SDR

Psychometrics and personality questionnaires

***Paul Barrett** and **Rosalie Hutton** discuss psychometric properties of personality questionnaires*

THIS ARTICLE CAME about quite by accident. From what started as a relatively straightforward psychometrics exercise gradually turned into something of much greater import. We were originally interested in whether or not job applicant responses on personality questionnaires are so distorted (in comparison to volunteer data) that the fundamental psychometric properties and/or structure of a questionnaire can be substantively changed. If this is the case, then we might have to consider the validity of scoring a scale of items with a score-key that has only been established using non-applicant/volunteer data.

Schmit and Ryan (1993) examined the NEO Five-Factor Inventory structure in job applicant and volunteer samples. That is, they factor analysed each set of responses from the job applicants and from their volunteer university student sample. They found that the NEO items from the volunteer sample factored into the expected five-factor model, but that the applicant data did not produce the same solution. The conclusion from this pioneering study was that to continue scoring questionnaire data using a score-key generated from volunteer data might no longer be considered a valid activity. They suggested that perhaps alternate score keys be used for both samples respectively.

Brown (1998) and Brown and Barrett (1999) subsequently investigated the differences between job applicant sample and volunteer sample questionnaire responses on the 16PF-5 personality questionnaire. The analyses ranged from comparing the scores on the 16PF-5 scales between applicants and non-applicants, through to factor analysis of the scale scores, in an effort to replicate the published global factors. The data samples comprised a sample of 589 non-applicant (N=403 training course delegates), mixed gender individuals who completed the 16PF-5. The job applicant sample consisted of 506 mixed gender participants, the majority of whom were graduate applicants to a merchant bank. The authors were also given access to the ASE UK volunteer standardisation sample (N=1575) 16PF-5 scale-score correlation matrix. In order to really have the best chance of finding the expected factor solutions, we used the published factor solution (in Table 1.4 of the *US Technical Manual* for the 16PF-5) as a 'target', and then tried to get as near as possible to this target solution with our factor analysis. We also went as far as using the factor correlations from the *US Technical Manual* Appendix 1B to help define the target factor solution for the 16PF-5.

The results of these analyses led us to conclude that a high degree of similarity existed overall in the 16PF-5, given the particular datasets. However, the results for Anxiety and Tough Minded global factors were nevertheless awkward. The scale Q1 shared the largest second order factor weight with scale I on Tough Minded (0.5) in the factor scale equations given on page 16 of the *US Technical Manual*, yet was barely identifiable as a significant variable in the applicant data. Likewise with scale L on the Anxiety second order, although this was not so severely affected. On the basis of these results, and those of Schmit and Ryan (1993), there did seem to be some substantive evidence that applicant factor structures of questionnaires might be different in certain

respects to those of volunteers. It was noted that two scales currently used in the second order factor equations would not have been identified under the rules for identification used in the *US Technical Manual* (factor loadings >0.30), and therefore not weighted at all in the equations. Figure 1 below shows the full impact on the global equations if we had adopted this criterion for composition of the global factor scores. What we see here is that for the global factor Tough Minded, two out of the four scales which should be contributing to the meaning of this scale have 'disappeared' in the applicant sample. Further, three out of the remaining four factors lose a 'global' factor weight. Given that these global factors are defined by four scales, surely the loss of one of these defining scales might be expected to have some effect on the interpretation of the meaning and predictive facility of these global factors?

Given the crucial import of these results to the future use of personality questionnaires in staff selection, we decided to proceed further with similar comprehensive analyses of applicant and non-applicant data samples, using two additional popular questionnaires, Saville and Holdsworth's OPQ normative Concept Model 5.2 Questionnaire and Psytech International's 15FQ questionnaire (details in Barrett & Hutton, 2000). The results from the 15FQ analyses indicated that there were no loading differences between the applicant and non-applicant data. And it is here that our study took a surprising turn.

