The anglicization of American personality tests: A comment on the debate

Paul Kline

I HAVE BEEN asked to write a summarizing commentary on the debate concerning the anglicization of personality questionnaires, as an academic with perhaps a smaller axe to grind than those concerned in the commercial development of questionnaires.

One of the problems with a debate of this sort is that it involves a number of truisms with which no rational person could disagree:

1. Since cultures differ, tests developed in one culture may not be satisfactory in another.
2. It is wrong to use tests which may not work efficiently for selection or any other purpose.

From this it is clear that the argument is essentially about methods, since it follows that either tests developed in one culture must be shown to work in alien cultures before use, or separate tests for each culture should be used, thus involving problems of comparison. If the first option is chosen, as it appears to have been in this debate, then technical issues are raised of how cultural equivalence can be illustrated.

Although there is no reference to it in this debate this is one of the oldest issues in cross-cultural psychology. Perhaps the reason that it is not more widely known is that it was given a ridiculous label (not rare in psychology of any sort) - the emic-etic dilemma. Some cross-cultural psychologists argued that comparison across cultures was impossible. Members of a culture could only be assessed within the context of that culture. Etic researchers believed that were common dimensions - human universals, envy, greed and aggression come easily to mind. I raise this issue because it is only an assumption that personality traits can be equated even in cultures as similar as America and the UK.

Berry & Dasen (1974) have argued that before cross-cultural comparisons can be made three criteria have to be met - functional equivalence, conceptual equivalence and metric equivalence.

Functional equivalence

Cleanliness is a good example. An item concerned with cleaning under stair rods might work in one culture but even in urban Scotland with a preponderance of bungalows there would be a difficulty. A different but equivalent item would be required.

Conceptual Equivalence

This concerns the actual phraseology or translation of the items. The examples in the debate of baseball and wrestling illustrate this difficulty.

Metric equivalence

This involves the demonstration that items are equivalent across cultures.

Before I discuss how all these aspects of equivalence may be met , it is worth noting that cross-cultural psychologists generally argue that the ideal solution in comparing cultures is to use emic tests (tests specially developed for each culture). This, however, raises a further problem of the meaning of any such comparisons.

I shall deal only briefly with conceptual and functional equivalence because although essential, these are subjective and the empirical test of whether such equivalence has been met lies in the demonstration of metric equivalence. Items may fail in metric equivalence because of defects in conceptual and functional equivalence. These latter are the guidelines for adapting items to a new culture. In fact, as is clear from the

examples of items in the debate, minor changes to the wording are attempts to produce conceptual equivalence while the attempt to yield functional equivalence results in items with no obvious similarity.

Thus the central issue of this debate concerns the demonstration of metric equivalence and cross-cultural validity. This latter is essential since metric equivalence is necessary but not sufficient for validity.

Different ways of demonstrating metric equivalence

Classical item analysis

This involves the demonstration, in each culture that an item correlates with the total score, similarly in both cultures. In addition the proportion putting the keyed response in both cultures should be the same. Items are only selected for the scales if they meet these criteria. Without splitting hairs about what is meant by similar correlations and proportions, this method has an obvious difficulty. It assumes that there are no cultural differences on the variables. If Americans, for example, were markedly more extraverted than British people, to select items with identical item analytic indices in these cultures would be bad measurement. If we do not make this assumption, then any differences in the proportions putting the keyed responses to items in the two groups may be a reflection of an inadequate item. From this it is clear that the use of classical item analysis in the demonstration of cross-cultural equivalence makes assumptions either about the similarity or lack of it which need to be justified.

For practical purposes of test construction and use, it is sensible to include items in scales which correlate highly with the total scale score in both groups even if there are differences in the proportions putting the keyed response. Of course, large samples are necessary to reduce statistical error. Item differences will be ironed out in the British norms.

Factor analysis

Here the technique is to factor the items and to select items which have similar factor loadings in both cultures. Again from the statistical nature of factor analysis and the inevitable error in factor loadings what is regarded as a similar structure is somewhat subjective. This is an effective method especially since cultural differences between items are often reflected in the fact that poor items in the new culture may load on more than one factor. Again item differences between the cultures undetected by this method can be ironed out in the establishment of new British norms.

