Journal of Occupational and Organizational Psychology (1996), 69, 1–19 *Printed in Great Britain*
© 1996 The British Psychological Society

An evaluation of the psychometric properties of the concept 5.2 Occupational Personality Questionnaire

P. Barrett*

Department of Psychology, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

P. Kline

Department of Psychology, University of Exeter, Exeter, Devon EX4 4QG, UK

L. Paltiel

Psytech International Ltd, Icknield House, Eastcheap, Letchworth, Herts SG6 3DA, UK

H. J. Eysenck

Department of Psychology, Institute of Psychiatry, De Crespigny Park,
Denmark Hill, London SE5 8AF, UK

Using three samples of applicant data, encompassing over 2300 participants, the Concept Model 5.2 Occupational Personality Questionnaire (OPQ) was examined for scale discriminability at the item, scale and factorial level. Item analysis and maximum likelihood factor analysis indicated that the OPQ questionnaire provided good, low complexity measurement on 22 out of the 31 scales. Nine exhibited poor signal-to-noise ratios, high item complexity indices, and insufficient number of keyed loadings on the appropriate factor. On the basis of the results below and from those reported by Matthews & Stanton (1994), it was argued that the test requires further development in conjunction with the revision of the psychological and measurement models specified as underlying its construction.

Introduction

The Concept 5.2 Questionnaire (Saville, Holdsworth, Nyfield, Cramp & Mabey, 1993) is one of a series of questionnaires that are subsumed under the general product title of 'Occupational Personality Questionnaire' (OPQ). The OPQ was developed from a model of personality that was initially generated from a review of existing questionnaires and personality theories, some work-related information and feedback from organizations, and from some repertory grid data generated by company employees. Using this model as a basis for test construction, Saville *et al.* created several hundred trial items that were tested within various companies and organizations in the UK. From the various analyses

*Requests for reprints.

implemented on these items, 31 scales were retained that provided the operational definition of the OPQ model of personality. The Concept 5.2 OPQ is the normative response questionnaire that is described as the one that most comprehensively measures the OPQ model of personality. From these scales, a variety of questionnaires were also introduced, some ipsative, some normative, some based upon more 'conceptual' and work-oriented validity, others on factor-analytic methodology. Addressing this latter point, it is noted that within the manuals for the test series, the OPQ concept and factor model questionnaires are described as having been derived using different techniques of test construction. However, there seems to be some confusion within the OPQ manuals themselves and within Peter Saville himself over this issue. Although Saville & Sik (1995) repeat the assertions that the concept model was *deductively* derived (subjective, rational, or theoretical derivation), and the factor model *inductively* derived (mathematical analysis of covariance between items, as well as theoretical derivation), it would appear that the same methods of analysis as used for inductive analysis were used to analyse the 'deductive' questionnaire. The only 'deduction' taking place in the development of the items and scales was that implemented in order to generate items hypothesized to measure a collection of psychological behaviours. Exactly the same as that required to generate data for inductive analysis. Barrett & Paltiel (submitted) make this point in more detail.

With regard to the logic of scale construction/item selection as outlined in Section 2 of the test manual (Saville *et al.*, 1993), *The Development of the OPQ*, paragraph 10, page 7 of this section states:

> A good item was taken as one which was closely related (i.e. had a high correlation) with other items in its own scale, but was not closely related to items in other scales. A good scale was one which was internally consistent and which was internally consistent across the four different item formats.

Factor analyses of items were carried out on three sets of data provided by subjective parcelling of items into small clusters consisting of three items. Two of the datasets were ipsative in nature, requiring preference choices to be made between items. No item factor analysis was undertaken on the various datasets. The description of the factor analyses indicated that factor solutions between two and 19 factors were generated, using Promax oblique rotation to simple structure in each case. From these solutions, factorial models with four, five, eight, 11, and 19 factors were chosen. No objective criteria were provided for selection of these numbers of factors. No higher order factor analyses were reported that might have suggested such a hierarchical set of solutions. A 'conceptual' model was used as the criterion for the selection of the various factor structures. Finally, after cross-validating these factors against previous datasets that contained the items used, the final factor models were chosen that contained four, five, eight, 10, and 17 factor scales. Four non-quantitative criteria were quoted as the 'filters' through which this final set of factor models were chosen.

There are two published studies on the 30 OPQ concept model scales. The first examined the scale factor structure of the test on a sample of 94 undergraduates (Matthews, Stanton, Graham, & Brimelow, 1990). Tests of factor extraction quantity indicated five factors to be extracted. The four, eight, 10, and 17 factor models were not replicated, neither was the 14 factor factorial model. Although the number of participants was low in this study, Barrett & Kline (1981) have previously demonstrated that this quantity, although borderline, is sufficient to permit some degree of confidence in the overall analysis results and extraction procedures. In a more recent paper, Matthews & Stanton (1994) carried out both an item level and scale factor analysis of the Concept 5.2 OPQ, using the bulk of the standardization sample of the test (2000 participants). The critical results reported in this study were that some of the concept model scales could not be distinguished clearly and that only a five or six factor solution appeared with any clarity. A 21 factor solution was produced but seven of these 'factors' had only six or fewer items loading greater than .3. In all, 175 from the 248 items in the test were retained as part of the 21 factor solution. In addition, factor similarity analysis (computed by splitting the 2000 participant sample into two groups of 1000 and factoring them separately) yielded many congruence coefficients below .80 (nine out of 21) when comparing factor patterns but only five when comparing factor structures. The mean factor pattern congruence was .79, and for the factor structure, .86. This implies that considerable covariance exists across items and factors, this covariance being partialled out within the factor pattern loadings but remaining within the factor structure loadings.

These results suggest that some of the concept model scales are confounded with items that share variance (and likely semantic interpretation) across scales. In addition, it appears that 73 of the items do not load on any factors within a 21 factor solution as derived by Matthews & Stanton. This is almost one-third of all the items in the Concept 5.2 scales. Regardless of whether one considers a scale as a factor or discrete measurement quantity, there is something very wrong with a test that claims to measure 31 separately named and semantically distinguishable concepts but can only be objectively shown to distinguish perhaps 21 discrete, mathematically distinct entities. Essentially, there appears to be a fundamental discrepancy between what is being subjectively modelled and what is actually being demonstrated by the data, using factor analysis to mathematically distinguish dimensions or facets of personality. Note further that the purported factor structures of the Octagon, Pentagon, and Factor Model OPQ tests are not supported by the empirical results reported within the two studies carried out by Matthews *et al.*

Since no detailed item-level analyses have ever been reported by the test constructors, the psychometric measurement properties of the tests are unknown except insofar as the internal consistency of most of the scales is high, especially for short six- or eight-item scales, and that the test–retest coefficients are also high. These two statistics suggest that the scales are measuring behaviour in a consistent and repeatable fashion. What is not known is just how much overlap in measurement exists between the normative measurement scales of the OPQ 5.2.

