The insignificance of personality testing

Steve Blinkhorn and Charles Johnson

The business world now sets great store by personality tests when assessing job applicants. But the evidence for their predictive value is frequently overstated and wrongly assessed.

PERSONALITY testing has never been uncontroversial in psychology. Many academic psychologists would be surprised to find that in the business world personality testing in recruitment and selection is considered to be the pith and essence of psychology. But in recent years there has been a dramatic growth in the use of personality tests for these and related purposes. About 50 per cent of companies in the United Kingdom claim to make use of personality tests at some point in their selection or assessment processes. Some of these tests have claims to origins in psychological theory; others make technical, methodological or empirical claims; still others rejoice in being unencumbered by any such distracting irrelevances. But even where the proposed basis for their use is demonstrated empirical predictive utility, we have severe doubts as to the accuracy of the claims.

The normal basis for a discussion of the use of psychological tests is statistical findings, usually expressed as correlation coefficients. But many proponents of personality testing adopt an approach to correlation that would have left its inventor Carl Pearson gasping and which beggars the contribution of R. A. Fisher to significance testing. But few of the punters have the kind of background in statistics that enable them to evaluate claims, and the numbers can be made to look truly impressive.

Caricature

Personality tests are scarcely new. Their origins in the first half of this century can be caricatured as a response to the apparent success of intelligence testing. "We have the technology" came the cry, and elder statesmen, now revered, turned their skills, their factor analytic routines and their research grants, to investigating the structure of human personality through the medium of questionnaires. It is not all that they did by any means, but it is the part which has survived to fascinate and intrigue. And although the popular imagination may focus on Rorschach ink-blots and the like as the type for personality tests, the big money is with multiple-choice questionnaires.

Early systems for describing the structure of human personality showed a certain enthusiasm for scientistic mystification. How do you rate on Rhathymia, Hypochondriasis and Affectia? More recently there has been a fashion for gritty everyday language. Submissiveness, Independence and Data Rationality. But regardless of terminology, the evaluation of the validity of a test has been conducted in terms of correlation coefficients.

For many years we took the view that the relatively poor showing of personality tests as compared with tests of mental ability was down to the need for extensive technical improvement. Ability tests show a rather monotonous tendency to correlate around 0.30 with various job performance measures. Personality tests do distinctly worse. Reasonable suggestions were made that the problem lay in: details of construction leading to poorer psychometric properties; the clinical orientation of many of the items in tests; the scarcity of well-conducted validity studies; the use of discredited theoretical approaches to the structure of personality.

All of these observations may well have been accurate, but all have now been addressed, and word is abroad that the current generation shows distinctly better validity than what went before. We beg to differ. We have conducted an informal survey of research on the predictive validity of three of the personality tests most widely used for employee selection and assessment which have claims to a serious and respectable pedigree, and we find a disconcerning approach to the analysis and interpretation of statistical data.

It has to be said that in restricting ourselves to these three (the California Psychological Inventory, the 16 Personality Factor Questionnaire and the Occupational Personality Questionnaire) we were consciously choosing to operate at the top end of the market. We maintain a 'Black Museum' of tests guaranteed to terrify the methodologically squeamish. Most of these, because of the way in which they are constructed, yield scores which are mathematically interdependent, and so unsuitable for the usual forms of statistical analysis. But the three we chose to look at are serious measures, constructed by sober-minded teams aiming for quality rather than just a quick buck.

The general style of these tests is as follows. Some activity (for example, going to parties) or choice (for example, parties versus reading a book) is briefly described, and the candidate is asked to choose from a limited number of options which describes him or her best. Responses to perhaps 10 or 15 such items are scored according to predetermined rules to make up one of the 'scale scores'. These scores can be presented in a graph (or 'profiled') or, increasingly, used to drive software to produce written reports.

The first thing that would strike a statistically competent but psychometrically unblooded observer is the sheer number of scores assigned on the basis of each of these tests — between 16 and 30. Linear algebra may make a great deal of sense as a way of capturing a handful of dimensions, but what precedents do we have for conceiving of a 30-dimensional space?

Big Bang

When, as is often the case, scores are expressed on 10-point scales, there is a simple calculation which dramatizes the point. Imagine a computer system printing individual reports based on personality profiles. It drives an 8-page-per-minute laser printer, and on average each report consists of eight pages. Being an exceptionally reliable piece of kit, this system needs no maintenance. We set out to print reports based on all possible profiles for 16 scales, and wisely do so at the time of the Big Bang. In round figures, as we enter the last decade of the 20th century, there are only 500,000,000,000,000 of the 10^{16} possible profiles left to print. The system in the next room, which is working on a 30-scale test, is barely into its first lap.

In truth it is difficult to know how one would justify the underlying mathematical model. We doubt that, even just splitting each scale at the mid-point to yield 2^{16} possible profiles, any existing database actually contains a representative of each. And in typical validation studies, we find samples of between 30 and 150 subjects.

But there is more. Only rarely are investigators satisfied with a single-criterion measure against which to validate their preferred test: five or six are not uncommon. The enterprise then becomes a fishing expedition. Take 30 test scores and half a dozen criterion measures on 50 or 100 subjects; calculate product-moment correlations between all scores and all criteria. Then look for your trusty table of critical values of the correlation coefficient and note and significance level of each. Report only those which are 'significant' at the 0.05 (or one-star), or 0.01 (two-star) or better level. Hey presto! Your test is now valid.

Do not on any account point out that you have calculated more correlations than there were subjects in your sample.

Never hypothesize in advance the magnitude or direction of the correlation, for that way your hypotheses may be shown to be wrong rather than merely not proven. Avoid mentioning how many 'nonsignificant' correlations were found.

