CAN A SINGLE ITEM REPLACE AN ENTIRE SCALE? POP vs. the OPQ 5.2

Paul Barrett and Laurence Paltiel

The Occupational Personality Questionnaire (Concept 5.2, OPQ; Saville et al., 1993) contains 248 items measuring 31 scales. Each scale has eight items. Responses to each item are on a normative five-point rating scale. Reliability coefficients (alpha) range from .57 to .88, with a median alpha of .75. Since alpha is known to depend on scale length as well as internal consistency, scales of just eight items may achieve high levels of reliability (greater than.7) due to item redundancy, e.g. where the items within a scale are simply reworded counterparts of one another. Instead of measuring a broad dimension of behaviour, it is possible that just one rather specific behavioural item is being assessed - using eight very similar items to achieve this.

If this is the case with the OPQ, then it should be possible to replace each of the 30 personality scales (excluding the social desirability Validity' scale) with a single composite item that captures the essential meaning of the scale and its constituent items. Thus the 240 items (30 scales) of the OPQ could be replaced by just 30 items. If item redundancy is a significant factor in the reliability of the OPQ scales, the following 'hypotheses' should hold true:

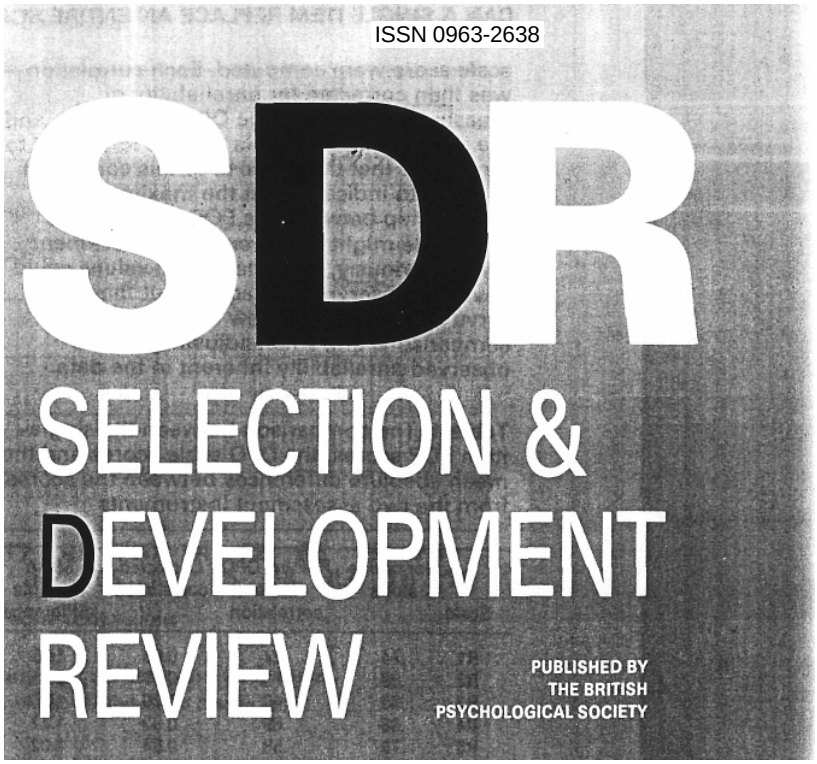
1. *The alpha coefficient should be strongly and positively related to the correlation between the single item and the OPQ scale score.*
   To justify replacing an eight-item scale with a single item, it is first necessary to show that the single item correlates highly with the full-scale score. The stronger the relationship between this correlation and the alpha coefficient, the greater the item redundancy and the more likely it is that just one item is sufficient to capture that behaviour measured by the scale.
2. *The difference between a single-item score and the full-scale score should be negligible.*
   Where both scores are transformed to comparable scales of measurement, we should observe only small differences between them, and the larger the alpha, the smaller the difference,

Questionnaire generation and administration

A 30-item questionnaire (the 'POP' questionnaire) was generated by the authors, based simply on a perusal of the items composing each scale of the OPQ. Each POP item attempted to capture the composite meaning of all eight items in an OPQ scale. Actual OPQ items were not used. Some examples of POP questionnaire items are presented in Appendix 1. The items all used an 11-point response format ranging from 'strongly disagree' to 'strongly agree' (see Appendix 2). Both the OPQ concept 5.2 and the POP questionnaire were administered concurrently to 420 managerial applicants (359 male; 61 female) as part of a corporate selection exercise.

