The British Psychological Society

OCCUPATIONAL PSYCHOLOGY CONFERENCE

3–5 January 1996
The Grand Hotel Eastbourne

BOOK OF PROCEEDINGS

The British Psychological Society

Division & Section of Occupational Psychology

The Psychometric and Practical Implications of the use of Ipsative, forced-choice format, Questionnaires

Authors:

Sean Hammond
University of Surrey
Department of Psychology
Guildford
Surrey
United Kingdom
Fax: 01483-32813

Paul Barrett
University of Canterbury
Department of Psychology
Private Bag 4800
Christchurch
New Zealand
Fax: 01064-3-364-2181

Summary

The consequences of interpreting psychometric tests that are constructed using ipsative, forced-choice responses, are examined with regard to the distortion of psychological measurement caused by such tests. Using Gordon's Survey of Interpersonal Values (SIV: both normative and ipsative versions), the HBS5 (Hammond-Barrett-Savage 5 factor model normative and ipsative adjective checklist), the OPP (Psytech's Occupational Personality Profile), and some computer-simulated datasets, it was demonstrated that:

1. Ipsative scores cannot be interpreted according to the classical test model which underlies nearly all normative questionnaires.
2. Ipsative test scores are not equivalent in meaning to normative scores. The psychometric structure of the measurement made by both forms of test is different.
3. Deleting a scale in order to adjust for the perfect collinearity between scales does enable more complex covariance-based analyses to be undertaken, however, the solutions so produced are fundamentally different from the normative data solutions.
4. Within the ipsative SIV, systematic bias was shown to be present amongst the forced choice triads.

On the basis of the reported results, we conclude that that there is little justification for the continued use of the ipsative, forced-choice item format, psychological test.

Introduction

Cattell (1944) introduced the term 'ipsative', and defined it as measurement relative to other measures within the individual. Ipsative scores reflect only relative strengths of traits within the individual. An ipsative scale uses the behaviour of the individual to create its own standard. For example, a patient's condition may be viewed as having either improved or declined relative to the patient's own average or relative condition. In contrast, normative scales measure absolute differences, and reflect an underlying continuum common across all people, as in measures of IQ.

The widespread use of ipsative measurement came about as a response to rater bias exhibited in questionnaire personnel ratings. Travers (1951) credits Paul Horst with first proposing the idea of forced-choice format to counter 'leniency' and other errors in rating. The forced-choice method involves the presentation of items that have been matched for preference value (e.g. social desirability) yet discriminate differentially on a set criterion, such as leadership quality for a specific task (Gordon, 1951). However, two important assumptions underlie this approach. Firstly, all the choices must be as high in apparent validity as each other, and secondly, the ratees will, on average, ascribe equal status to the irrelevant qualities (Guildford, 1954).

In a review of the properties of ipsative and normative measurement, Hicks (1970) concluded that ipsative measures possess such extensive psychometric limitations that their continued use was undesirable. The mathematical arguments and empirical evidence of others such as Clemans (1956, 1966), Horn (1966), Radcliffe (1963), and Closs (1976), left little doubt as to the severe problems for any assessment specialist attempting to use classical methods of test validation on an ipsative questionnaire. However, these arguments did not lead to the abandonment of existing ipsative tests, rather with the introduction of the Saville and Holdsworth OPQ series, an increase (rather than decrease) in the use of ipsative tests has recently taken place in the field of occupational selection testing.

Johnson, Wood, and Blinkhorn (1988) recently re-stated the arguments for the abandonment of ipsative testing via questionnaire, and provided some empirical examples of the error-prone consequences of their use. This article was, perhaps, the strongest indictment of ipsative measurement that has appeared to date. Saville and Willson (1991) responded to this article by attempting to demonstrate that ipsative tests manifest equal, if not superior, validity to normative tests. Using a novel, if somewhat ill-specified computer-generated dataset, they showed that under certain conditions ipsative and normative tests will yield equivalent psychometric parameters. In addition, they went on to show that, with certain real datasets, the expected statistical results from Johnson et al were not observed. However, these conclusions have been challenged by Cornwell and Dunlap (1994) who carried out a re-analysis of the Saville and Willson data and found little support for their claims.

