Multivariate Experimental Clinical Research, Volume 3, Number 4, 1978, 195-204

RELIABILITY AND SCOPE IN PERSONALITY ASSESSMENT: A COMPARISON OF THE CATTELL AND EYSENCK INVENTORIES

Samuel E. Krug[1]
Institute for Personality and Ability Testing

ABSTRACT

One questionnaire designed to measure primary personality traits (the 16 PF) and another designed to measure second-order traits (the EPI) were compared in terms of reliability and comprehensiveness. Reliability differences between the two tests appeared to be due to their relative length rather than any inherent characteristics of the underlying traits. Two approaches for evaluating the overlap between the two scales were explored. The theoretical conclusion appears to be that the reliable portions of the EPI scale scores can be entirely reconstructed from a knowledge of the 16 PF scores. In contrast, the EPI is far narrower in its scope and able to explain less than a third of the reliable variance in the 16 PF.

INTRODUCTION

In their recent study of the personality structure of British undergraduates Saville and Blinkhorn (1976a) devoted a significant portion of their report to a careful and penetrating analysis of the *16 Personality Factor Questionnaire* (16 PF). As the authors point out, the test, although originally designed for and standardized upon the adult American population, is one of the most widely used inventories in Britain and Ireland. Consequently, a major aim of their research, which encompassed the standardization of Forms A, B, C, and D of the test on a broad, nationally representative sample of British adults and university students, was to provide a technical evaluation of the test utilizing British rather than American data. The total research program, which has continued for more than half a decade, is too extensive to summarize here, and the interested reader will wish to consult the original sources (Saville, 1972, 1973; Saville & Blinkhorn, 1976b, 1976c).

One chapter in their report examines the reliabilities of the 16 PF scales. Although the findings were generally positive, Saville and Blinkhorn voiced concern about the adequacy of certain scales. This issue has important, practical consequences for the 16 PF test user since it bears on the confidence that can be placed on individual test scores. Theoretically, the issue is central to the continuing disagreement between Cattell and Eysenck regarding the precision with which primary traits can be assessed. Eysenck, while not denying the utility of primary traits, has generally argued that such traits lack the replicability and reliability of secondaries or type-level concepts such as extraversion and anxiety (Eysenck & Eysenck, 1969). Cattell, on the other hand, argues that the use of secondaries alone tends to lose important, reliable information contained in the primaries. The question of replicability has been satisfactorily dealt with by Cattell (1973), but some confusion still exists regarding the second issue.

This data set is also particularly valuable because both the 16 PF and the *Eysenck Personality Inventory* (EPI) were administered in their entirety. One chapter of their report is concerned with the overlap between the two inventories, but the principal emphasis is on how much of the 16 PF is predictable from the EPI. How much of the EPI is predictable from 16 PF scores remains unanswered. Consequently, the present paper is devoted to further exploration of these two issues: the comparative reliabilities of primary and secondary trait scales and redundancy between the 16 PF and EPI.

RELIABILITIES OF THE PRIMARY TRAIT SCALES

Reliability and validity are each crucial characteristics of tests. Of the two, reliability or consistency is often the more difficult to estimate properly, mainly because it involves quantities that cannot be directly observed or calculated (e.g., variation in the true score component). According to classical test theory the reliability coefficient is the ratio of true-score to observed-score variance or the proportion of relevant (i.e., non-error) variation in obtained test scores. Since we can't measure true scores, we must proceed by indirect means and under certain assumptions to estimate this variance ratio.

The approach Saville and Blinkhorn have adopted is a well-established one psychometrically: estimating reliability as the correlation between alternate forms of the same test. They note that "alternate forms reliability is indubitably highly relevant to personality questionnaires and their construction." It should be noted that what they call alternate forms reliability is equivalent to what Cattell (1973) has termed *equivalence* and which he has distinguished from *homogeneity* and *dependability*, correctly pointing out that these different varieties of reliability have quite different meanings.