Results

To test hypothesis 1 above, the correlations between each POP item and its respective OPQ



VOLUME 12 NUMBER 6
DECEMBER 1996

"CAN A SINGLE ITEM REPLACE AN ^ENTIRE SCALE? POP vs.'the OPQ 5.2 >By Paul Barrett and Laurence Paltiel 1

THE BPS LEVEL A OPEN LEARNING PROGRAMME - A review .Reviewed by Jeannette James ,4

TRAINING COURSES ON TESTING AND RELATED TOPICS 6

I

Consulting Editors: Dr S. Blinkhorn, Dr V. Dulewicz, DrN. Anderson

Co Editors: Roderic Vincent and Sean Boyle
Correspondence to:
Roderic Vincent
500 Chiswick High Road
London W4 5RG
Tel: 0181-956 2353
Fax: 0181 -956 2344
email: 100416.1423 @compuserve.com

Sean Boyle
76 Banbury Road
Oxford OX2 6JT
Tel: 01865 516202
Fax: 01865 510812
email: 101520.756 @compuserve.com

scale score were computed. Each correlation was then corrected for unreliability of measurement of both the OPQ scale score and the single item (see Table 1)[1]. It is important to remember that the purpose of this correction is solely to indicate what the maximum relationship between the POP item and the OPQ scale might be, *given no measurement error.* Obviously, these corrected values could never be observed in practice, but they do serve to demonstrate how close the comparisons are, when adjusting for the observed unreliability inherent in the data.

Table 1. The comparison between the single-item POP scores, the OPQ scale scores and the mean absolute differences between the scores from the two assessment instruments

	Normative alpha Scale	POP vs. OPQ scale correlation	Corrected correlation	Mean absolute diffference
R1	.74	.64	0.93	1.05
R2	.88	.73	0.97	0.81
R3	.63	.48	0.76	1.25
R4	.86	.69	0.93	1.47
R5	.75	.58	0.84	1.02
R6	.86	.64	0.86	1.31
R7	.75	.68	0.98	1.55
R8	.65	.58	0.90	1.35
R9	.77	.51	0.73	0.92
T1	.87	.88	1.00	1.07
T2	.88	.84	1.00	0.95
T3	.83	.79	1.00	1.30
T4	.73	.64	0.94	1.15
T5	.74	.56	0.81	1.50
T6	.62	.48	0.76	1.15
T7	.75	.68	0.98	1.43
T8	.84	.74	1.00	0.93
T9	.57	.51	0.84	1.31
T10	.74	.65	0.94	1.01
T11	.80	.63	0.88	0.99
F1	.83	.57	0.78	1.52
F2	.73	.56	0.82	1.82
F3	.83	.62	0.85	1.88
F4	.76	.66	0.95	1.37
F5	.73	.60	0.88	1.04
F6	.60	.27	0.44	2.50
F7	.79	.82	1.00	1.21
F8	.71	.64	0.95	2.32
F9	.63	.54	0.85	2.16
F10	.76	.34	0.49	2.87

'The single item unreliability could only be approximated. This was achieved by calculating actual item test - retest indices for three sets of personality questionnaire data, where the test -retest interval was approximately the same (10 - 14 weeks). The three datasets were composed of the USA version of the Eysenck Personality Questionnaire Revised (EPQR: /V= 165, 106 items, two-choice response), the Psytech Occupational Personality Profile (OPP: A/= 337, 98 items, five-choice response) and Gordon's Survey of Interpersonal Values (SIV: Normative version, A/ = 96, 90 items, five-choice response). Altogether, 20 scales and 294 items were used . The results of this analysis indicated that the EPQR median test - retest coefficient was .59, the OPP was .63 and the SIV .64. The minimum and maximum coefficients respectively for each test were EPQR: -.02 to .87; OPP: .32 to .86; and SIV: .37 to .82. From these data it was decided to use the most liberal estimate of test - retest; hence we took the median value of .64 as a reasonable estimate of overall expected item test - retest for the POP items.

The median correlation (uncorrected between the POP items and the OPQ scales is .64. The median of the corrected values is .89. Allowing for unreliability of measurement, half of the OPQ scales correlate above .90 with a single-item response. Figure 1 shows the plot of the alpha coefficients (in decreasing size order) against the POP vs. OPQ scale score correlations. A Spearman rank order correlation between the two sets of data (alpha coefficients and POP vs. OPQ scale correlations) was computed as a direct test of hypothesis 1. The observed correlation was -.69 ($p<.0001$, one-tail).