We think that this kind of misuse of a hypothesis test is scandalous, and bamboozles an unsophisticated public with pseudoscience. The distribution of correlation coefficients in multivariate designs is a rather intractable problem, but that is no excuse for pretending that it can be ignored, and that the repeated application of a test designed for a single correlation can simply be prolonged until a 'significant' result is obtained.

To get an empirical perspective on the issues we turned to published sources, and started counting stars. For instance, the publishers of the Occupational Personality Questionnaire have published a review of validity studies, of which 28 were germane to this issue. Variations in the way results are reported make it difficult to be precise, but on average in any given study about 6.5 per cent of correlations are marked as significant at the 5 per cent level or better. In other words, there appears on the surface to be some slight effect beyond what one would expect in a random system. But both test scales and criterion measures tend to be heavily correlated amongst themselves, which will tend to increase the incidence of significant correlations.

Elsewhere we have found 11 published studies concerning the California Psychological Inventory (CPI) and four concerning the 16 Personality Factor Questionnaire (16PF) which address the use of the tests in a business rather than an educational or clinical setting, and where adequate data are presented. For the CPI, 9.9 per cent of correlations (that is, for two of the 20 scales) were 'significant', and for the 16PF, 7.3 per cent (that is for one of the 16 scales).

These results suppose that the most appropriate null hypothesis for a list of many correlations is a run of zeroes, but of course it is not. As has often been noted, the distribution of sample product-moment correlations is complex, and depends on both sample size and the underlying correlation in the population. Assuming zero correlations in the population, however, we can derive an expected distribution of (ordered) correlations using the method of rankits. Ordering observed correlations in the same way, we can gain an impression of the extent to which observed correlations depart from what might be expected by chance alone. Table 1 gives an example from a real study of 38 staff supervisors.

TABLE 1 Expected and observed correlations between 30 test scales and overall job performance (sign and decimal point omitted)

Expected	39	33	29	26	24	22	21	19	18	17	16	15	14	13	12
Observed	32	32	32	28	26	23	22	20	18	18	17	16	16	15	15
Expected	11	10	09	08	08	07	06	05	05	04	03	02	02	01	00
Observed	13	13	13	11	10	09	09	08	07	07	06	03	03	02	02

The first three observed correlations are 'significant' at the 0.05 level, but clearly the two distributions are as close as makes no difference. Yet on the basis of such data, claims are made for the predictive validity of tests.

The basic concern of those who would use personality tests is usually whether one can have confidence that predictive relationships between test scores and criteria would recur in a cross-validation. In other words, has a real underlying relationship been identified? Going for the largest correlations looks the safest route. But what does one expect the largest correlations to be? Not zero, even on a null hypothsis of zero underlying correlations. Assume 30 test scales, a single criterion and a sample of 50. Ignoring sign (because we are talking about fishing trips with no hypothesized direction for relationships), the expected value of the smallest correlation is never zero, and the expected value of the largest is 0.34. The 95th percentile point of the distribution of the largest correlation in a set of 30, again ignoring sign, is about 0.42.

Looking at correlations in order of magnitude introduces a further complication in that under the null hypothesis of zero correlations in the population, expected values are not independent. In fact they are positively correlated, implying that one chance high positive correlation is likely to be accompanied by others. This is independent of any associations attributable to correlations amongst test scales (and is believed by writers of textbooks — for example, H.A. David (*Order Statistics;* Wiley, 1981) — to be intuitively obvious).

To be concrete, there is little evidence of enduring relationships between personality test scores and measures of success at work. For instance, Table 2 shows a typical case of results failing to cross-validate, extracted from the validity review mentioned earlier. The test has 30 scores, of which 21 showed no 'significant' correlations with the extent to which sales targets were met.

So these sales managers did better if they were indecisive in 1985, but this did not carry over into 1986, though it began to be important to lose emotional control as that year wore on, particularly if optimism took hold, critical attitudes prevailed and modesty were out of the window. This is of course nonsense. The truth of the matter is that these correlations, despite their stars, are well within the bounds of what chance might throw up.

There are more sophisticated forms of validity delusions. Readily available statistical software allows easy computation of multiple correlations, which may lead to the claim that as a set the scales predict job performance, in accordance with a linear regression equation. And indeed, large multiple correlations are frequently found. But large numbers of predictors in combination with small sample sizes inevitably yield large multiple correlations. For example, when population correlations are all zero, the expected value of the multiple correlation coefficient between 30 predictors and one criterion on a sample of 50 subjects is 0.77. What psychologist would not crack a bottle of bubbly on the strength of such a result?

We are not suggesting that personality tests have no uses, or that there are no stable underlying aspects of temperament which are important in the determination of behaviour. Indeed, for counselling purposes, or in other situations where self perception is as important as the truth, they may be invaluable. But we see precious little evidence that even the best personality tests predict job performance, and a good deal of evidence of poorly understood statistical methods being pressed into service to buttress shaky claims. If this is so for the most reputable tests in the hands of specialists, one may imagine what travesties are committed further down market. But we leave this as an exercise for the reader. □

TABLE 2 Observed correlations between test scores and sales performance over time

Test score	1985 Sales target	1986 (first half) Sales target	1986 (second half) Sales target
Persuasive	—	0.25*	0.35*
Outgoing	−0.23*	—	—
Affiliative	0.26*	—	—
Modest	—	—	−0.39**
Emotional control	—	—	−0.29*
Optimistic	—	—	0.28*
Critical	—	—	0.29*
Achieving	—	—	0.32*
Decisive	−0.27*	—	—
	(*n*=88)	(*n*=91)	(*n*=56)

Steve Blinkhorn and Charles Johnson are directors of Psychometric Research and Development Limited, Brewmaster House, The Maltings, St Albans, Hertfordshire AL1 3HT, UK.