Personality without borders

Do questionnaire languages and smart phones bias results?

Rob Bailey, MSc., OPP Ltd

With increased globalisation and use of internet technology affecting workforce mobility and online access to psychometrics, there is greater demand upon psychologists and test-providers to perform multinational and multilingual personality assessment. It is not enough to assume that assessment in one language will be equivalent to another: there are linguistic, cultural and psychological reasons why results might differ.

Similarly, as mobile devices become a major route to internet services, it would be risky to assume that putting a personality instrument on a smart phone app will yield personality measurement equivalent to administration of the same assessment over a website.

Introduction

With increased globalisation and use of internet technology affecting workforce mobility and online access to psychometrics, there is greater demand for psychologists and test providers to perform multinational and multilingual personality assessment. It is not enough to assume that assessment in one language will be equivalent to another: there are linguistic, cultural and psychological reasons why results might differ. Similarly, as mobile devices become a major route to internet services, it would be risky to assume that putting a personality instrument on a smart-phone app will yield personality measurement equivalent to administration of the same assessment over a website.

Support for the equivalence of multilingual personality measurement has been found for several questionnaires, e.g. Bartram (2008). This current research was necessary to establish that equivalence can be found also in the 16PF® questionnaire, rather than assume the research of others generalises to the 16PF assessment. It was also important for a specific reason: the nature of 16PF questions differs somewhat from many other questionnaires, which are often more simple in the phrasing of their items. An example of the more idiosyncratic items for the 16PF is: "In dealing with people, it's better to: A. 'put all your cards on the table'. B. ? C. 'play your hand close to your chest'". Such an item creates a greater challenge for translators to create a strictly equivalent item in the new language.

As for the medium of questionnaire administration, only recently have the first studies of equivalence in web vs. smart phone administration been published, e.g. Illingworth et al (2014). This current research was done to establish whether or not these are generalisable findings.

Established methods for evaluating equivalence are:

- Differential Item Functioning (DIF) to check that individuals who have the same level of a particular trait will answer the questions in the same way, irrespective of a supposedly irrelevant factor (e.g. language or medium in which the test is presented)
- Analysis of scale level differences (i.e. not just at the item level)
- Confirmatory Factor Analysis (CFA) to check invariance in the factor structure across test versions.

In addition to researching equivalence, the findings were compared to other differences that can be expected in personality data; these are differences between genders, ethnic minorities and whites, and people of different occupations. If differences are found to be greater between languages or nations (either through measurement bias or genuine personality variation) than those found between occupational groups, then this would render a questionnaire unsuitable for international use (particularly recruitment), as any attempt to recruit for occupationally relevant traits could be overwhelmed by bias associated with the languages or nationality.

16PF questions differ from many other questionnaires, which are often more simple in the phrasing of their items

As the intention of this work was to lead to an international norm group for the 16PF Questionnaire, for this to be ethically viable, the following hypotheses would need to be met:

1. Any DIF between languages will not affect measurement substantially at a scale level

2. Differences in the means of the individual traits, between languages, mostly will be of small effect size

3. The language differences will be of a similar size to ethnic and gender differences

4. The differences in means for the languages will be smaller than the differences found for different occupational groups

5. The results of web and mobile administration of the 16PF Questionnaire will be equivalent.

The innovative qualities of this work were to: evaluate a personality questionnaire with more complex items than some already studied; explore the differences between mobile and web administrations; and compare the language results with a range of other demographic and occupational differences, in order to give a clear comparison and context for the size of each difference.

Method

Language Equivalence Study

The analysis was at 3 levels: item level (DIF, using the lordif package in R); scale level (t-tests); factor structure level: with Exploratory Factor Analysis (EFA). Questionnaire administration was online, collected via client use of the 16PF.

Questionnaire, mostly assessing working-age managers/professionals who were applying for jobs. From this data, sets of 500 people were 'randomly' chosen, with the following constraints: for each language sample N=500, 50% female, 50% male, with proportional representation of ethnic minorities, where possible.

Data from the UK English questionnaire was used as a benchmark for the DIF analysis. This is because UK and US questionnaires differ by only a few words and both have been used as the source for translations into other languages.