In the case of the 16 PF, however, considering the relatively unique test development and item selection procedures adopted by Cattell, we may reasonably inquire whether equivalence is, in fact, the most appropriate method. In the test handbook and later writings, Cattell states that the principal criterion by which items were selected was their correlation with the pure factor each scale was designed to measure (Cattell, Eber, & Tatsuoka, 1970; Cattell, 1973). The assignment of items to different forms seems entirely based on factor validity. No attempt was made to equate item content across the different forms. On the contrary, such a pursuit would appear to be directly opposed to the principle of "balancing specifics," as Cattell calls it. Perhaps as a result, Cattell himself rarely speaks of the different forms as being parallel. More commonly he urges test users to treat the various forms as a series of extensions. He repeatedly urges test users to administer as many forms as time permits, add the sets of scores together, and work with the resulting composites.

It is true that in many cases parallel-forms estimates of reliability are entirely appropriate. For example, in the case of achievement tests, especially highly speeded ones, if we were to estimate reliability as the correlation between successive presentations of the same form, performance on the retest would undoubtedly be influenced by memory, learning, practice, and other non-error variance scores and seriously bias our estimate of reliability.

However, one immediate consequence of pursuing the "balanced specific" principle is that parallel-forms methods underestimate the true reliability of the test. In Cattell's thinking, a good test for, say, the dominance trait ideally samples dominance behavior across a wide variety of specific contents (e.g., at work, at home, or at a club) and role situations (e.g., as a parent, father, or co-worker). In attempting to sample broadly across the universe of items for a particular common factor, he systematically excludes items which share specifics. While the various test forms may be parallel with respect to their common factor saturation, they are not so with respect to specific factors. More formally, the usual ratio we wish to estimate is:

$$\frac{\sigma_c^2 + \sigma_s^2}{\sigma_c^2 + \sigma_s^2 + \sigma_e^2} \tag{1}$$

where σ_c^2 is the common factor variance, σ_s^2 the specific factor variance, and σ_e^2 the error variance. But the correlation between two test forms constructed according to the principles Cattell espouses may be expressed as

$$\frac{\sigma_c^2}{\sigma_c^2 + \sigma_s^2 + \sigma_e^2} \tag{2}$$

which, except in the improbable case where specifics vanish, must always be less than the quantity given above as (1).

Two questions might reasonably be asked: (1) how successful has Cattell been in pursuing this principle of balanced specifics, and (2) what method more appropriately estimates reliability in the case of scales constructed in this manner?

With respect to the first question, consider that in terms of the common factor model, if two forms of a test share only a single common factor and no specifics, then their intercorrelation may be expressed as the product of their correlation with the factor. That is,

$$r_{xy} = r_{xf}\ r_{yf}$$

where r_{xy} is the interform correlation and r_{xf} and r_{yf} are the respective factor validities. If the two forms are otherwise parallel, then $r_{xf} = r_{yf}$ and $\sqrt{r_{xy}} = r_{xf} = r_{yf}$. That is to say, the square root of the equivalence between the two forms is equal to the factor validity of either form. The first row of Table 1 gives the square root of correlations between corresponding scales on Forms A and B of the 16 PF calculated by Saville and Blinkhorn. Row 2 gives the factor validities for Form A of the test independently calculated on an American sample of undergraduates. Separate samples were deliberately employed to avoid the charge that the two coefficients may be to some extent interrelated in a single sample. As can be seen, the values in row 2 correspond quite nicely with the theoretical predictions in row 1. There is, in fact, a difference of only .09 on the average between the two sets of values, and the congruence between the two vectors is .99.

But if this is the situation and the intercorrelation of two forms underestimates the real reliabilities, then what recourse is there to obtain appropriate values?

Table 1

RELATIONSHIPS BETWEEN 16 PF INTERFORM (A/B) CORRELATIONS AND FACTOR VALIDITIES

Coefficient	A	B	C	E	F	G	H	I	L	M	N	O	Q_1	Q_2	Q_3	Q_4
$\sqrt{r_{AB}}$ [a]	80	60	73	77	84	75	89	77	60	48	51	81	69	69	76	84
r_{Af} [b]	79	35	70	63	83	67	92	70	49	44	41	71	62	70	68	57

[a]Square roots of interform correlations based on 1148 male and female undergraduates (Saville & Blinkhorn, 1976a).

[b]Correlations between scale scores on Form A and pure factors based on 958 male and female American undergraduates. Source: IPAT (1972).