For the OPQ questionnaire, I (Paul Barrett) reached for the manual to specify the 'target' factor model for the Concept OPQ scales. But then, Helen Baron's (1996, p.22, third paragraph) words came to mind... '*The attempt at confirmatory factor analysis is also misguided. OPQ scales are divided into three broad domains: relationships with people, thinking style, and feelings and emotions. There is no claim that these domains are unidimensional or even that they represent higher order factors. They are merely collections of scales which relate to different aspects of behaviour*'. Essentially, the Concept OPQ consists of 31 scales. End of story. Unlike the 16PF5, NEO, and 15FQ, no empirical structuring of the covariance between the scales is imposed (i.e. the 'factors' in a factor analysis).

Figure 1: Global factor equations for the US normative sample data, and those that would have been produced had we used the same criteria but with the UK applicant data.

The global factor scores

P.14, first paragraph, US Technical Manual, 'global factor equations were developed using only those primary scales having a loading of .30 or greater...'

From Table 1.4, p.16, *US Technical Manual*

Extraversion	=	.3A + .3F + .2H - .3N - .3Q2
Anxiety	=	-.4C + .3L + .4O + .4Q4
Tough Minded	=	-.2A - .5I - .3M - .5Q1
Independence	=	.6E + .3H + .2L + .3Q1
Self-Control	=	-.2F + .4G - .3M + .4Q3

From the UK applicant dataset SEM analysis

Extraversion	=	.3A + .3F + .2H - .3N - .3Q2
Anxiety	=	-.4C + .4O + .4Q4
Tough Minded	=	-.2A - .5I
Independence	=	.6E + .3H + .3Q1
Self-Control	=	.4G - .3M + .4Q3

The three-domain concept model proposed for broad interpretation of clusters of scales is entirely subjective. So, we sat back and wondered just how we might proceed with examining whether there were any differences in the psychometric structure of the OPQ (in applicant and non-applicant data), given that no *a priori* psychometric structure is proposed for the Concept 5.2 questionnaire. Well, in this situation, it is clear that imposition of a 'structure' upon data by solely empirical means may be quite arbitrary. Perhaps a useful exposition of this fact is the paper by Maraun (1997), showing that the use of another kind of structural analysis with NEO questionnaire data reveals a two-dimensional structure, whereas, the use of more conventional factor analysis reveals the five dimensions we know as the five factor model. To say which is the 'more correct' solution is actually a very awkward problem with no simple answer. Further, it is clear that interpretation by a practitioner of the 31 OPQ scales does not rely upon an empirically constrained clustering of scales (as with the 16PF5 global second orders), but on a more qualitative appreciation of how scales might interact in a meaningful way for a particular candidate. Therefore, the only quantitatively mediated concern for a practitioner in these circumstances is that the scales are making reliable measurement of some proposed construct. Everything else is of qualitative import. So, the analysis finally undertaken for the OPQ using applicant and volunteer data was a comparative analysis of internal consistency reliability. Only one scale 'Independent' (R3) showed any difference; an alpha of 0.63 for the UK normative sample and an alpha of 0.45 for our applicant sample.

As we pondered these results, and especially those from our initial 1999 study using the 16PF-5, we seemed to have arrived at a position where we had shown that the 16PF5 global factor equations may be incorrectly specified for applicant data, but in practice, this result has no apparent detrimental effect. Further, from Helen Baron's statement concerning the OPQ, it was clear that qualitative interpretation of the scales and their relationships was paramount. In short, for the 16PF-5 we had demonstrated a substantive decrement in measurement breadth for certain global factors - which has no apparent practical effect. For the OPQ, we are led to conclude that we could never empirically demonstrate a loss of measurement 'breadth', as there is no empirically defined model of 'breadth' for this test. However, it gets worse, much worse. Remember the normative (Saville and Blinkhorn, 1981) 16PF Form A reliabilities and test-retest coefficients - especially those for scales A-outgoing (0.37), M-Conceptual (0.21), N-Restrained (0.27), and Radical (0.39). Classical psychometric test theory would tell us that these scales are making very poor measurement. Yet, the test sold well, and probably hundreds of thousands of candidates world-wide were assessed. It apparently made no difference in practice whether the scale alphas were 0.8 or 0.2! As Blinkhorn has pointed out (private communication), a high-profile consultancy used factor N on the 16PF as a key interpretative variable - yet we can see that its reliability is so low as to be almost non-existent. Exactly how are practitioners able to continue successfully using tests that seem to defy the principles of measurement upon which the tests have been constructed?

