Journal of Occupational Psychology, 1988, **61**, 153–162. Printed in Great Britain

Spuriouser and spuriouser: The use of ipsative personality tests

Charles E. Johnson
Psychometric Research & Development (PRD) Ltd, 36–38 London Road, St Albans AL1 1NG, UK

Robert Wood
PRD Ltd and University of London Institute of Education

S. F. Blinkhorn
PRD Ltd and Hatfield Polytechnic

One of the more worrying features of recruitment and selection practices in the United Kingdom is the misuse of ipsative personality tests. Employers are understandably attracted by claims that these quick and easy to administer tests will give valid insights into the personality of job applicants. However, on the evidence we have seen, the publishers and the promoters of these tests are either unaware of, or do not understand, or are choosing to ignore their limitations. This is not to say that ipsative tests have no utility but that the claims made for their validity and reliability and their applicability to inter-individual comparisons are misleading. Failure to take account of the mathematical properties of ipsative measures leads users to treat them as if they are normative measures, with startling consequences which ought to be obvious but unfortunately are not.

Ipsative personality tests are commonly encountered in recruitment and selection practice. Examples of such tests are:

Kostick Perception and Preference Inventory (PAPI) (Kostick & Coules, 1980)
Personal Profile Analysis (PPA) (Hendrickson and Associates, Inc., 1981)
PAL Personality Profile System (PPS) (Cleaver, revised Oddy, 1985)
Cleaver (Donelly, Mahan & McManus, 1965)
Performax Personal Profile System (Geier, 1979)
Concept Four and Octagon Two forms of the Occupational Personality Questionnaires (OPQ) (Saville & Holdsworth Ltd, 1984).

All these tests are self-appraisal personality inventories that are claimed to be useful in selection, counselling and training. The PAPI uses a forced-choice format. Statements exemplifying particular scales are paired with statements exemplifying different scales and the task is to choose the statement which best describes your personality. There are 20 scales in all, covering the concepts work direction, leadership, activity,

social nature, work style, temperament and 'followership'. All the scales are supposed to be specific to the work situation. These scales are subdivided into 10 role scales and 10 need scales.

The PPA, PPS, Cleaver and Performax all belong to a family of tests which it is claimed are based on the work of Marston (1928). They are all variants of the DiSC system which measures four personality traits. The letters in the acronym stand for the traits dominance, influence or inducement, steadiness or submission and compliance. Although there is some variation in the trait names, the adjectives used as items in the various inventories are almost identical. The PPS and Cleaver have the same manual, while the PPA uses this manual with a few wording changes. All the tests are administered in the same way using a forced-choice format whereby respondents have to choose from four adjectives the one which is most like them and the one which is least like them. The adjectives are exemplars of the four different traits, although in some of the tests a few adjectives score on more than one trait. The most like, least like and total scores are often interpreted separately.

The OPQ is actually a suite of nine tests which vary in the number of scales measured (from six up to 30 plus a social desirability scale) and in item formats (multiple choice, forced-choice and ranking). At the most general level, the 30 scales are grouped into those dealing with ways of relating to others, those dealing with style of thinking and those dealing with feelings. Like the PAPI, the scales are all supposed to be work-related. The 30 scale versions are the basic versions from which others are derived, the basic scales being combined into superordinate scales in accordance with the results of factor analyses. Only two of the test versions are ipsative. The Concept Four version has 30 scales (plus the social desirability scale) and involves ranking four statements which are each exemplars of different scales. The Octagon Two version has eight scales and involves a forced choice between two statements which are each exemplars of different scales.