Note: Decimal points have been omitted.

One strategy, entirely consistent with trait theory, is to readminister the same form or combination of forms after a period of time too short for any real personality change to have occurred. In this way specifics are shared across occasions. From this perspective the reliabilities of the primary trait scales are quite excellent and substantially higher than those reported by Saville and Blinkhorn. Comparisons are set out in Table 2. Applying the contemporary reliability standards adopted by Saville and Blinkhorn, it appears that all

Table 2

16 PF RELIABILITIES: COMPARISON OF TEST-RETEST AND PARALLEL FORMS METHODS

Form(s) Used	A	B	C	E	F	G	H	I	L	M	N	O	Q_1	Q_2	Q_3	Q_4
									TEST-RETEST METHODS							
A + B[a]	89	65	87	88	90	88	93	89	87	82	76	89	83	85	78	91
A[b]	86	79	82	83	90	81	92	90	78	75	77	83	82	85	80	72
C + D[c]	82	76	83	77	80	83	86	83	75	68	67	79	75	68	77	82
C[d]	69	61	71	63	67	71	75	71	60	52	50	65	60	52	63	69
									PARALLEL FORMS METHOD							
A with B[e]	64	36	54	59	70	56	80	60	36	23	26	65	47	48	58	70
C with D[f]	53	26	51	46	45	51	52	50	30	30	14	55	39	38	42	51

[a]N = 146 American adults and undergraduates

[b]N = 243 Canadian high school students

[c]N = 150 American undergraduates.

[d]N = 150 American undergraduates. These values were obtained by applying the Spearman-Brown formula to the values on the preceding row assuring a test of half length

[e]N = 1148 British undergraduates

[f]N = 1158 British undergraduates.

Note: Values in rows 1-3 are taken from IPAT (1972). Values in rows 5 and 6 are from Saville and Blinkhorn (1976a).

scales of the 16 PF, even those of Form C containing only six items each, exceed the requisite levels.

What of the EPI? Are scales set up to assess secondaries such as Eysenck's Extraversion (E) and Neuroticism (N) inherently more reliable than those for primaries? The average, reported test-retest coefficient for the EPI scales is about .85 (Lanyon, 1972). However, the EPI E and N scales each contain 24 items, four times the number included in Form C scales. Consequently, if we adjust for length, the expected reliability of a six-item EPI scale is .59, a value exceeded by all but three of the 16 PF Form C primary scales.

The conclusion would seem to be that secondaries are not inherently more reliable. In practice, they frequently appear to be more reliable, but only because the tests themselves are longer. For example, one test for the measurement of depression, which has been identified and replicated as a second-order factor among a set of normal and clinical primaries, has been shown to have an average reliability of .93 across various estimation methods and samples (Krug & Laughlin, 1976, 1977). However, this scale contains 36 items, and the higher reliability must necessarily be attributed to the greater length. This is, in fact, almost exactly the value one would expect to achieve if the typical 16 PF scale were 36 items in length.

The degree of correspondence between certain scales of various forms of the 16 PF has one practical corollary for test users involved in sequential experiments requiring retesting: use the same form or combination of forms at both points in time. Saville and Blinkhorn perceptively make this their conclusion, but the point bears repeating. With the exception of Factor B (intelligence), it is difficult to see how memory of specific item content will prove to have serious impact on the results. If the interval is quite short—a day or less—it may be appropriate and helpful to modify the instructions just enough to remind examinees not to be influenced by their earlier answers. However, in reality it seems unlikely that the test, which has been shown to measure relatively stable response patterns, would actually be used to measure such transitory effects. In such situations, tests specifically designed to measure psychological mood states are more likely to be appropriate.

The decision to use a test that targets on primaries rather than on secondaries must therefore be made on grounds other than the inherent reliability of the latter. An obvious criterion is, and always has been, the scope or comprehensiveness of the test, and this leads rather directly into the second major issue of this paper.

RELATIONSHIP BETWEEN THE CATTELL AND EYSENCK INVENTORIES

Although the principal focus of the Saville and Blinkhorn study was on the 16 PF itself, and a full investigation of the relationships between the 16 PF and EPI would have been somewhat tangential, the question is still intriguing.