To test hypothesis 2, the OPQ scale scores were linearly transformed from their full-scale score to one that matched the 11-point POP response scale. Then the POP scores were subtracted from the OPQ rescaled score, and their absolute (unsigned) values noted. The mean difference score was then computed across all respondents. These results are also in Table 1.

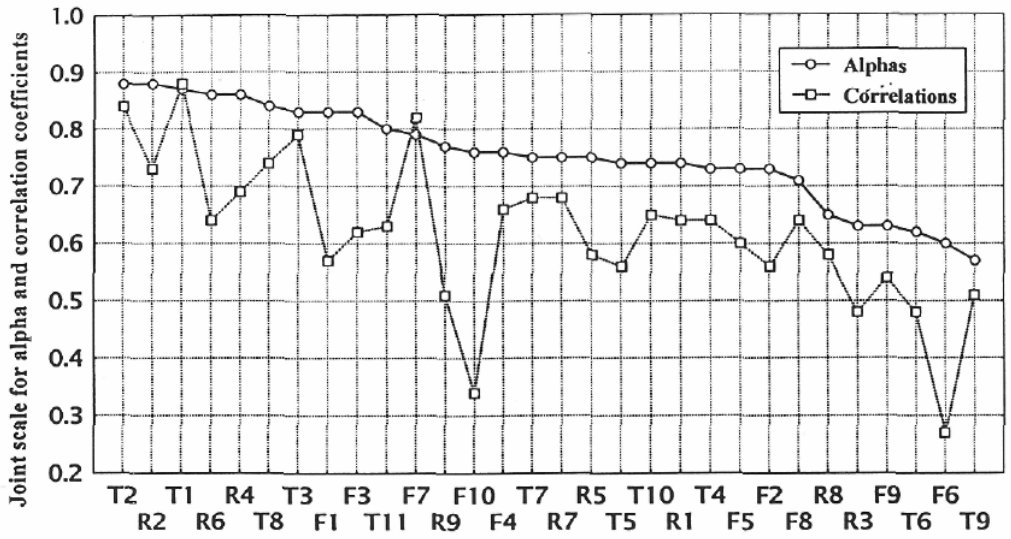
It can be seen that the mean absolute differences between the POP and OPQ rescaled scores are fairly low, with exception of scales F6, F8, F9 and F10 (all of which had mean absolute differences greater than 2.0). The mean of these data is 1.41, with median =1.30. Figure 2 compares the profiles of the mean POP scores and the corresponding OPQ rescaled scores. On the same graph is shown the difference between these two profile plots, scale by scale. This shows clearly where the similarities and dissimilarities occur over the length of the profile plot. To test the specific prediction in hypothesis 2, that the size of the score differences should be negatively related to the size of the respective alpha coefficient, we again computed a Spearman rank order correlation between these two items of data, which resulted in a coefficient of - .35 (p = .030, one-tail). This is not so compelling as the previous 'order-effect' coefficient, but the result does remain conceptually important.

Next, we looked at the range and variability of the difference scores — as an aid in judging the distribution of error around the median value for each scale. Since absolute differences tell us nothing about the proportion of over/or underestimated scores, we used the actual differences. It was found that the interquartile range (the middle 50 per cent) of errors around the median is confined to +1.0, or less, for 26 of the 30 scales.