The problems with ipsative tests are, in fact, well documented, but an informal survey of textbooks of psychometrics—which are, of course, mostly American and concentrate on achievement and ability tests—shows that, with the exception of Guilford (1954), the topic is largely ignored. Where it is covered, as in Anastasi (1976), the coverage tends to be thin. The facts concerning ipsative tests are, however, uncontroversial and show that:

1. they cannot be used for comparing individuals on a scale by scale basis;
2. correlations amongst ipsative scales cannot legitimately be factor analysed in the usual way;
3. reliabilities of ipsative tests overestimate, sometimes severely, the actual reliability of the scales: in fact, the whole idea of error is problematical;
4. for the same reason, and others, validities of ipsative tests overestimate their utility;
5. means, standard deviations and correlations derived from ipsative test scales are not independent and cannot be interpreted and further utilized in the usual way.

What is an ipsative measure?

The strongest ipsative property is that scale scores for an individual always add to the same total. If on an achievement test scores consisting of number right, number

wrong and number omitted are assigned to individuals, then the sum of the three scores will always be the same, i.e. equal to the number of items in the test.

The sort of tests which typically result in ipsative measurement are those using forced choices between scales, or the ranking of scales. Not all scoring systems are fully ipsative. Hicks (1970) lists seven possibilities for reducing ipsativity:

1. respondents only partially order item alternatives, rather than ordering them completely;
2. scales have differing numbers of items;
3. not all alternatives ranked by respondents are scored;
4. scales are scored differently for respondents with different characteristics, or are referred to different normative transformations on the basis of respondent characteristics;
5. scored alternatives are differentially weighted;
6. one or more of the scales from the ipsative predictor set is deleted when data are analysed;
7. the test contains normative sections.

Tests to which such procedures are applied, however, still retain most of their ipsative properties, with attendant (and largely predictable) consequences, as will be seen.

Why can't you treat ipsative measures as if they were normative measures?

In his book *Psychometric Methods*, J. P. Guilford writes as follows about ipsative measures:

> The ordinary factor analysis by what is known as the R technique calls for normative measurements. It would be wrong to use *ipsative* measurements in the intercorrelation of experimental variables. Ipsative measurements for each individual are distributed about the mean of that individual, not about the population mean. Individual differences in ipsative measurements have little meaning because there is not a single scale for all individuals. Ipsative scores arise when traits for an individual are ranked for that individual, directly or by some other procedure such as pair comparisons. The forced-choice type of item, in which we have something approaching pair comparisons, often gives scores with strong ipsative properties. They should not be used for correlations of variables over a population of individuals. Ipsative scores are the appropriate ones to use when we intercorrelate persons. Such intercorrelations are appropriate for application of the Q technique in factor analysis. (Guilford, 1954, p. 528).

In factor analysis, there should be no reason for relationship except the existence of common factors. Spurious correlations arise when correlated variables are not experimentally independent. That is precisely what happens when ipsative measures are prepared for a conventional R factor analysis. The immediate consequence of making up different scores from the same scales is a sharing of specific variance across scores, thus distorting immediately estimates of reliability and validity. With spurious correlations present among the scales, ordinary factor analysis breaks down, actually in a most spectacular way with the built-in dependencies among scores producing

mixtures of positive and negative correlations resulting in degenerate and illegal solutions. (Technically, the matrix is singular and therefore has no regular inverse. In order to produce an inverse, it is necessary to delete a variable Clemans, 1966).

Writing out the basic R factor analysis model (also the classical true score test theory model) should make it clear why the estimation methods suitable for that method cannot be used for ipsative measures.

Suppose there are $m < p$ factors $f_1, f_2, \ldots, f_m$, then the score X_{ij} of any particular individual j on scale i can be written as

$$X_{ij} = \sum_{k=1}^{m} \lambda_{ik} f_k + \varepsilon_{ij},$$

where the λ_{ik} are the factor loadings and the ε_{ij} are the specific factors or residuals. These ε_{ij} are assumed to be independent of all other ε and of the f_k. This immediately rules out ipsative measures because the constraint on the X_{ij} always adding to the same constant means that whatever errors are present must be correlated. As ipsativity diminishes, the effect moderates, but is there all the same.