In their investigation Saville and Blinkhorn sought to determine how much of the reliable variance in the 16 PF scales was not redundant on the EPI scales. Their emphasis on reliable variance is important because one test may appear to be totally independent of a second, but be so unreliable as to make it useless. Its unreliability may, in fact, be the source of the apparent

independence since a test's validity may not exceed the square root of its reliability. There is, however, a paradox in their presentation. On the basis of their Table 5.6 one would be led to the conclusion that less than 10% of the reliable variance in the typical 16 PF primary scale is not predictable from the EPI. In a later chapter they report that extraction of the Eysenck factors from the 16 PF intercorrelations only accounts for about 20-23% of the reliable variance among the full set of scales. This latter finding undoubtedly led them to conclude that "in comparison with the Eysenck Personality Inventory, Cattell's 16 PF covers more ground" (p. 163). Still, there is quite a bit of difference between 90% redundancy by the first method and 20% redundancy by the second, and a comparison of the two methods is in order.

A careful examination of the method by which Saville and Blinkhorn arrived at the first set of values suggests that these values are, indeed, quite conservative. Their procedure was as follows. First, the 16 PF-EPI intercorrelations were corrected for attenuation (i.e., for unreliability of the component scales). Next, each 16 PF scale was predicted from the optimum linear combination of the EPI scales. The resulting multiple correlations were squared (to obtain the proportion of 16 PF variance predictable from the EPI) and subtracted from unity to obtain the unshared^2 variance proportion. Since all unshared variance is not necessarily reliable, they adjusted by multiplying each of the values obtained thus far by the square of the corresponding interform correlation.

The first thing to keep in mind is that correlations corrected for attenuation no longer reflect interscale relationships directly but instead estimate what the relationships would be if the two scores were uncontaminated by measurement error. Observed scores may have seemingly little overlap, while at the same time the underlying traits are highly related. However, strictly speaking, once the correlations are corrected for attenuation we are no longer talking about the tests as they exist, but as they might exist. Further, working with such "idealized" matrices can generate computational problems since the corrections may destroy the positive definiteness of the correlation matrix, thereby leading to inconsistent results. In several cases it is obvious that their multiple correlations, which can theoretically only range between 0 and 1, have exceeded this upper limit, leaving us with negative variance estimates.

Secondly, note the reliability estimates used in their final correction were interform correlations which, as Table 2 and the earlier discussion point out, may seriously underestimate the true reliabilities of the scales. Finally, since the reliability coefficient by definition already represents the systematic (i.e., non-error) proportion of the total test variance, the appropriate correction is to subtract the *squared* multiple *R* from the reliability coefficient. In effect then, the values they report are adjusted for reliability three times, and the reliabilities used for adjustment are probably conservative.

When these values are recalculated without attentuation corrections and with appropriate reliabilities, the results are as shown in Table 3. These figures agree with Saville and Blinkhorn's in showing, for example, the 16 PF anxiety primaries to be most redundant on Eysenck's scale, while Factors B (intelligence), G (conscientiousness), and I (sensitivity) have the largest proportion of unshared, reliable variance. The most obvious difference is in the

magnitude of the values. While Saville and Blinkhorn suggest that less than a tenth of the reliable variance associated with the typical 16 PF scale is independent of the EPI, the values in Table 2 suggest that more than half of the variation in the 16 PF scales is reliable and unshared.

Table 3

PROPORTIONS OF 16 PF SCALE SCORE VARIANCE (Forms C & D)[a] THAT IS BOTH RELIABLE AND DISTINCT FROM THAT CONTAINED IN THE EPI SCALES

A	B	C	E	F	G	H	I	L	M	N	O	Q_1	Q_2	Q_3	Q_4
.64	.75	.36	.61	.34	.73	.41	.82	.68	.67	.55	.22	.70	.53	.58	.33

[a]16 PF-EPI intercorrelations were not available for Forms A and B. The values given here are reasonable estimates of values for Forms A and B and may, in fact, be lower-bound estimates.

To complete the picture we must, of course, ask how much of the reliable variation in EPI Factors E (Extraversion) and N (Neuroticism) is unpredictable from 16 PF scores. Results are .19 for men and .21 for women with regard to extraversion and .03 for men and .18 for women with regard to neuroticism (anxiety). That is, about a tenth of the variation in the Eysenck scales is reliable and independent of the 16 PF.