Conclusions

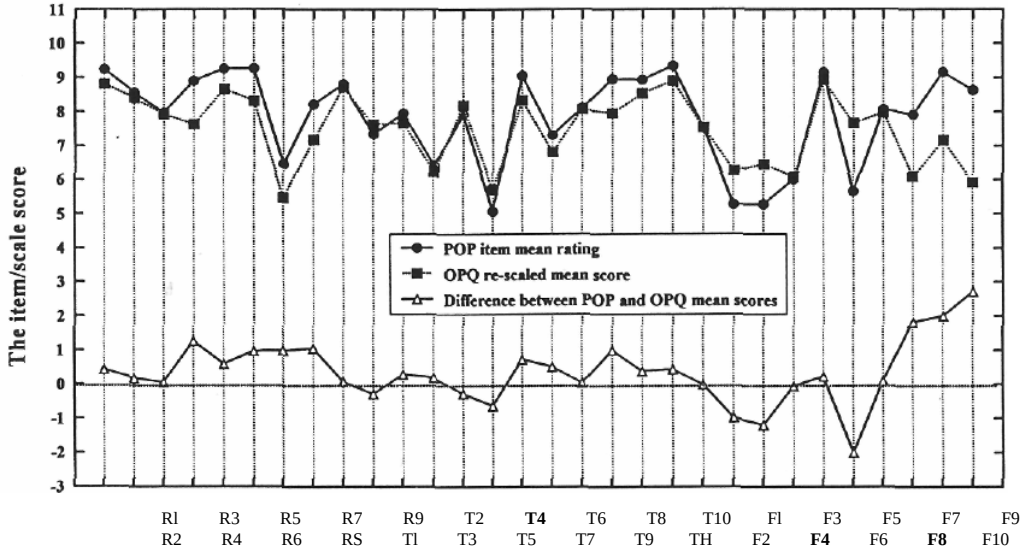
The results above indicate that a single item can capture virtually all of the measurement 'breadth' of many of the OPQ concept 5.2 scales. For one-third of the scales, corrected correlations of .95 or above between a single POP item and the eight items of the OPQ scale were obtained, which suggest that these scales are redundant. This is not to recommend that in fact just one item is used in future, because the inherent measurement unreliability of a single item is significant. However, it is a moot point whether up to eight repetitive items are needed. In two recent papers (Barrett & Paltiel, 1995a,b) we have shown that we could reduce 17 OPQ scales from eight to four or five items, while maintaining alpha reliabilities at a

SDR
December 1996
Volume 12 (6)



OPQ full-scale The OPQ Concept 5.2 scale designations

Figure 1. The POP/OPQ full-scale correlations and the normative OPQ alphas.
The OPQ scales are in descending order of size of the alpha coefficients.



The OPQ Concept 5.2 scale designations

Figure 2. Comparison between the POP item score and corresponding OPQ scale
(The OPQ scale scores have been rescaled into the 1-11 range of the POP items).

comparable level to those for the full scale, and achieving correlations of .92 or more between the long and short versions of the scales in every case bar one.

Saville & Sik (1995a,6) have responded to our previous observations, indicating that it is better to err on the side of redundancy than miss out a key item or concept. There are many issues in the 19956 rejoinder which we would like to comment upon. However, we think the issues can now best be resolved by the reader. If you have an OPQ 5.2 nearby, look at the actual items of the OPQ scales that correlate .95 and above with a single POP item, then decide for yourself if the arguments of Saville & Sik still hold. We agree that some scales are broader than others; we also agree that more than one item is required to reliably assess any behavioural construct. However, we disagree that significant reductions in questionnaire length cannot be made, whilst preserving all the measurement properties in the OPQ to which Saville & Sik allude. This would be a trivial point to argue except that such reductions can result in many kinds of indirect financial benefits to the test user.

References

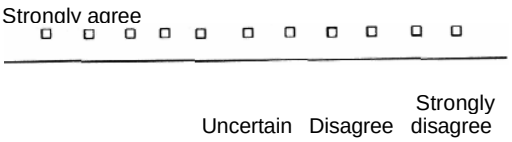
Barrett, P. & Paltiel, L. (1995a). The concept **5.2:** Insufficient measurement **or broad** occupational assessment? *Occupational Psychology Conference Book of Proceedings* (January), 45 - 50.

Barrett, P. & Paltiel, L (19955). Reductio ad absurdum: A reply to Saville & Sik. *Selection and Development Review,* 11 (4), 3-5.
Saville, P. & Sik, G. (1995a). Reductio ad absurdum? *Selection and Development Review,* 11 (3), 1 -3.
Saville, P. & Sik, G. (1995W. Hitting the target: A rejoinder to Barrett & Paltiel. *Selection and Development Review,* 11 (4), 6.
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.

Paul Barrett is at the Department of Psychology, University of Canterbury/ Christchurch, New Zealand.

Laurence Paltiel is with Psytech International Ltd, The Grange, Church Road, Pulloxhiil, Beds MK45 5HE.

Appendix **1: Example POP items**

1. I am at ease in social settings (RG: socially confident)
2. Variety and"change appeal to me {T6: change oriented)
3. I am usually critical of people's ideas (F6: critical)
4. I am modest about my achievements (R7: modest)
5. I enjoy the discussion of hypothetical issues {T7: conceptual)

Appendix *2:* The response scale used for the POP items