Test theory, like the common factor model, supposes that there is a degree of random error in all test scores. The purpose of esimating the reliability of a test is to quantify this error. Ipsative tests, by definition and by necessity in their construction, have no random error component as such. If there are k scales, then the score on any scale can be perfectly predicted from the other $k-1$ scales. Since retest, alternate form and internal consistency estimators of reliability share a common theoretical justification, and this justification does not apply to ipsative tests, presentation of reliability coefficients calculated in the ordinary way using data from ipsative tests is merely an exercise in arithmetic with no significance for practice. What can be said is that any coefficients so calculated are not independent of each other, nor are they independent (in a mathematical sense) of the means and variance of the scales to which they refer, contrary to what is the case with ordinary tests.

Ipsative measures need to be modelled quite differently from normative measures. The point of departure would be the recognition earlier that the ipsative test is effectively a variant on the paired-comparisons technique, in which case the relevant methodology applies (e.g. David, 1963). The result of scaling, which is what is entailed, is a point estimate for each person on some notional scale. If the ipsative test data cannot support scaling on one dimension alone, then multi-dimensional scaling comes into play. Guilford's suggestion of Q factor analysis is not the answer, as at first appears, simply because the continued presence of dependencies in the data matrix corrupts estimation procedures just as it does for R factor analysis.

Consequences of treating ipsative measures as normative

More specific predictions of the technical consequences of using ipsative measures as if they were normative can be made through mathematical derivations. The following results are due to Clemans (1966):

1. the sums of the columns, or rows, of an ipsative covariance matrix must equal zero;
2. the sums of the columns, or rows, of an ipsative intercorrelation matrix will equal zero if the ipsative variances are equal;

3. the average intercorrelations of ipsative variables have $-1/(m-1)$ as a limiting value where m is the number of variables;
4. the sum of the covariances obtained between a criterion and a set of ipsative scores equals zero;
5. the sum of ipsative validity coefficients will equal zero if the ipsative variances are equal.

If any reassurance is needed that mathematical results actually work out in practice, empirical verification is presented in the next section.

Illustrative data sets

Of the tests mentioned at the start of this paper, some published data are available for the PAPI, OPQ, the PPS and the PPA. We have also had access to data for the Cleaver and the PPA. Of these, the PAPI and the ipsative versions of the OPQ are examples of purely ipsative measurement. The Cleaver, PPS and PPA are examples of partially ipsative measurement. For example, the Cleaver, although it uses a forced-choice format, incorporates features 2 and 5 from Hicks' list.

The first of the mathematical peculiarities of ipsative tests is that average intercorrelations of their scales can be predicted from the number of scales using the simple formula shown above. Actual average intercorrelations sometimes differ from those predicted if the scale variances are unequal but the average is always negative with the predicted correlation as a limit. Table 1 shows average intercorrelations for the above tests. Note that these tests vary in degree of ipsativity yet the outcomes are much the same.

Table 1. Predicted and actual scale intercorrelations

Test	Predicted	Actual	Source	*n*
PAPI (roles)	−0.11	−0.11	PAPI technical manual	300
PAPI (needs)	−0.11	−0.09	PAPI technical manual	300
Cleaver	−0.33	−0.33	HAY–MSL	206
PPS	−0.33	−0.29	Paltiel (1986)	452
PPA	−0.33	−0.31	Saville & Holdsworth Ltd	222
PPA	−0.33	−0.28	Paltiel (1986)	467
PPA	−0.33	+0.04	PPA technical manual	100
Concept Four (OPQ)	−0.03	−0.03	Saville & Holdsworth Ltd	440

In every case except one the actual average intercorrelation is very close to the predicted average. The technical manual for the PPA has, however, two matrices of scale intercorrelations where the average correlation is positive. This is mathematically impossible so users should be aware that these matrices either contain errors or have been transformed in some way. In fact, the point is made in the manual that in further field trials much larger negative correlations were found between some of the scales, which is borne out by the other data sources for the PPA.