The figures in Table 3 are in better agreement with the results from Saville and Blinkhorn's second approach, which was multivariate in nature. Using a method for extracting arbitrary, orthogonal factors (Cooley & Lohnes, 1971) from the intercorrelations among the 16 PF and EPI scales, they first extracted a factor which was precisely collinear with one of the Eysenck scales and a second component which was targeted on the remaining EPI scale. The reason that the second factor is not precisely collinear with the second scale is that this method of diagonalizing a matrix requires successive components to be orthogonal within the total test space. Even though E and N correlate only —.08, the second factor must be orthogonal to the first factor across the 16 PF scales as well. This frequently results in a factor which blends extraversion and anxiety in a concession to the mathematics of the method and, as a result, may not be the most appropriate method for evaluating overall relationships between the two domains of psychological variables. Nevertheless, when these two components were extracted, Saville and Blinkhorn found that at least four or five more factors were necessary to account for covariation among 16 PF primaries. This number is in reasonably good agreement with Cattell, who has steadily argued that no less than eight secondaries are required to account for the covariation among the 16 primaries (1973).

The canonical correlation model is particularly appropriate for examining the interrelationships between two variable domains and avoids certain technical problems Saville and Blinkhorn encountered. When used along with the "redundancy index" developed by Stewart and Love (1968), it becomes even more powerful and still avoids theoretical controversies regarding methods of extraction and rotational techniques if the joint matrix were to be fully factor analyzed. The redundancy indices provide overall measures of how much variance in one set of variables is accounted for by the other set.

The results from canonical analyses of these data are shown in Table 4.

Table 4

CANONICAL FACTOR STRUCTURE MATRICES

	Uncorrected				*Corrected*			
16 PF	**Males**		**Females**		**Males**		**Females**	
Trait	**I**	**II**	**I**	**II**	**I**	**II**	**I**	**II**
A	21	46	41	36	13	53	19	51
B	−06	−18	−09	−02	−03	−19	−06	−06
C	73	−16	73	−41	86	00	88	02
E	16	53	23	40	07	60	03	48
F	52	68	65	52	39	81	31	77
G	−14	−37	−21	−34	−10	−53	−04	−50
H	54	52	71	31	43	65	45	62
I	−09	−17	−02	00	−09	−28	−02	−01
L	−22	26	−15	38	−30	26	−31	29
M	−06	−14	−15	02	−04	−16	−13	−06
N	−38	−18	−33	−05	−37	−30	−26	−23
O	−80	15	−77	36	−88	−02	−85	−08
Q_1	09	29	09	29	05	36	−05	34
Q_2	−06	−52	−34	−23	03	−58	−19	−38
Q_3	27	−46	16	−58	37	−47	40	−47
Q_4	−72	36	−56	59	−78	23	−72	24
EPI Trait								
E	39	92	65	76	22	98	22	98
N	−95	31	−81	58	−99	13	−99	13
Canonical Correlations								
	.92	.79	.85	.77	1.00	.87	1.00	.91

Note: Decimal points have been omitted when reporting structure values.

All analyses were conducted separately for men and women in order to maintain consistency with the earlier work. First, the correlation matrices provided by Saville and Blinkhorn were analyzed as is. All canonical coefficients were statistically significant far beyond the usual levels. The first four columns of Table 4 give the resulting canonical factor structure values (i.e., correlations between the scales and the canonical variates). The canonical correlations are substantial and suggest considerable overlap between the two tests. The redundancy indices, however, indicate that the overlap is decidedly one-sided: while about 70% of the information contained in the EPI scales is already contained in the 16 PF (74% for men, 66% for women), only 22% of the information contained in the 16 PF is also contained in the EPI (23% for men, 21% for women).

What of the unexplained 30% of the variance in the Eysenck scales? Two possibilities exist: either this represents non-systematic (i.e., error) variance and thereby does not permit prediction or it is reliable variance of a restricted psychological nature that does not enter the broad personality sphere on which the 16 PF was thoroughly grounded.