Given a test measurement model and corresponding psychological model that assumes discriminability between the behaviours/traits (as within the familiar *domain sampling* trait model founded upon classical test theory), then significant item-level overlap might be considered indicative of poor test development and/or a poor psychological model. Why is this? Well, within a domain sampling model, it is required that items measure a piece of behaviour that is essentially unidimensional and homogeneous. That is, the behaviour to be measured is not a function of more than one causal or latent factor. If it is, then interpretation of the item or scale score is now more complex, as any score on the item or scale is now a function *not* of the assumed unidimensional latent trait underlying the measure, but of two or more latent traits. It is acceptable for the dimensions/domains to be correlated, but it is not acceptable for items within a domain to be also a direct measure of another domain. We have, in fact, strayed from a fundamental tenet of classical test theory. However, let us assume a questionnaire where items like this are not rejected from some test scales—so we now have scales which correlate at about .3 and above, which contain some items that correlate with their own scale scores and with others. This is really now a matter of theory—if my model proposes correlated dimensions, then I have to accept that some item complexity will probably be apparent. However, I also have to wonder whether the dimensional correlation should be so high—or whether it is my items (or some subset) that are introducing the correlation because they are not sufficiently distinct in meaning. Further, I need to consider whether the items are actually composing a dimension or are better thought of as a more specific, meaningful cluster that simply measure a single piece of behaviour. It is simply prudent and efficient psychometric analysis to seek to minimize item overlap in order to both clarify and differentiate the meaning of each scale composed of such items, and to determine whether what were thought to be general scales of behaviour might be better viewed as item 'parcels', measuring a single, specific behaviour. This in turn forces a re-evaluation of the model of personality that should be guiding scale development.

If I change my measurement model from a domain sampling one to say a circumplex one (such as a circumplex model of personality put forward variously by Wiggins (1982), Peabody & Goldberg (1989), and Hofstee, De Raad & Goldberg (1992)) which makes few constraints upon the amount of overlap between traits, then item-level complexity becomes a function of the spatial separation distance in the circumplex model. That is, the closer the spatial proximity of two traits within the circumplex model space, the more overlap might be expected between items measuring the two concepts. However, we have now left the domain sampling model far behind. Personal and occupational profiling with this form of model is fundamentally different to that currently being used by domain sampling trait models (i.e. spatial mapping vs. conventional linear profiling). However, the OPQ, according to the personality model underlying the construction of the test, the measurement model used, and the recommended practical use and interpretation of test results, is a domain sampling test.

The primary aim of this study is to examine the psychometric properties and discrete measurement capability of the OPQ Concept 5.2 Questionnaire, i.e. to what *quantifiable* extent can the concept scales of the OPQ be said to be measuring behavioural traits that are uncontaminated to some *specified* degree with other behavioural trait measures. A related aim is an attempt to identify 31 scales of items from an item factor analysis of a large sample of OPQ data.

In order to achieve these aims, it has been necessary to develop some item analysis parameters that are related directly to the level of 'complexity' or relationship between items and non-keyed scales. These parameters are defined within the global framework of 'signal-to-noise' analysis. That is, they all variously index the ratio of keyed scale item indices to non-keyed scale item indices. All parameters vary between 0 and 1, with 0 indicating no quantitative information available to distinguish a scale of items from any other scale or set of items in the test. A value of 1.0 indicates perfect discriminability between the scale of items and all other items and scales in the test. A value of .5 can be viewed as indicating 50 per cent discrimination between a scale and the 'background noise' or non-keyed items or scales. As with Kaiser's (1974) scaling of his index of factorial simplicity (another signal-to-noise procedure that looks specifically at item complexity across item vectors within a factor analysis), we also choose .60 as the lower bound for minimally acceptable parameter values. From the rationale above, this might well be considered a conservative value. Equivalent parameter values from evoked potential analysis in electroencephalography are generally above .8.

Specifically, the following measures were generated:

C_66%_ISNR

The computational formula is:

$$\text{C\_66\%\_ISNR} = \left[\frac{ITC_i^2}{ITC_i^2 + \left\{ \sum_{j \neq i}^{N} nsITC_j^2 / (N-1) \right\}} \right] / K,$$

where ITC = the keyed scale item–total correlation i. It is the item–total correlation for an item within the scale in which that item is identified as being a member

$nsITC$ = the correlation between an item and a scale score on which it is not assumed to be associated (non-keyed)

N = the number of scales in the test

K = the number of non-keyed scale ITCs ≥ 66% of the size of the keyed scale ITC

Thus for each item, this parameter indexes the ratio of squared 'keyed scale' correlation to the squares of the non-keyed ITCs, modifying this ratio by dividing it by the number of 'salient' non-keyed ITCs. This correction is required since as the number of scales increases in a test, the effect of one or two high correlations across other scales can be swamped in the calculation of a mean value. 'Salient' is defined as those correlations greater than or equal to two-thirds of the mean ITC for a keyed scale, with a hard lower bound of .15 (i.e. if two-thirds of the size of the keyed ITC is less than .15, then it is set to .15). In other words, a subjective decision is made here in deciding that a non-keyed ITC is critical when its size is greater than this value. Essentially this measure treats as a 'signal' the keyed scale ITC, and 'noise' as the remaining correlations across the non-keyed scales that are at least two-thirds the size of the keyed-scale ITC.

SQUAL

The *S*cale *QUAL*ity index. This parameter is an attempt to provide a single parameter that indexes the measurement quality of a scale of items as a whole, taking into account scale-item complexity, the signal-to-noise ratio of the scale, and the disparity of ITCs below the mean ITC within the scale. In essence it is an attempt to capture the many essential psychometric properties of a scale of items as a unitary parameter.