Some of these tests have more than one scoring system whereby statements which are perceived as being most like the respondent and those which are perceived as least like are summed separately. However, as Table 2 shows, the scales derived from these scoring systems still show the symptoms of ipsative measurement.

Table 2. Average scale intercorrelations for most and least scores

Test	Predicted	Actual	Source
Cleaver (least)	−0.33	−0.30	HAY-MSL
Cleaver (most)	−0.33	−0.29	HAY-MSL

There are several implications of this aspect of ipsative measurement. One is that it is highly unlikely that any really large positive correlations will be found between scales, even when they are measuring practically the same characteristic. Another is that, because the correlations between scales tend towards zero as the number of scales increases, an illusion of scale independence is created.

The interdependencies of ipsative scales can be well seen in the way items appear to weight on scales. Table 3 shows biserial correlations between the first 12 descriptors in the Cleaver and the four scales. It also shows the scales on which they are supposed to score.

Table 3. Sample item–scale correlations for the Cleaver [source: Hay-MSL ($n = 206$)]

Descriptor	Target scale	D	i	S	C
1	i	0.32	0.29	−0.39	−0.22
2	S	−0.71	−0.19	0.72	0.43
3	C	−0.47	−0.27	0.42	0.42
4	D	−0.54	−0.32	0.50	0.59
5	D	0.63	−0.29	−0.38	−0.24
6	i	−0.32	0.69	0.00	−0.16
7	S	−0.49	−0.35	0.52	0.46
8	C	0.26	0.28	−0.22	−0.43
9	S	0.45	0.38	−0.52	−0.45
10	C	−0.57	−0.01	0.41	0.44
11	D	0.61	−0.05	−0.42	−0.42
12	i	0.03	0.26	−0.12	−0.20

The significance of the degree of interdependence apparent in these results is its effect on reliability. Any consistency within one scale automatically creates consistency in some or all of the other scales with the net result that all reliability coefficients, particularly those estimating internal consistency, will be inflated.

The interdependent nature of ipsative scales has other consequences which affect multivariate analyses. If the test is purely ipsative, every scale score can be exactly predicted by a combination of the other scales, which means there is always at least one degree of freedom less than there should be. In some cases, depending on how the respondent approaches the task, there may be even fewer degrees of freedom.

The effects on and the consequences for factor analysis are fascinating, not least because correlation matrices derived from ipsative measures would appear to be acceptable for R factor analysis on the typical tests for assessing psychometric adequacy. For example, the Bartlett test of sphericity can be used for checking that variables are sufficiently related to justify factor analysis. The statistic is approximately distributed as chi square and, for example, with the role of the PAPI, a chi square value greater than 80.51 would be significant at the 0.001 level. The test actually gives a value of 5087 for the PAPI role scales which is so high it ought to alert the researcher to peculiarities in the data.

The peculiarities in fact show up unmistakably when the uniquenesses of the scales and the residual matrices are examined after factor analysis. Table 4 shows these for (a) Cleaver 'most like' scores, and (b) the PAPI. The program used fits least squares and maximum likelihood solutions; the number of factors in each case was chosen on the basis of the Kaiser–Guttman criterion.

Table 4. Uniquenesses and residual matrices from two-factor analyses

Scale	Uniqueness	Residual matrix									
(*a*) Cleaver ' most like ' scores (source: Hay–MSL, $n=206$)—two-factor solution											
				D		i		S		C	
D	−0.0005			0.00		0.00		0.00		0.00	
i	0.0004			0.00		0.00		0.00		0.00	
S	0.3482			0.00		0.00		0.35		−0.33	
C	−0.7513			0.00		0.00		−0.33		0.75	
(*b*) PAPI–roles (source: PAPI technical manual, $n=560$)—five-factor solution											
		G	L	I	T	V	S	R	D	C	E
G	0.9586	0.92	0.00	−0.23	0.00	−0.08	0.00	−0.18	0.00	0.00	−0.11
L	−0.0004	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.9172	−0.23	0.00	0.84	0.00	−0.10	0.00	−0.17	0.00	0.00	−0.21
T	−0.0003	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.9010	−0.08	0.00	−0.10	0.00	0.81	0.00	−0.15	0.00	0.00	−0.21
S	−0.0001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.9447	−0.18	0.00	−0.17	0.00	−0.15	0.00	0.89	0.00	0.00	−0.08
D	−0.0001	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C	−0.0003	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.8677	−0.11	0.00	−0.21	0.00	−0.21	0.00	−0.08	0.00	0.00	0.75