To explore this point, a second set of canonical analyses were carried out. This time the correlation matrices were first corrected for attenuation in order to represent relationships among the traits as they might be if there were no errors of measurement. Once again all coefficients were statistically significant and the structure values are given as the last four columns of Table 4. Again, the redundancy indices are informative, this time showing that while 32% of the reliable variance in the 16 PF is shared with the EPI, 99% of the reliable variance in the EPI is shared with the 16 PF.

Using Gleason's (1976) recently demonstrated interpretation of the redundancy index, it would appear, theoretically at least, that the EPI scales can be entirely reconstructed from knowledge of the 16 PF trait scores. The EPI is far narrower in its scope and able to explain less than a third of the reliable variance in the 16 PF. Furthermore, it seems quite improbable that the addition of a third factor, psychoticism, to E and N will close the gap, since work by Cattell and Bolton (1969) and Krug and Laughlin (1977) has shown that psychoticism is relatively independent of the normal personality traits tapped by the 16 PF. Instead, it keeps company with a set of 12 specialized traits covering depression and other clinical manifestations and which are required, along with the 16 PF source traits, to account adequately for pathology (Delhees & Cattell, 1971). A trinity of traits may be theologically orthodox. However, such a position appears quite unable to explain the complexities of normal human personality functioning.

NOTES

1. Requests for reprints should be sent to Samuel E. Krug, Test Services Division, Institute for Personality and Ability Testing, 1602 Coronado Drive, Champaign, Illinois 61820.

2. Saville and Blinkhorn refer to "unique, reliable variance." However, since the discussion revolves around factor-analytically developed tests, the term "unique" may be somewhat confusing. The term "unshared" avoids this confusion.

REFERENCES

1. Cattell, R. B. *Personality and mood by questionnaire.* San Francisco: Jossey-Bass, 1973.

2. Cattell, R. B., & Bolton, L. S. What pathological dimensions lie beyond the normal dimensions of the 16 PF? A comparison of the MMPI and 16 PF factor domains. *Journal of Consulting and Clinical Psychology,* 1969, *33,* 18-29.

3. Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. *Handbook for the 16 PF.* Champaign, Ill.: IPAT, 1970.

4. Cooley, W. W., & Lohnes, P. R. *Multivariate data analysis.* New York: Wiley, 1971.

5. Delhees, K. H., & Cattell, R. B. *Handbook for the Clinical Analysis Questionnaire.* Champaign, Ill.: IPAT, 1971.

6. Eysenck, H. J., & Eysenck, S. B. G. *Personality structure and measurement.* San Diego: EDITS, 1969.

7. Gleason, T. C. On redundancy in canonical analysis. *Psychological Bulletin,* 1976, *83,* 1004-1006.

8. IPAT Staff. *Manual for the 16 PF.* Champaign, Ill.: IPAT, 1972.

9. Krug, S. E., & Laughlin, J. E. *Handbook for the IPAT Depression Scale.* Champaign, Ill.: IPAT, 1976.

10. Krug, S. E., & Laughlin, J. E. Second-order factors among normal and pathological primary personality traits. *Journal of Consulting and Clinical Psychology,* 1977, *45,* 575-582.

11. Lanyon, R. I. Review of the *Eysenck Personality Inventory* in O. K. Buros (Ed.), *The seventh mental measurement yearbook.* Highland Park, NJ: Gryphon Press, 1972, pp. 163-166.

12. Saville, P. *The British standardization of the 16 PF: supplement of norms, Forms A and B.* Windsor: NFER, 1972.

13. Saville, P. The standardization of an adult personality inventory on the British population. *Bulletin of the British Psychological Society,* 1973, *26,* 25-29.

14. Saville, P., & Blinkhorn, S. *Undergraduate personality by factored scales.* Windsor: NFER, 1976. (*a*)

15. Saville, P., & Blinkhorn, S. *Undergraduate norms to the 16 PF, Forms A and B.* Windsor: NFER, 1976. (*b*)

16. Saville, P., & Blinkhorn, S. *Undergraduate norms to the 16 PF, Forms C and D.* Windsor: NFER, 1976. (*c*)

17. Stewart, D. K., & Love, W. A. A general canonical correlation index. *Psychological Bulletin,* 1968, *70,* 160-163.