The formula is:

$$\mathrm{SSNR} = \frac{\left(\sum_{i=1}^{S} ITC_i^2 / S\right)}{\left(\sum_{i=1}^{S} ITC_i^2 / S\right) + \left(\sum_{j=1}^{NS} nsITC_j^2 / NS\right)} \qquad \mathrm{C\_SSNR} = \mathrm{SSNR} - \left(\mathrm{SSNR}\left(\frac{K}{S}\right)\right)$$

$$\mathrm{CB} = 1.0 - \left(1.0 - \left(\frac{\sum_{i=1}^{NB} \mathrm{C\_66\%\_ISNR}_i < 0.5}{NB}\right)\right) * \left[\frac{NB}{S}\right]$$

$$\mathrm{CR} = 1.0 - \left(1.0 - \left(\frac{\sum_{i=1}^{NL} \mathrm{ITC}_i < \mathit{BoundValue}}{(NL * \mathit{BoundValue})}\right)\right) * \left[\frac{NL}{S}\right]$$

$$\mathrm{SQUAL} = \mathrm{C\_SSNR} * \mathrm{CB} * \mathrm{CR}$$

where *ITC* = the keyed scale item–total correlation *i*. The item–total correlation for an item in the scale in which that item is identified as being a member

nsITC = the correlation between an item and a scale score on which it is not assumed to be associated (non-keyed)

S = the number of items in a target scale

NS = the number of non-keyed-scale items in the test, i.e. the number of items remaining in the test after those in the target scale are excluded

K = the number of items which correlate ≥ the specified value of either the mean target scale ITC, or the .5 bound respectively

NB = the number of items whose C_66%_ISNR value is less then .5 in a scale

NL = the number of items in a scale whose ITC is less than the mean ITC for that scale

CB = the correction for high complexity items

CR = the correction factor for low-ranging ITC disparity

BoundValue = the mean ITC for a scale or .5 if the mean ITC is greater than .5.

Thus for each scale, the ratio of mean squared ITCs to mean squared non-keyed ITCs is indexed as a Scale Signal-to-Noise Ratio (SSNR). Once again a correction is applied based upon identification of salient correlations that are greater than or equal to a specified bound value. For the OPQ analyses, two specified bound values were used: the mean ITC (corrected item–total) correlation for each scale and a constrained maximum value of .5 or the mean ITC (if lower) for each. In the case of the OPQ, for each target scale there are eight items, with 240 non-keyed items. Non-scale items which correlated significantly with the keyed scale score would have little effect on this uncorrected SNR parameter as the division by 240 of the sum of squares would decimate the effect of the few salients. Therefore, the scale SNR is corrected by treating the salients as of equal 'signal' strength to a keyed item ITC. The logic of this is that if four 'external' items correlate as highly with the scale score as do the eight items within the scale, then the C_SSNR parameter would indicate a 50 per cent level of 'noise' in discriminating the scale from the remainder of the test items. If greater than eight items correlate higher than the bound value with the target scale score, the parameter is set to 0. *The scale measure can no longer be identified as distinct from a subset of the remaining items in the test.* Two other correction factors are then applied to the C_SSNR parameter, the first (CB) is a correction based upon the relative size of the C_66%_ISNR coefficients in a scale that are less then .5 in size. The correction is then weighted by the number of these 'bad items' in order to provide some degree of sensitivity. The second correction parameter (CR) is also a weighted factor that indexes the relative disparity in ITCs below the mean ITC in a scale. This coefficient is sensitive to low ITCs within a scale that may itself contain many high ITCs. The SQUAL parameter is thus a complex function of signal-to-noise, item complexity, and ITC disparity within a scale.

The term 'quality of measurement' has been chosen to best represent the meaning to be attributed to this complex parameter. The parameter is obviously not capable of determining the *utility* of measurement made by a scale of items.

As part of the SQUAL calculations, the items which are not part of that scale are correlated with the scale score. The number of items correlating higher than the specified bound value are noted. If five or more such items correlate in this way, they are treated as a scale and the correlation between the targeted scale score and 'new' scale score is computed, as is the 'new' scale alpha coefficient. Of course, the 'new' scale can be composed of any of the remaining 240 OPQ items that are not keyed on the target scale. This analysis is useful in highlighting alternative representations of a target scale generated from other items in the same test.

Test Quality Index

The mean of the SQUAL indices for a test. A summary parameter that indexes the measurement quality of a test as a whole.

Test Complexity Index

This is a summary parameter that attempts to describe the complexity of a test as a single numerical index. It is computed by summing the number of C_66%_ISNR coefficients with values of less than .5 (less than 50% 'signal' in an item) and dividing this sum by the number of items in the test. This value is expressed as a percentage and provides another summary parameter indexing the discriminability of test items within a test as a whole.

Method

Participants

Six hundred and twenty-one (gender unidentified) individuals aged between 18 and 50 years provided item-level data on the OPQ Concept 5.2 Questionnaire (Sample 1). The majority of the participants completed the questionnaire as part of a job application process within various companies within the UK. Another 390 male and 30 female applicant forms (age not identified) from within other companies' data were also used (Sample 2). The questionnaires were administered according to Saville & Holdsworth's procedural guidelines. A third sample of 816 male and 44 female job applicants provided scale level only data (Sample 3).

Questionnaire

The OPQ Concept 5.2 Questionnaire normative questionnaire contains 248 items assessing 31 scales, eight items per scale, with items answered using a five-point rating scale. Table 1 provides a list of the scale names (in addition to the coefficient alphas for the combined Sample 1 and Sample 2 data).

Factor analyses

In order to compute two of the measures detailed below, maximum likelihood factor analyses (MLFA) of the OPQ item data were undertaken. Principal components analysis (PCA) and image component analysis was also undertaken in order to allow computation of two tests of factor extraction quantity: the Velicer MAP test (Velicer, 1976) and Autoscree (Barrett & Kline, 1982). Factor rotations used hyperplane maximized direct oblimin rotation with hyperplane bandwidth set at ±0.1, and the δ parameter swept from -10.5 to $+0.5$ in steps of $+0.5$. With regard to the use of MLFA, principal component analysis generates components that account for the maximum variance possible within the *original sample* matrix. These components will explain as much or more variance than any other factoring method. However, they explain *sample* variance. That is, the loadings so produced by PCA are not necessarily the most likely to represent the population values. PCA does not take into account that the sample is drawn from a population, it simply computes its parameters based upon the sample data at hand. MLFA on the other hand, is a factor solution based upon the principle that the sample matrix of correlations is a sample from a population matrix of such values. The aim is therefore to produce maximum likelihood estimates of the factor loadings such that these loadings are those that are most likely to occur given the properties of the sample correlation matrix (means, variances, and covariances). Basically, MLFA is an estimation of population values from sample values, much in the same way as the sample mean and standard deviations are maximum likelihood estimates of the population mean and SD. Whereas PCA will maximize the variance explained by each component (without regard to best reproducing the actual correlations within the correlation matrix), MLFA will attempt to directly reproduce the actual correlations within the correlation matrix, from their common and unique factors. In conjunction with the MINRES factor analysis technique, MLFA is now generally considered to be the only other acceptable exploratory common factor analysis model.