These data sets are typical of the analyses we have carried out on the ipsative tests to which we have access. There are three common features which stand out. First, many of the scales appear to have zero or even slightly negative uniqueness. The notion of negative variance is of course nonsensical, and these may be just the result of

rounding errors. A zero uniqueness implies exactly that: all the variance attributable to such a scale is shared with other scales. These findings are examples of Heywood cases, and thus indicate improper factor solutions. The second feature is that the residual matrices include many zeros which implies that there is no measurement error on some of the scales. This is highly unlikely. The third feature is that neither increasing the number of scales (as in the PAPI) nor using partial scores (as in the Cleaver ' most like ' scores) improves the situation.

A final consequence of ipsative measurement, but one which is crucial, is that the sum of the covariances between pure ipsative scales and any criterion always equals zero, and the sum of the correlations tends to zero. If a test is partially ipsative or if the scale variances are unequal, there may be some departure from zero but typically not much. Table 5 shows correlations between the Cleaver and second order factors from Cattell's 16PF.

Table 5. Correlations between the Cleaver and the 16PF (Source: Hay–MSL)

16PF	D	i	S	C	Total
Exvia	0.23	0.70	−0.43	−0.49	0.01
Anxiety	−0.14	−0.10	0.17	0.07	0.00
Cortertia	0.06	−0.05	0.05	−0.12	−0.06
Independence	0.48	0.16	−0.42	−0.34	−0.12
Discreteness	−0.20	−0.33	0.16	0.38	0.01
Prodigal subjectivity	0.01	0.09	−0.03	−0.05	0.02
Superego strength	−0.15	−0.15	0.13	0.23	0.06
Neuroticism	−0.24	−0.26	0.27	0.26	0.03
Leadership	−0.08	0.32	−0.07	−0.09	0.08
Creativity	0.19	−0.09	−0.13	−0.08	−0.11
School achievement	0.24	−0.21	−0.14	0.05	−0.06

Similar results can be seen in the correlations between the PAPI scales and the Eysenck Personality Inventory (Johnson, 1986) but generally the technical manuals for the various tests do not show the complete matrices of validity coefficients. The most telling example can be found by close inspection of a table in Paltiel (1986). He presents a matrix of correlations between the PPS and OPQ Octagon Two. Since both tests are ipsative, to varying degrees, we would expect the correlations in both the rows *and* the columns of the matrix to sum to approximately zero, and this is precisely what happens. The import of this feature of ipsative tests is that any or all of the apparently significant correlations between scales may be artifacts and all are, in any case, dubious.

Other interesting validity phenomena can be observed. For example, Johnson (1986) has identified an interesting paradox whereby there are six PAPI scales where the same trend is visible across three experimental groups varying in business experience yet half the scales are negatively correlated with the other half. This happens because the scale means and the scale correlations are also interdependent. The

message is, however, clear. The standard statistics used in the evaluation of tests are not appropriate with ipsative tests. In some cases the authors of these inventories have made the claim that they are not tests. This is a perfectly reasonable claim to make but they should not then present standard statistics in an attempt to give the inventory credibility.