Some psychometrists would recommend that confirmatory factor analysis be used to demonstrate the equivalence of a scale in another culture. This is a possible technique given that large samples (more than 500) are used to reduce the error in the maximum likelihood analyses and also given that the factor loadings which the analysis is to confirm are neither too rigorous or too lenient. For example if we were to insert the American loadings as the target for the analysis in the British sample, we would almost certainly fail. If all that was required was that some items load zero, others greater than .3, it would be easy to confirm the metric equivalence. (See Kline, 1993 for further details and pitfalls in confirmatory analysis.)

Even if by item analysis or factor analysis scales have been shown to be metrically equivalent it is still necessary to demonstrate that these scales are valid in the new context. As Cattell (1978) has argued this is best done by showing that the correlations between the scales and external criteria are the same in the two cultures and that the same groups are discriminated by them. If this is done then it makes sense to argue that the scales are genuinely equivalent.

Standardization

As has been argued the item analyses and factor analyses, even if equivalent, still allow mean differences between the American and British groups. However since we have no way of knowing whether these are real or due to item imperfections, standardization in the new culture is essential. This will mean that the standard scores are then comparable across cultures, given that the scales have been demonstrated to be valid, as discussed

above. Indeed, for applied use scales with item analytic or factor analytic equivalence, demonstrated validity in the cultures and adequate norms within the cultures are certainly useful and usable.

However, before concluding I want to make two further points. There was reference, in this debate to science and scientific rigour. In my experience (as with BSE) an appeal to science is often a defence, a defence which hides ignorance and flawed reasoning. To an extent this is true of this debate.

The first concerns the use of item characteristic curve theory. Exponents of this approach regard it as having displaced the classical methods (Hambleton et al., 1991). Rasch scales are said to be population free and item free, i.e. Rasch analysis computes indices of item difficulty which are independent of the sample completing the test and indices of participants' status on the variable which are independent of the items in the scale. Thus Rasch scaling would appear to be well suited to the cross-cultural application of tests. Items which differed within the two cultures would be eliminated. However, there are problems with this approach, which are too technical for discussion here but see Kline (1998) although it might prove useful in the field of aptitude and abilities. It certainly merits further research.

The second point is the more important. Occupational psychologists claim that psychometric measurement is scientific and regard it as the application of science to selection. However, as Michell (1997) demonstrated, there is no little delusion in this claim. Psychometric tests differ from scientific measures in crucial ways: there are no true zeros, only the assumption of equal intervals and no units of measurement. Item characteristic curve theories have established better scaling properties for tests but these still have no units of measurement. Without these meaningful measurement is highly problematic. Recently, I have attempted to outline how psychometrics should proceed in order to approach the precision of the natural sciences but that task lies ahead (Kline, 1998).

Conclusions

Until psychologists possess tests with units of measurement, the problems of validity and test equivalence, not to say standardization will remain. Notice that a ruler, whether in metres or yards, is cross-culturally valid, and requires no standardization or evidence of validity. Until that happy day, we will have to rely on item and factor analyses followed by validation and standardization of the scales in each culture. If this is well done then selection and appraisal will be on a sound basis but a basis that is essentially pragmatic and empirical rather than scientific.

References

Berry, J.W. & Dasen, P.R. (1974) **Culture and Cognition: Readings in Cross-Cultural Psychology.** London: Methuen

Cattell, R.B. (1978) **The Scientific Use of Factor Analysis.** New York: Plenum Press.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991) **Fundamentals of Item Response Theory.** Newbury Park, CA: Sage

Kline, P. (1993) **The Handbook of Psychological Testing.** London: Routledge

Kline, P. (In Press) **The New Psychometrics: Science, Psychology and Measurement.** London: Routledge

Michell, J. (1997) Quantitative science and the definition of measurement in psychology. **British Journal of Psychology, 88,** 355-383

Dr Paul Kline *is Professor of Psychometrics at the University of Exeter*