Factorial signal-to-noise parameters

Index of Factorial Simplicity (IFS). This parameter was introduced by Kaiser (1974). It is a measure of the complexity of an item based upon its factor loadings. The formula is:

$$\mathrm{FS} = \frac{F\sum_{j=1}^{F} v_{ji}^{4} - \left(\sum_{j=1}^{F} v_{ji}^{2}\right)^{2}}{(F-1)\left(\sum_{j=1}^{F} v_{ji}^{2}\right)^{2}}$$

where F = the number of factors
v = the loading for variable i on factor j

The parameter varies between 0 and 1, with 0 indicating maximum item complexity and 1 indicating maximum simplicity (all but one item loading exactly zero). A value below .5 is considered unacceptable by Kaiser, with values above a minimum of .6 considered as acceptable. From the individual item IFS coefficients, we can compute the mean IFS for a test, or for a scale.

Essentially, this parameter indexes the rotational simple structure for a test scale; a good scale is one where the loadings of its items are significantly higher on a single factor than across all other factors. The IFS indexes the degree to which this aim is achieved in practice. There is of course one problem with the measure, that is, it cannot distinguish whether a scale of items is loading on the same specific factor, only if the items are loading significantly on one or more factors. Thus, if we had five items (which form a scale) loading on five factors, each item loading .5 on a different factor than the rest, and all other loadings of .0, the IFS coefficients would all be 1. However, the proposal that the five items measure a single dimension of behaviour would not be valid! This underlines the interpretation of the IFS as a measure of item complexity independent of item–scale composition.

Factorial Absolute Signal-to-Noise Ratio (ANR). This coefficient is based upon Fleming's (1985) measure of the index of fit for factor scales. The scale signal-noise-ratio (SSNR) coefficient, as detailed above, is a direct analogue and use of Fleming's formula, where item–total correlations were defined as the values to be squared. Fleming used factor loadings as the basis for his signal-to-noise ratio. However, in the factor analytic domain, it was decided to modify Fleming's formula by using absolute value loadings rather than squared loadings. This provides a closer analogue as to how a user interprets columns of factor loadings and is generally more sensitive to the size ratio of salient and non-salient loadings. The coefficient is modified for the same reasons, and using the same formula as for the SSNR parameter above. It must be reiterated that the use of the uncorrected Fleming formula is of little value. As the ratio of the number of items in a scale to the number remaining in the test becomes larger, the sensitivity of the coefficient to index any useful information decreases correspondingly. The correction 'bound' value for the ANR parameter uses the conventional lower bound of .3 for treating a factor loading as significant. Non-scale variables which load equal to or greater than this value on a scale factor are summed as significant non-salients. Thus the C_ANR parameter is the direct factorial equivalent of the C_SSNR parameter. Finally, an additional correction is made to the C_ANR parameter that indexes the number of items within a scale that load less than .3 on the scale factor. This correction has to be implemented in order to adjust for the specific case where only some of the keyed items for a scale actually load significantly on a factor. Given no other items load significantly on this factor, it is possible to still maintain a high signal-to-noise ratio even though maybe only half the number of keyed items load above .3. Thus, a correction is applied in the same way as that for the SQUAL parameter, using only the CR parameter described above, with loadings replacing the ITC values, and the *BoundValue* replaced with a constant of .3. The signal-to-noise parameter is adjusted for non-keyed salient loadings and for the quantity of keyed items loading less than .3 on the scale factor. The absolute noise ratio is thus corrected for non-keyed items loading too highly on a factor, as well as corrected for keyed items loading too low. C_ANR values above .7 are invariably generated by item factors where most, if not all, the keyed items load greater than .3 and are the only such loadings on a factor.

Results

Table 1 provides the corrected mean item–total correlations, mean inter-item correlations, and coefficient alphas computed from the combined data of Samples 1 and 2 ($N = 1041$). These are compared to the alphas from the UK normative sample for the OPQ scales, based upon 2987 individuals. All mean correlations were computed using the Fisher *z* distribution transform.The alphas for the scales Democratic and Caring have been replaced with those published earlier by Saville & Holdsworth, based upon a sample of 2306 individuals (SHL update, 1992). The 1993 Concept Model values are in error—as can be easily verified by computing the standard error of measurement (SEM) values for the two scales using both the 1993 and 1992 values. Using the 1992 values, we are best able to approximate the 1993 manual SEMs.

As can be seen from this table, the alphas are in broad agreement with those from the normative sample. Although values less than .70 have been highlighted as indicating sub-optimal consistency, this is perhaps questionable given the short length of the test scales. However, with alphas greater than .8, the counter-argument is that the eight item scales

Table 1. Mean item–total correlations (ITC), mean inter-item correlations (*R*), and alpha coefficients for the joint Sample (1 + 2) item data (*N* = 1041). The alpha coefficients are compared to OPQ normative data provided in the Concept Model Test Manual (*N* = 2987)

		Joint sample data			
Scale[a]		Mean ITC	Mean *R*	Alpha[b]	Normative alpha[b]
R1:	Persuasive	.50	.32	.79	.74
R2:	Controlling	.51	.33	.79	.88
R3:	Independent	.28	.15	**.55**	**.63**
R4:	Outgoing	.64	.48	.88	.86
R5:	Affiliative	.48	.30	.76	.75
R6:	Socially confident	.58	.41	.83	.86
R7:	Modest	.65	.48	.88	.75
R8:	Democratic	.34	.19	**.64**	**.65**[c]
R9:	Caring	.40	.25	**.69**	.77[c]
T1:	Practical	.72	.55	.90	.87
T2:	Data rational	.70	.54	.90	.88
T3:	Artistic	.63	.46	.87	.83
T4:	Behavioural	.43	.26	.73	.73
T5:	Traditional	.45	.28	.75	.74
T6:	Change oriented	.31	.17	**.61**	**.62**
T7:	Conceptual	.47	.29	.76	.75
T8:	Innovative	.59	.42	.85	.84
T9:	Forward planning	.31	.16	**.60**	**.57**
T10:	Detail conscious	.52	.35	.80	.74
T11:	Conscientious	.47	.30	.76	.80
F1:	Relaxed	.59	.40	.84	.83
F2:	Worrying	.43	.26	.73	.73
F3:	Tough minded	.56	.38	.83	.83
F4:	Emotional control	.58	.40	.84	.76
F5:	Optimistic	.47	.30	.76	.73
F6:	Critical	.32	.17	**.61**	**.60**
F7:	Active	.55	.36	.82	.79
F8:	Competitive	.52	.33	.79	.71
F9:	Achieving	.38	.22	**.69**	**.63**
F10:	Decisive	.44	.26	.74	.76
F11:	Social desirability	.40	.23	.70	**.67**