In summary, standard statistical techniques cannot be used with data from ipsative tests. Correlations, of any sort, between ipsative scales are uninterpretable because the scales are mathematically interdependent. Similarly, correlations between ipsative scales and external criteria are uninterpretable. It follows, therefore, that any method which relies on the analysis of correlation matrices is also inadmissible. Therefore partial correlations, multiple correlations, multiple regressions, reliability coefficients, discriminant analyses, cluster analyses and, as we have seen, factor analyses will produce results which are at odds with statistical and test theory and thus are very misleading. Only by deleting a variable or variables can strict dependencies be removed but it will be appreciated that the variables left still have shared specific variance etc., and so problems of interpretation remain. Finally, because of their mathematical properties, scale means are also uninterpretable and tests of differences between group means such as *t* tests and analysis of variance are meaningless.

What are ipsative measures good for?

Guilford maintains that ipsative measures are the appropriate ones to use when correlating persons and when doing Q factor analysis. The idea of correlating persons is odd at first but makes good sense when it is realized that experimenters are more likely to have many data on a few persons than few data on many persons. What a Q factor analysis can bring out, in principle, is evidence of personality types or syndromes. Persons having outstanding combinations of traits in common will show these as factors (Guilford, 1954, p. 529). We stress ' in principle ' because it is not clear to us how any factor analysis can go through when the same dependencies remain in the data. Although it is now people who are being correlated, rather than variables, the same mix of positive and negative correlations can be expected to occur, and this is indeed what happens. Q factor analysis remains a possibility, but with normative or very weakly ipsative measures.

It is certainly permissible to compare individuals in terms of score profiles or patterns, because then the absence of a common metric does not matter. Obviously this is only non-trivial once the number of variables equals three or more. When the number is more than three, interpretation becomes difficult, but when it is exactly three there is a neat way to plot the trivariate scores using triangular graph paper and so-called barycentric coordinates. Wood (1976) has an example using number right, number wrong and number omitted.

There are psychological benefits to be reaped from the use of the forced-choice technique but in measurement terms these are compromised by the practice of concocting different kinds of scores from the several scales and therefore building in dependencies. There may be psychological merit in such scoring systems but fresh ways will have to be found of demonstrating it. Manipulating ipsative measures as if they were normative measures is an exercise in futility, like cheating at patience.

References

Anastasi, A. (1976). *Psychological Testing*, 4th ed. New York: Macmillan.

Clemans, W. V. (1966). An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, **14.**

David, H. A. (1963). *The Method of Paired Comparisons*. London: Griffin.

Donelly, N., Mahan, T., Jnr & McManus, L., Jnr (1965). *Self-description: A Technical Manual*. J. P. Cleaver.

Geier, J. G. (1979). *The Personal Profile System Manual*. Performax Systems International. Princeton, NJ: J. P. Cleaver & Co. Inc.

Guilford, J. P. (1954). *Psychometric Methods*, 2nd ed. New York: McGraw-Hill.

Hendrickson, T. M. and Associates Inc. (1981). *Personal Profile Analysis: A Technical Manual*. Thomas International Management Systems.

Hicks, L. E. (1970). Some properties of ipsative, normative and forced-choice normative measures. *Psychological Bulletin*, **74**, 167–184.

Johnson, C. E. (1986). A critique of the Perception and Preference Inventory. *Guidance and Assessment Review*, **2**, (6), 3–7.

Kostick, M. M. & Coules, J. (1980). *Kostick's Perception and Preference Inventory (PAPI) Manual*. PA International.

Marston, W. M. (1928). *Emotions of Normal People*. New York: Harcourt, Brace.

Oddy, K. (1985). *PAL Personality Profile System (PPS) Manual*. Birmingham: Oddy.

Paltiel, L. (1986). Self-appraisal personality inventories. *Guidance and Assessment Review*, **2**(3), 3–7.

Saville & Holdsworth Ltd. (1984). *Occupational Personality Questionnaires Manual*. Esher: Saville & Holdsworth Ltd.

Wood, R. (1976). Inhibiting blind guessing: The effect of instructions. *Journal of Educational Measurement*, **13**, 297–307.

Received 14 December 1986; revised version received 10 April 1987