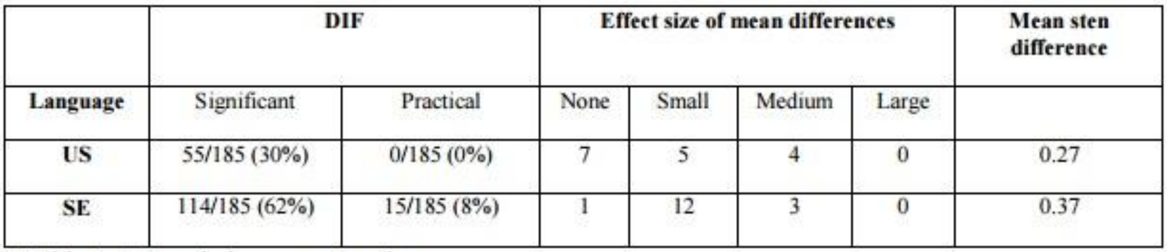
Mobile vs. Web Study

Data were collected from public users of a short form of the 16PF Questionnaire (62 items), who completed the questionnaire for free and voluntarily to learn more about their own personality. The questionnaire was available via a website and via a smart-phone app (for both iPhone and Android). 330 individuals were selected from app users and 330 from web users; samples were balanced to each include 165 men and 165 women.

Results

Languages

Significant DIF was found in a high number of items for each language; for example, 30% of items when comparing UK vs. US data, and 62% for UK vs. Sweden. These two comparisons are presented here as examples, due to little space to explore all language results. The US and Sweden examples are chosen as they represent different ends of the differences found. However, once the effect size of the DIF was estimated in lordif, a very different pattern was found: very few items showed DIF big enough to have created a practical effect at the scale level.



	DIF		Effect size of mean differences				Mean sten difference
Language	Significant	Practical	None	Small	Medium	Large	
US	55/185 (30%)	0/185 (0%)	7	5	4	0	0.27
SE	114/185 (62%)	15/185 (8%)	1	12	3	0	0.37

Table 1: Example language results

T-tests showed many significant differences in mean scale scores across languages; however, again, effect sizes (Cohen's d) were mostly small. When all language data were pooled to create an international norm group, stens for each language could be calculated. Following this, it was possible to see how much, on average, each language differed from the mean for all languages. The average difference from the mean was 0.33 stens (the Standard Deviation is approximately 2 stens on the 16PF traits). The average sten differences for each language from the international mean were as follows: UK 0.22; US 0.27; DK 0.29; NE 0.33; FR 0.34; NO 0.35; DE 0.35; SE 0.37; IT 0.38; ES 0.42.

Exploratory Factor Analysis of 16PF data normally produces an approximation of the Big Five factor structure. However, as the samples were mostly managerial and professional applicants, the data showed a pattern of more socially desirable responses. This meant EFA produced factors grouped more by socially desirable traits than the Big Five. The plan to use Confirmatory Factor Analysis to fit the Big Five model to each language will be replaced in the near future by a different CFA approach: looking for invariance in the relationships between traits.

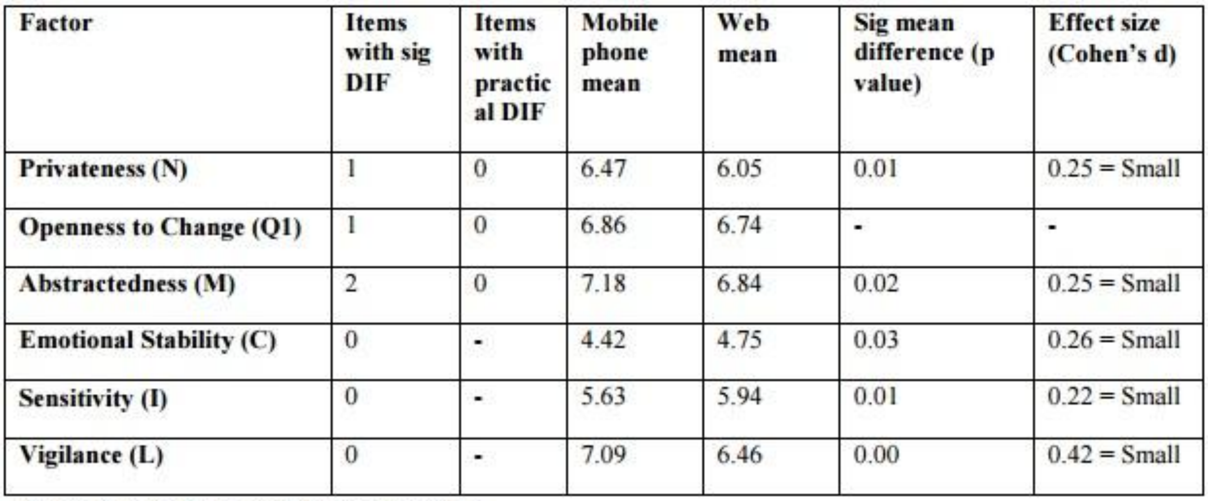
Mobile vs. web study

DIF: Only 4 items were flagged for significant DIF between web and mobile administration; however, when looking at effect size, none show practically meaningful DIF. (See Table 2)

T-tests: There were 5 significant differences; however, effect sizes are small. (See Table 2)

Demographics: there was a clear age difference between web (average age 30 years) and mobile phone users (average age 25 years).