[a] The copyright OPQ scale names and normative alphas are reproduced with permission from Saville & Holdsworth Ltd.
[b] Figures in bold indicate coefficients less than .70
[c] These two values are taken from an earlier (SHL Update) published norm of 2306 individuals. The 1993 manual values are in error (see text)

may be too internally consistent to support the proposition that they are measuring a general domain of behaviour. The mean inter-item correlations, however, do not support this criticism except perhaps with regard to the Practical and Data rational scales where the parameter values are .55 and .54, with alphas of .90. Note also that the mean item–total correlation (ITC) for these scales are .72 and .70 respectively. Table 2 presents the results from the item complexity analysis, that is, examining each scale in terms of the non-scale

Table 2. Item Complexity Analysis noting the number of non-keyed scale items that correlate greater than or equal to the keyed scale mean item–total correlation (ITC). The second set of values is for the case where mean ITC values are constrained to be .5 or less

		Mean ITC Bound			.5 Bound		
Scale		Items	Alpha	Corr.	Items	Alpha	Corr.
R1:	Persuasive						
R2:	Controlling						
R3:	Independent	7	*.69*	*.53*	7	*.69*	*.53*
R4:	Outgoing				5	*.82*	*.73*
R5:	Affiliative				1		
R6:	Socially confident	3			5	*.80*	*.80*
R7:	Modest						
R8:	Democratic						
R9:	Caring						
T1:	Practical						
T2:	Data rational						
T3:	Artistic						
T4:	Behavioural						
T5:	Traditional						
T6:	Change oriented	4			4		
T7:	Conceptual						
T8:	Innovative						
T9:	Forward planning	18	*.82*	*.65*	18	*.82*	*.65*
T10:	Detail conscious						
T11:	Conscientious						
F1:	Relaxed				3		
F2:	Worrying	6	*.81*	*.66*	6	*.81*	*.66*
F3:	Tough minded						
F4:	Emotional control						
F5:	Optimistic	1			1		
F6:	Critical						
F7:	Active						
F8:	Competitive						
F9:	Achieving	1			1		
F10:	Decisive						
F11:	Social desirability	1			1		

Notes. The column headed 'Items' provides the number of non-keyed scale items that correlate according to the criterion.
The column headed 'Alpha' (= reliability) provides the alpha reliability for the items scored as a new scale. Five or more items were required for the creation of a 'new' scale.
The column headed 'Corr.' (= correlation) provides the correlation between the items scored as a new scale, and the designated scale.

items which correlate (higher than the specified bound values) with the target scale score. The two bound values were the mean ITC for a scale and a value of .5 (or the mean ITC if lower.

These results show that when looking at non-scale items which correlate higher with a scale than that scale's mean ITC five scales have significant item intrusion (Independent, Socially Confident, Change Oriented, Forward Planning, and Worrying). Where five or more such items correlate with a target scale, these items are themselves formed into a scale, scored, and these scale scores correlated with the target scale scores. In addition, an alpha is computed for these 'new' scales. The column headed 'Alpha' provides the alpha coefficient for this new 'scale', the column entitled 'Corr.' provides the correlation between the 'new' scale and the target scale. In almost every case, the new scale alpha exceeded the target scale alpha, across both bound values. As the maximum bound is reduced in size (where relevant) from the mean ITC to .5, the number of these item 'incursions' increases dramatically. This analysis provides a simple but powerful method to view the level of 'common' variance amongst items and scales in a test. The results from this table indicate that there is considerable measurement overlap across many items in the test. That is, although the items may correlate higher with their own score than with another scale score, 16.5 per cent of the items in the test correlate higher than the mean ITC of a non-keyed scale, and up to 21 per cent correlate higher than .5 (or their mean ITC) with non-keyed scales.

Table 3 provides the SQUAL ratios for the OPQ scales, based upon a .5 item complexity bound. Scales with values below .60 are considered as exhibiting poor discriminability of item content and poor measurement quality. Values below .5 are indicative of serious degradation in the measurement properties of a scale. We reiterate, the rationale for these values is based upon the properties of signal-to-noise ratios, where a value of .5 is representative of 50 per cent 'signal' and 50 per cent 'noise'. Here 'signal' is defined as the clarity with which keyed items correlate with their own keyed scale vs. the 'noise' of their correlations with other scale scores.

As a comparison to these item-level indices, the factor-level indices IFS and C_ANR were both computed for the maximum likelihood factors. A problem encountered in the factor-based measures is that associated with the number of factors extracted and rotated. Compression of the factor space tends to decrease both IFS and C_ANR, while excessive expansion is likely to also decrease the C_ANR, while the IFS might be expected to be reasonably stable. Thus, four rotation solutions were computed based upon Matthews & Stanton's (1994) extraction of 21 factors, the Velicer MAP test indicator of 26 (PCA) and 28 (image) factors, and Autoscree indicators of 17 and 21 factors for PCA and image respectively. From these solutions, it was hypothesized that a full 31 factor rotation might provide the optimal C_ANR parameters for the OPQ scales. Further, as a by-product of the use of MLFA, it is possible to compute a test for the statistical significance of the number of factors extracted. For the dataset used ($N = 1041$, variables $= 248$), it was found that 30 factors were considered 'significant' at $p = .05$. Thirty-one factors yielded a p value of .21. However, since the test is notably sample-size dependent, the sample size was reduced to 400 and the analyses recalculated. This yielded a decision of 26 factors to be retained. At $N = 300$, the decision was 23 factors to be retained. Since little could be inferred from these values other than more than 20 or so factors seemed a reasonable number to extract, it was decided to proceed with the rotation of 31 factors.