Consider for a moment what happens when you want to measure the accuracy of machining of nuts and bolts in a factory. Several devices will be available that range from a simple ruler, a type of vernier gauge, through to laser interferometry. Each of these kinds of devices can be made to differ in 'look and feel', and will differ drastically in price. They can also differ substantively in accuracy. If you are the owner of the factory making these products, what is your first priority, 'look and feel', price, or accuracy? You know that if the dimensions and threads cut on the nuts and bolts do not match within certain tolerances, then you are producing expensive scrap metal. You also know that even if the threads are cut to within tolerance, the actual physical dimensions of the products may be inaccurate, meaning that they cannot be used for the purpose for which they were originally ordered. So, the first priority is accuracy of measurement. Then comes price, and finally 'look and feel'. Here, the outcome of 'getting it wrong' is

clear and substantial. Now think of how you buy a personality questionnaire for selection purposes. Is the Myers Briggs 'more accurate' than an OPQ 5.2? Does the 16PF5 Extraversion scale make more accurate measurement than the Eysencks' EPQR Extraversion scale? Do some use Hogan Assessment Systems' 'Dark Side' questionnaire because it measures the dark side of personality much more accurately than the Eysenck Personality Profiler? As we ask these questions, it is clear how inappropriate they are for the area. There is actually no way of answering such questions as accuracy in this area is not verifiable in the manner of a scientific measuring instrument. Instead, our best bet seems to be to rephrase the questions in terms of 'which scale best predicts specific criterion x'. The papers by McHenry (1997), and Hogan, Hogan and Trickey (1999) seem to follow this line of reasoning.

Many are now familiar with the arguments by Kline (1998, 2000) and Barrett (1998) concerning the measurement problems of psychometrics - and some have rejected these arguments within the pages of this publication (Duncan, 1999). However, we contend that the dilemmas highlighted within this paper are the direct result of confusing attributes that we normally apply to physical measurement, with what takes place in psychometrics. Surprisingly for some, we agree with the position of the Hogans, McHenry, Duncan and Trickey. When it is impossible to make decisions upon the basis of the accuracy of measurement of measuring instruments, the next sensible approach is to differentiate tests on their ability to predict certain outcomes. You no longer have to be concerned with what precisely it is you are measuring, but merely that whatever it is, it predicts a valuable outcome. Of course, face validity generally allows us to make some coherent semantic generalisations - even though some discrepancies do occur from time to time (such as is currently occurring with the evolving construct of Emotional Intelligence).

So far so good - but is it? The predictions we are talking about are quantitative - generally constructed using regression and correlation analysis, against specific job criteria. There is no qualitative or narrative 'interpretation' of test scores taking place, merely linear numerical operations. It seems that to continue with the use of this kind of 'prediction' or 'criterion' keying paradigm would require that practitioners should work as actuarial analysts. Forever locked into using decision analysis and decision-theoretic statistics to assist optimal candidate selection, for that is where the validity for test use is now coming from. However, most practitioners (as evidenced by the recent 1999 Occupational Psychology Conference debate, and papers by Ridgeway (1998) and Maddocks (1998)) use, or desire to use personality questionnaires in order to assist in the more 'clinical' or 'therapist' interpretation of candidate attributes. This renders quantitative outcome evaluation of the use of a personality test in a total candidate decision-making process as completely impossible under current conditions of usage (and with current practitioner training). Hence, it now comes as no surprise that attempting to analyse personality questionnaire data as though it consisted of quantitatively measured variables, is actually of no relevance to real-world test practice (except for marketing, and other qualitative exercises). It is no wonder that when numerical index 'discrepancies' occur in analyses such as our own, there is no apparent impact anywhere except amongst a few esoteric psychometricians in the UK who might sometimes read this kind of work.

