.14
24	Worrying	0.21	0.09	0.14
25	Tough Minded	-0.06	-0.28	-0.20
26	Optimistic	-0.03	0.29	0.16
27	Trusting	0.18	0.12	0.14
28	Emotionally Controlled	-0.09	0.11	0.03
29	Vigorous	-0.16	0.23	0.07
30	Competitive	-0.09	-0.14	-0.12
31	Achieving	-0.20	-0.17	-0.18
32	Decisive	0.31	0.20	0.24
	Neuroticism	0.23	0.11	0.16
	Extraversion	-0.07	-0.08	-0.08
	Openness	-0.35	-0.06	-0.18
	Agreeableness	0.05	0.14	0.10
	Conscientiousness	-0.01	0.01	0.00

Table 4.
ALPHA coefficients and Standard Error of Measurement (SEm) for the UK Standardisation Sample (supervised paper-and-pencil) and the combined online (unsupervised) sample.

		Combined web-based sample, N=768		Standardisation Sample, N=807	
	Scale	**Alpha**	**SEm**	**Alpha**	**SEm**
1	Persuasive	0.82	2.08	0.81	2.18
2	Controlling	0.81	2.03	0.87	2.13
3	Outspoken	0.75	2.17	0.76	2.35
4	Independent minded	0.69	2.21	0.72	2.28
5	Outgoing	0.81	2.08	0.85	2.22
6	Affiliative	0.79	1.88	0.82	2.03
7	Socially Confident	0.75	1.96	0.83	2.12
8	Modest	0.83	1.93	0.81	2.06
9	Democratic	0.70	2.15	0.68	2.16
10	Caring	0.73	1.98	0.78	2.04
11	Data Rational	0.87	1.98	0.88	2.02
12	Evaluative	0.64	2.14	0.67	2.18
13	Behavioural	0.80	2.10	0.82	2.22
14	Conventional	0.72	1.99	0.74	2.15
15	Conceptual	0.76	2.19	0.79	2.32
16	Innovative	0.87	1.91	0.88	2.00
17	Variety Seeking	0.73	2.16	0.72	2.17
18	Adaptable	0.82	2.07	0.82	2.06
19	Forward thinking	0.78	1.95	0.75	2.09
20	Detail Conscious	0.77	2.17	0.80	2.35
21	Conscientious	0.77	1.86	0.82	1.94
22	Rule Following	0.83	1.83	0.84	1.95
23	Relaxed	0.78	1.90	0.85	2.08
24	Worrying	0.87	1.90	0.88	2.02
25	Tough Minded	0.69	2.08	0.82	2.12
26	Optimistic	0.79	2.01	0.80	2.06
27	Trusting	0.78	1.89	0.81	2.00
28	Emotionally Controlled	0.80	1.97	0.85	1.95
29	Vigorous	0.73	2.03	0.75	2.19
30	Competitive	0.87	2.07	0.86	2.16
31	Achieving	0.73	2.11	0.79	2.22
32	Decisive	0.81	2.17	0.80	2.14
	AVERAGE	**0.78**	**2.03**	**0.80**	**2.12**

Table 5.
Goodness-of-fit Statistics for a Model constraining each sample within pairs to have equal scale intercorrelations.

Mode	Samples	Group	CFI	RMSEA
Matched	A-F	Man&Prof	0.974	0.023
Matched	B-G	Man&Prof	0.946	0.031
Matched	C-H	Man&Prof	0.942	0.034
Matched	E-J	Graduates	0.944	0.031
Matched	D-I	Graduates	0.938	0.034
Matched Average			**0.949**	**0.031**
Supervised P&P	A-B	Man&Prof	0.959	0.029
Supervised P&P	A-C	Man&Prof	0.963	0.028
Supervised P&P	B-C	Man&Prof	0.913	0.039
Supervised P&P	D-E	Graduates	0.968	0.024
Supervised Average			**0.951**	**0.030**
Unsupervised online	F-G	Man&Prof	0.919	0.038
Unsupervised online	F-H	Man&Prof	0.934	0.035
Unsupervised online	G-H	Man&Prof	0.962	0.027
Unsupervised online	I-J	Graduates	0.921	0.037
Unsupervised online Average			**0.934**	**0.034**
Grand Average			***0.945***	***0.032***