Table 3, as well as showing the SQUAL values, also details the IFS and C_ANR coefficients for each scale, as well as the number of absolute valued non-keyed item loadings greater than .3 on each scale, and the number of keyed items loading on each scale. The reported IFS and C_ANR parameters were computed over the factor pattern rather than factor structure matrices as in all cases, the simplicity and SNR ratios were significantly greater for pattern loadings. This is due to the fact that in every matrix computed some

Table 3. Scale Quality Indexes (SQUAL) based upon corrected scale signal-to-noise ratios (C_SSNRs) computed using a .5 bound, Kaiser indices of Factorial Simplicity (IFS), factor-based corrected absolute noise ratios (C_ANR), the number of non-keyed and keyed items loading greater than .3 on each scale factor

Scale		SQUAL[a]	IFS[a]	C_ANR[a]	Non-keyed >| .3|	Keyed >| .3|
R1:	Persuasive	**.56**	**.55**	.73	0	4[b]
R2:	Controlling	.75	.73	.92	0	7
R3:	Independent	**.05**	**.49**	**.53**	0	2[b]
R4:	Outgoing	**.24**	.60	**.11**	7	7[b]
R5:	Affiliative	.61	.64	.86	0	7
R6:	Socially confident	**.15**	.69	**.11**	7	7[b]
R7:	Modest	.88	.84	.96	0	8
R8:	Democratic	.81	.62	.89	0	7
R9:	Caring	.80	.66	.83	0	5
T1:	Practical	.89	.90	.96	0	8
T2:	Data rational	.91	.88	.96	0	8
T3:	Artistic	.91	.89	.95	0	8
T4:	Behavioural	.86	**.57**	.87	0	7
T5:	Traditional	.82	.75	.93	0	7
T6:	Change oriented	**.21**	**.45**	**.41**	0	2[b]
T7:	Conceptual	.76	.67	.89	0	7
T8:	Innovative	.87	.81	.94	0	8
T9:	Forward planning	**.00**	**.34**	**.00**	8	1[b]
T10:	Detail conscious	.83	.67	.90	0	7
T11:	Conscientious	.73	.65	.79	1	8
F1:	Relaxed	**.46**	**.46**	.75	1	7
F2:	Worrying	**.12**	**.55**	**.49**	2	3[b]
F3:	Tough minded	.80	.76	.93	0	8
F4:	Emotional control	.88	.80	.93	0	7
F5:	Optimistic	**.52**	.66	.89	0	6
F6:	Critical	**.56**	**.53**	**.53**	0	3[b]
F7:	Active	.74	.72	.87	0	6
F8:	Competitive	**.57**	.71	.80	0	6
F9:	Achieving	**.52**	**.33**	.67	0	3[b]
F10:	Decisive	.78	.67	.80	1	8
F11:	Social desirability	.64	**.58**	.77	1	6

[a] Figures in bold indicate coefficient considered less than optimal (below .60).
[b] Indicates a poorly defined scale (see text).

factor correlation was present between some of the scales. This invariably has the effect of boosting the majority of the loadings in a structure matrix given that the relationship between factors is built into the correlation between a factor and a loading. While arguments can be put forward for the superiority of interpretation of structure correlations vs. pattern beta weights, the specific aims of this particular factor-based analysis methodology demands that greater emphasis is placed on the relationship between an item and its factor, irrespective of factor correlations. Thus, the pattern loadings are defining the relative contribution of a factor to each of its variables, irrespective of the relationship between factors.

The indices in Table 3 show broad agreement between item-level SQUALs and the factor-based measures of complexity and C_ANR. Values less than .60, for all measures including Kaiser's IFS, are considered as indicative of poor scale discriminability and measurement quality. Using the data from this table in comparison with those in Tables 1 and 3, the reason why some of the scales are failing becomes apparent. The Independent scale has a low alpha of .55, a fairly low mean ITC of .28, and has seven item-analysis-level incursions which correlate on average at .53 with the scale score. Only two if its keyed items load greater than .3 on the factor. Outgoing and Socially Confident are scales with high alphas but are 'noisy' with items from each other scale. Note that the IFS coefficients are insensitive to this kind of overlap, but both the SQUAL and C_ANR parameters detect the cross-loadings on these scales. The Forward Planning scale has a moderately low alpha of .60, but with 18 item-analysis-level incursions, six alone from the Conscientiousness scale. The alpha for this new scale is .82. The SQUAL for this scale is .00, as is the C_ANR. Only one keyed item greater than .3 loads this factor, another eight non-keyed items also load it. The Relaxed and Worrying scales show almost the same effects as for Outgoing and Socially Confident, but to a lesser degree.

From this table, we have selected nine scales as being of dubious psychometric validity. This selection is based upon the consideration of low parameter values (<.60) and keyed-item loading counts of less or equal to only 50 per cent of keyed items being identified for a factor-scale (>.3). The scales considered of dubious psychometric validity are: Persuasive, Independent, Outgoing, Socially Confident, Change Oriented, Forward Planning, Worrying, Critical, and Achieving.

Table 4 provides a scale intercorrelation frequency count histogram for the OPQ data using the combined data of Samples 1, 2, and 3 ($N = 2301$). These data show that fewer than 1.5 per cent of scale intercorrelations are greater than .5 and above. The maximum correlation of .74 was observed between the scales Socially Confident and Outgoing. This is what would be expected from the complexity parameter values given in Table 3.

The overall Test Quality Index (TQI), Test Complexity Index (TCI), and IFS for the OPQ, based upon the combined Sample (1+2) dataset of $N = 1041$ cases, using the constrained .5 *BoundValue* are .62, 25.81 per cent, and .65 respectively. The OPQ, as would be expected from the various parameter analyses above, has summary values indicating marginal acceptability in terms of overall measurement quality. The TCI also suggests that this is where the test measurement is weakest—too many items are significantly associated with the measurement made by other scales on the test ('significantly' defined by items correlating greater than .5 or the mean item–total correlation with other scale scores). However, since these items are focused on a few scales, it would not be difficult to improve these properties of the test by judicious item selection that minimized the probably needless overlap between scales and non-keyed items.

Finally, since the essence of the Concept Model for the OPQ is the proposition that the scales are drawn from three psychological domains, rating, thinking and feeling, it was decided to test this proposition in a quantitative fashion, using structural equation modelling. Two simple models were specified according to the naming convention of the scales

Table 4. Scale intercorrelation histogram for the OPA data, Samples 1, 2 and 3 (N = 2301). Absolute value correlations are tabulated

Range	Quantity
.0 to .09999	153
.1 to .19999	163
.2 to .29999	85
.3 to .39999	41
.4 to .49999	16
.5 to .59999	4
.6 to .69999	2
.7 to .79999	1
.8 to .89999	0
.9 to 1.00000	0

R1, R2, R3, . . ., R9 from the relating latent variable manifest indicators, T1, . . ., T11 the thinking latent variable manifest indicators, and F1, . . ., F10 the feeling latent variable manifest indicators. The data used for the modelling were those based upon the scale scores from the combined Samples (1+2+3) with N = 2301. In Model 1, the three latent variables were specified as orthogonal to one another, in Model 2, they were allowed to correlate. Maximum likelihood estimation was used via the Statistica-SEPATH (Steiger, 1995) structural equation software. Figure 1 presents the path diagram and fitted coefficients for Model 1. The model chi-square was 10 938.109 based upon 402 d.f. ($p < .001$) with a Bentler–Bonnett normed fit index value of .463 and Comparative Fit Index of .504. The James–Mulaik–Brett Parsimony fit index was .46. None of these indices is even close to their minimal accepted values—generally all above .90. The Steiger–Lind RMSEA index was .118. The accepted value for this index is less than .10 (Steiger, 1995). For Model 2, as shown in Fig. 2, the model chi-square was 9655.022 based upon 399 d.f. ($p < .001$) with a Bentler–Bonnett normed fit index value of .528 and a Comparative Fit Index of .564. The James–Mulaik– Brett Parsimony fit index was .512. The Steiger–Lind RMSEA index was .114.