So, maybe we just consign these results to the wastebin. Or maybe we ask just why the BPS imposes Level A and some Level B training in quantitative psychometrics. It seems to matter little in reality as to whether personality tests have high or low alpha scales, use factor analysis models (or not), and use scores that are interpreted 'clinically-qualitatively'. Frankly, why bother with any practitioner training in personality test use at all, except that required to develop expertise in narrative feedback and qualitative assessment?

Now, in conclusion, let's be very clear about what we are saying. That personality test scores can be shown to predict certain outcomes is not in question - see Trickey and Hogan (1998), Ackerman, Kanfer and Goff (1995), and Schmidt and Hunter (1998). However, what we do question

is whether the 'objective' psychometrics and test theory parameters that are used to substantiate personality test scores is dwarfed by the degree of 'subjectivity' in current I/O usage. Dwarfed to such a degree that even measurement reliability of test scores is no longer of any serious practical concern to those who use tests on a day-to-day basis. Those who respond *'but this is crazy, how can we make better measurement of psychological attributes if it is unreliable?'* must first decide how they would ever recognise 'better measurement' in this field. There are solutions to this problem, but we have yet to be convinced that many others in this area even appreciate the magnitude of the issues accidentally exposed above.

References

Ackerman, P. L., Kanfer, R. & Goff, M. (1995). Cognitive and non-cognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied, 1,* 270-304.

Barrett, P.T. (1998). Science, fundamental measurement, and psychometrics. *Selection & Development Review, 14,* 4, 3-10.

Barrett, P.T. & Hutton, R. (2000). *The distortion of meaning and measurement in applicant sample personality questionnaire responses.* Paper presented at the 2000 BPS Occupational Psychology Conference, Brighton (available for download from http: //www.liv.ac.uk/~pbarrett/present.htm)

Baron, H. (1996). An evaluation of some psychometric parameters: A response to Barrett, Kline, Paltiel, and Eysenck. *Journal of Occupational & Organizational Psychology, 69,* 1, 21-23.

Brown, R.(1998). *A study investigating the differences between applicants and non-applicants on the 16 personality factor questionnaire - 5th edition.* MSc thesis, University of Sheffield.

Brown, R. & Barrett, P.T. (1999). *Differences between applicant and non-applicant personality questionnaire data: Some implications for the creation and use of norm tables.* Paper presented at the 1999 BPS Test User Conference, Scarborough.

Duncan, D. (1999). The future of psychometric testing: A user's view. *Selection & Development Review, 15,* 1, 16-17.

Hogan, R., Hogan, J. & Trickey, G. (1999). Goodbye Mumbo-Jumbo: The transcendental beauty of a validity coefficient. *Selection & Development Review, 15,* 4, 3-9.

Kline. P. (1998). *The New Psychometrics: Science, Psychology, and Measurement.* London: Routledge.

Kline, P. (2000). *Handbook of Psychological Testing* (2nd ed). London: Routledge.

Maraun, M.D. (1997). Appearance and reality: Is the big five the structure of trait descriptors?. *Personality and Individual Differences, 22,* 5, 629-647.

McHenry, R. (1997). *Quality standards for developing tests: An alternative to existing orthodoxy.* BPS Test User Conference, Stratford.

Maddocks, J. (1998). The significance of the personality test user. *Selection & Development Review, 14,* 5, 10-11.

Ridgeway, C. (1998.) The Assessor-Therapist. *Selection & Development Review, 14,* 16-17.

Saville, P. & Blinkhorn, S. (1981). Reliability, homogeneity and the construct validity of Cattell's 16PF. *Personality and Individual Differences, 2,* 325-333.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 2, 262-274.

Schmit, M. J. & Ryan, A. M. (1993). The big five in personnel selection: Factor structure in applicant and non-applicant populations. *Journal of Applied Psychology, 78,* 6, 966-974.

Trickey, G. & Hogan, R. (1998). We don't have a choice - personality matters. *Selection & Development Review, 14,* 6, 12-13.

Paul Barrett, *The State Hospital, Carstairs and University of Liverpool, and* ***Rosalie Hutton,*** *Psychometric Technology Ltd., Kings Bromley, Staffordshire.*