Completion rates differed substantially across the two platforms: 8.7% on mobile phones and 55.2% on the web.



Factor	Items with sig DIF	Items with practical DIF	Mobile phone mean	Web mean	Sig mean difference (p value)	Effect size (Cohen's d)
Privateness (N)	1	0	6.47	6.05	0.01	0.25 = Small
Openness to Change (Q1)	1	0	6.86	6.74	-	-
Abstractedness (M)	2	0	7.18	6.84	0.02	0.25 = Small
Emotional Stability (C)	0	-	4.42	4.75	0.03	0.26 = Small
Sensitivity (I)	0	-	5.63	5.94	0.01	0.22 = Small
Vigilance (L)	0	-	7.09	6.46	0.00	0.42 = Small

Table 2: mobile vs. web administrations

Other Comparisons

Gender and ethnicity showed comparable levels of difference to the languages (gender: 33% significant DIF items, 0.36 sten score difference on average; ethnicity: 23% significant DIF items, 0.33 sten score difference on average). Occupational differences showed the largest mean differences between groups. For example, GP and Neurosurgery trainees showed an average sten score difference of 0.51, applicants vs. general population group (gender balanced) showed a difference of 0.61 stens, police and IT (all male) showed 0.81 average difference in sten scores.

Discussion

All hypotheses were supported. The results for the mobile phone vs. web administrations were clear: an absence of DIF indicates that any differences found in the means of each group were due to the sample, not bias in the measurement properties of the media. As for the languages, the results were more mixed. They suggest that there is DIF present. Content analysis of the questions suggested translation and cultural issues had a part to play in the DIF (e.g. in one item of the Swedish translation an archaic word was used, and a Swedish cultural norm to follow rules caused higher scores on Rule Consciousness items).

> *It is clear that language differences are no greater than gender or ethnicity differences, and are smaller than differences between occupational groups*

However, effect size estimates suggest that DIF has little practical effect at the scale level, as observed both in lordif results and in the t-test results. When compared with demographic and occupational data, it is clear that language differences are no greater than gender or ethnicity differences, and are smaller than differences between occupational groups.

Limitations

The data came from a recruitment sample – with a clear social desirability signature in the trait scores. This is both a limitation (as the results cannot be said to generalise to a general population sample), but also a benefit, as the application of this work will be the assessment of individuals in international recruitment campaigns. Additionally, the analysis was limited to an EU/US sample; however, preliminary analysis suggests that Traditional Chinese, English for South African and English for India versions show less variation than some of the European languages. This analysis will be complete within 2014. Conclusions

Conclusions

Knowing a person’s job is likely to tell you more about their personality than knowing their nationality, gender or ethnicity. Asking people to complete personality assessments on mobile phones or the web appears to make little difference to the measurement, but mobile phone completion rates are lower and this format currently appeals more to a younger age group than web administration.

References

Bartram, D. (2008) Global Norms: Towards Some Guidelines for Aggregating Personality Norms Across Countries. International Journal of Testing, 8: 315–333, 2008

Illingworth, A.J., Morelli, N.A., Scott, J.C., Boyd, S.L. (2014) Internet-Based, Unproctored Assessments on Mobile and Non-Mobile Devices: Usage, Measurement Equivalence, and Outcomes. Journal of Business and Psychology

OPP specialises in personality assessment, enabling people and organisations around the world to increase their effectiveness through the innovative application of psychological tools and techniques.

The company's market-leading psychometric tools include the Myers-Briggs Type Indicator® and the 16PF® assessment, which provide high-impact results for recruitment and personal development, such as teambuilding, leadership development, communication and conflict resolution.

With more than 20 years' experience providing consultancy services and training programmes, our assessment tools and resources have transformed the businesses of thousands of organisations globally, including the majority of the FTSE 100.

www.opp.com

Oxford, UK – Chicago, USA – Paris, France – Amsterdam, The Netherlands