It is obvious looking at both figures that neither model is a suitable fit to the data. The path coefficients (factor loadings) from the latent variables to their respective manifest indicators show values well below that required to demonstrate model fit. This brief analysis is a clear signal that the fundamental three-domain OPQ concept model is unlikely to be verified by objective, quantifiable methodology. The use of the R, T, and F initials to classify scales is therefore seen as quite misleading and rather subjective.

Discussion

Similar to the results reported in the Matthews & Stanton paper, only 22 out of the 31 OPQ scales emerge clearly from an item and factor-level analysis of the questionnaire. These scales are distinguishable from one another with low item and scale complexity.

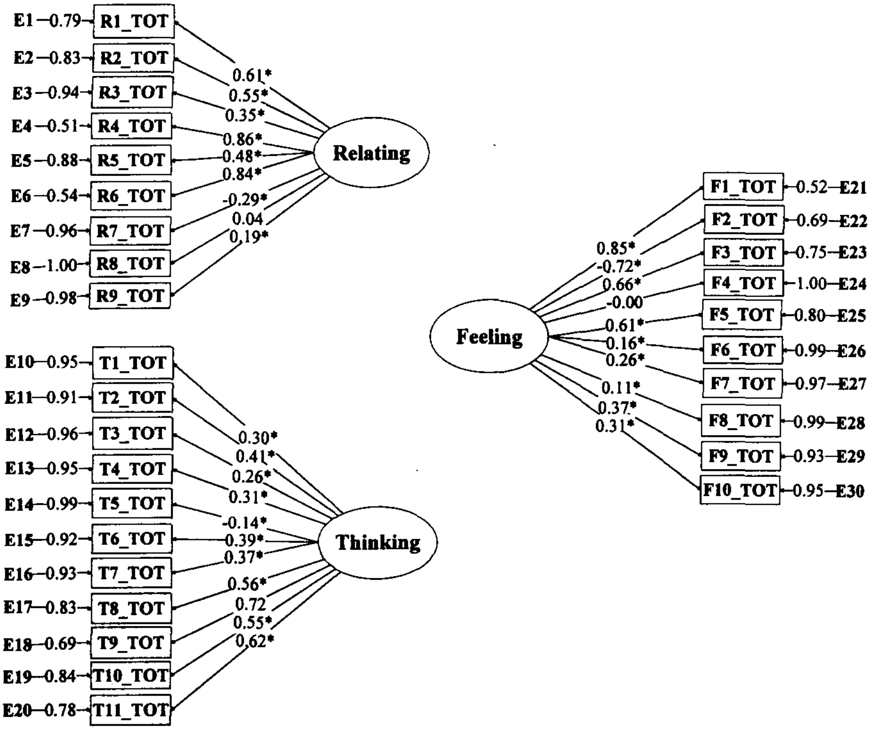

Figure 1. Structural equation model for the the three hypothesized domains underlying the OPQ Concept Model 5.2 Questionnaire. The latent variables are orthogonal to one another.

They can be said to be measuring relatively discrete segments of behaviour, with little overlap in measurement between scales. However, the nine scales of Persuasive, Independent, Outgoing, Socially Confident, Change Oriented, Forward Planning, Worrying, Critical and Achieving demonstrated either high degrees of item overlap, item complexity, poor factorial signal-to-noise ratios, or were simply unidentifiable as factors defined by keyed item loadings. It is within these nine scales of the test that many of the identified measurement problems with the OPQ reside. From the outline of the OPQ development in the introduction, and from the subsequent analyses above, it can be hypothesized that test development was driven primarily by a desire to maximize the alpha coefficient where possible, and to maximize the item–total correlation for each item in a scale, whilst maintaining a constant eight items per scale. The OPQ manual indicates that item selection for each scale ensured that each item correlated higher with its own scale than it did with other scales. This aim alone is insufficient in ensuring that the items are not correlating substantively with other scale scores. Items that correlate .6 with their own scale and .52 with another scale cannot be said to be making optimal, discriminable measurement. Some of the scales noted above as failing to be discriminable from this background 'noise' are full of item incursions that have this multiscale correlation property. These items inflate scale correlations which can then be used to demonstrate the notion of a higher-order 'superfactor' which contains 'primaries' In fact, what we have is the case of an item mix which, by definition of the admixture of 'common variance items'

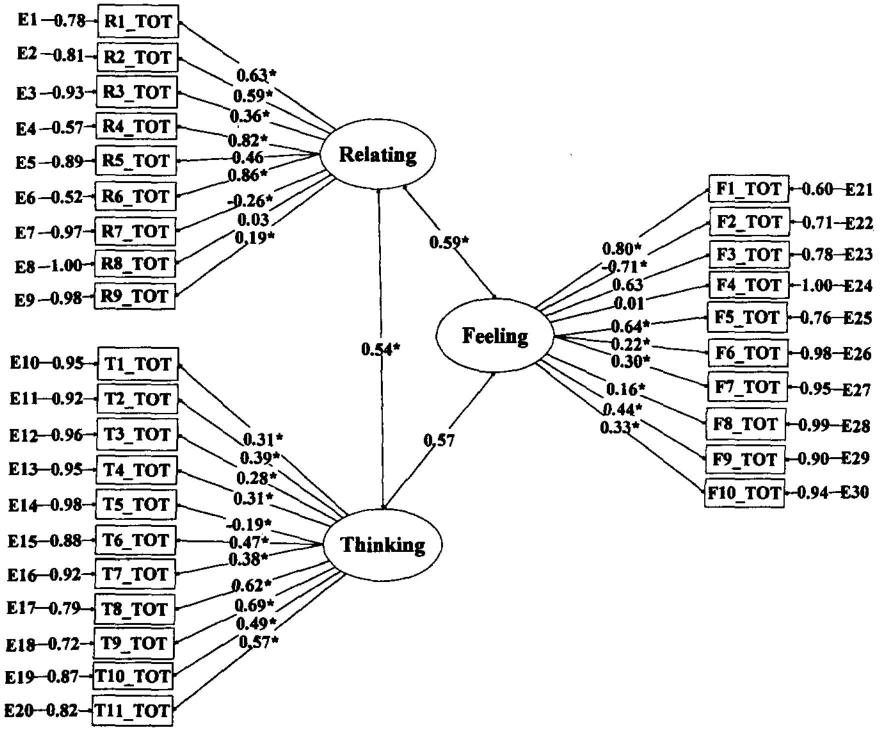

Figure 2. Structural equation model for the three hypothesized domains underlying the OPQ Concept Model 5.2 Questionnaire. The latent variables have been allowed to correlate between one another.

and actual discrete behavioural items, produces quasi-primaries which can then be correlated and defined to form a higher order factor or concept. The actual breadth of measurement though is more limited than one would expect from the use of the term 'higher order'.

Within the domain sampling model, increasing measurement covariance between items produces at best, redundancy of measurement and overlong tests, and at worst, incorrect interpretation of test scale scores based upon a scale name that implies discrete behavioural measurement but in fact is composed of items that share their variance with other test scales. Surprisingly, it appears that some test constructors do not see item overlap (indexed by item and factorial complexity) as a significant measurement flaw, rather, it is viewed as being of minor importance in comparison to maximizing internal consistency coefficients, with inter-scale correlations left 'floating free' and attracting concern only if they exceed some notional, generally unspecified, value. For example, in the paper by Matthews & Stanton (1994), they note a recent personal communication from Saville . . . 'Item analyses were used to ensure that the scales were reliable and not too highly correlated with one another, but it is not claimed that the scales are factorially pure'. Note that the criterion for 'too highly correlated' is not specified and that factorial impurity implies items that are measuring other aspects of behaviour that may not be specifically associated with their scale designation or name (Jackson, 1970). Given our arguments in the Introduction above, concerning the property of unidimensionality within classical test theory, it is clear that concentrating solely on internal consistency without also preserving homogeneity of measurement is not the optimal way to construct psychometric tests.

It might be argued that since some of the scales identified here as psychometrically 'poor' also possess reasonable to high alpha coefficients, then at least the measures being made are likely to exhibit reliability and some precision of measurement, even though confounded with the measurement made by other scales. Although the argument can be considered sound in principle, the consequences of such an argument in practice are not desirable. Without a very clear psychological model and measurement model guiding the test development process, highly confused measurement scales (such as those identifed in the OPQ) are likely to be the outcome. Although this may not affect the eventual use of the scales (as scales confounded by a similar item set will all correlate significantly positively with one another), it is quite unnecessary and is the result of suboptimal test development at both the psychological modelling and psychometric measurement levels. For example, to generate two scales say of worrying and relaxed requires the isolation of the common variance shared between the scales, partitioning this off as a new scale of X, and examining the specific scale components remaining and subsequently generating (if felt necessary or desirable) one or more highly focused scales that provide a greater fidelity of measurement than the two original 'mixed' scales. In the current combined sample scale score dataset of 2301 cases, worrying and relaxed correlate at −.632. If we correct for the unreliability of measurement, we have a correlation of .81. Not identical, but nevertheless indicating 65 per cent shared variance. More relevant perhaps is the recognition that 'relaxed' would normally be considered the opposite pole to 'worrying' on a bipolar trait of anxiety. The psychological model that proposes that the two concepts are separable would need to be extremely specific in order to allow measures to be taken that can discriminate usefully between the two concepts. It is to be noted that Jackson (1970) and Wolfe (1993) also make this general point rather more forcefully and in more detail in a recent review of scale and test construction procedures.

In conclusion, from the analyses above and taking into account Matthews & Stanton's results, it is probable that the OPQ Concept Model 5.2 is overlong, containing a proportion of redundant and quantifiably complex items. Furthermore, there is little evidence supporting the existence of 31 discrete measurement scales. Focusing more on the positive features of these analyses, it has been demonstrated that the OPQ is making high quality measurement on up to 22 out of the 31 scales currently existing within the test. It is possible that with further development of the test, some of the remaining items/scales might be successfully added to these 22 discriminable measures.

References

Barrett, P. T. & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study and Group Behaviour*, 1, 1–23.

Barrett, P. T. & Kline, P. (1982). An item and radial parcel factor analysis of the 16PF questionnaire. *Personality and Individual Differences*, 3, 259–270.

Barrret, P. T. & Paltiel, L. (submitted). The OPQ Concept 5.2 questionnaire: too many items for too few concepts? *Selection and Development Review.*

Fleming, J. S. (1985). An index of fit for factor scales. *Educational and Psychological Measurement*, 45, 725–728.

Hofstee, W. K. B, De Raad, B. & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, 63(1), 146–163.

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current Topics in Clinical and Community Psychology*, vol. 2. New York: Academic Press.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–36.

Matthews, G. & Stanton, N. (1994). Item and scale factor analyses of the Occupational Personality Questionnaire. *Personality and Individual Differences*, 16(5), 733–744.

Matthews, G., Stanton, N., Graham, N. C. & Brimelow, C. (1990). A factor analysis of the scales of the Occupational Personality Questionnaire. *Personality and Individual Differences*, 11, 591–596.

Peabody, D. & Goldberg, L. R. (1989). Some determinants of factor structures from personality-trait descriptors. *Journal of Personality and Social Psychology*, 57, 552–567.

Saville, P., Holdsworth, R., Nyfield, G., Cramp, L. & Mabey, W. (1993). *Occupational Personality Questionnaires: Concept Model Manual and User's Guide*. Esher, Surrey: Saville & Holdsworth Ltd.

Saville, P. & Sik, G. (1995). *Reductio Ad Absurdum? Selection and Development Review,* 11(3), 1–3.

SHL Update (1992). The Preliminary Results from the BRMB Survey of a Representative Sample of 2,838 UK Adults. Esher: Saville & Holdsworth Ltd.

Steiger, J. H. (1995). *Manual to Statistica-SEPATH*. Tulsa, OK: Statsoft Inc.

Velicer, W. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, 321–327.

Wiggins, J. S. (1982). Circumplex models of interpersonal behaviour in clinical psychology. In P. S. Kendall & N. J. Butcher (Eds), *Handbook of Research Methods in Clinical Psychology*. New York: Wiley.

Wolfe, R. (1993). A commonsense approach to personality measurement. In K. H. Craik, R. Hogan & R. N. Wolfe (Eds), *Fifty Years of Personality Psychology*. New York: Plenum.

Received 22 February 1994; revised version received 25 September 1995