With regard to ipsative vs normative measurement, the entire area of questionnaire measurement is now in a state of flux. The studies reported below attempt to address this issue from an independent standpoint, using both computer simulated and real data from not just one, but a range of ipsative and normative measures. The specific objectives were:

1. To examine the validity of using classical psychometric methods as a means of evaluating the reliability and validity of ipsative tests.
2. To assess the psychometric properties of ipsative measurement.
3. To compare and contrast the practical measurement implications of using ipsative vs normative questionnaires.

Respondents, Tests, and Procedures

The study involved 2 strategies, empirical and synthetic. The empirical strategy involved the administration of an ipsative and a normative version of a psychometric test of personality to a relatively large sample together with a number of other measures that were used as construct and measurement validation criteria. The synthetic strategy involved writing a computer program that simulated ipsative measurement. The focus here was to generate normative independent scores with a normal distribution by random number generation. These scores were then ipsatised to provide a directly equivalent ipsative form. The program generated scores from 1000 cases for a variety of ipsative conditions.

1072 respondents, drawn from Surrey and Luton University students, as well as from the Institute of Psychiatry's Biosignal Laboratory general population volunteer database, and from a sample of the general population within Guildford, composed the respondent sample. 637 were female, 435 were male, with a median age of 26, ranging from 17 to 72 years of age.

The main ipsative measurement device used was Gordon's (1976) Survey of Interpersonal Values (SIV). Fifteen items tapping each of the 6 scales are juxtaposed against each other in 30 triads, and the respondent must choose the most and least favourable in each set. A normative version of the SIV has been developed for research purposes (Knapp, 1964, Roberts, 1985) and has been psychometrically evaluated by Roberts (1985). A further ipsative test was developed specially for this study as a measure of the 5-factor model personality traits. This adjective checklist format test is called the Hammond-Barrett-Savage-Big-5 (HBS5) test. The ipsative version involves ranking 10 sets of 5 personality traits according to their accuracy in describing the respondent. The normative form simply asks that each of the 50 items be rated on a 5-point rating scale.

The Occupational Personality Profile (OPP: Paltiel, 1986) was also included as a validation measure. It was originally envisaged that the Jackson Personality Inventory might be used in this context but the OPP was chosen in preference due to the fact that it is widely used in job selection assessment in the UK, and that it has been developed primarily for use within the UK general population.

Tests were administered in group settings and also individually (self-completion). The subject pool was divided into a number of groups:

Group 1: received the *ipsative* SIV and 3 months later received the *ipsative* SIV again (N=185)

Group 2: received the *normative* SIV and 3 months later received the *normative* SIV again (N=170)

Group 3: received the *ipsative* SIV and 3 months later received the *normative* SIV (N=341)

Group 4: received the *normative* SIV and 3 months later received the *ipsative* SIV (N=376)

With the administration of the ipsative or normative forms of the test, respondents were also presented with a further psychometric test. This was either the FIRO-B (N=203) or the OPP (N=734). 139 students received the HBS5 ipsative form followed up 3 months later with the normative form.

Results

The results are described under 4 main headings:

1) Problems of Measurement

The first series of analyses was designed to examine the effects of a classical psychometric analysis on ipsative tests. Both the results for the SIV and HBS5 are reported below.

Table 1: Classical Psychometric Analysis of the SIV and HBS5

	Alpha Coeff.		**Mean Inter. r**		**Test Retest**	
Scale	**Ipsat.**	***Normat.***	**Ipsat.**	***Normat.***	**Ipsat.**	***Normat.***
SIV						
Support	0.76	0.90	0.17	0.38	0.47	0.84
Conformity	0.79	0.92	0.21	0.45	0.38	0.70
Recognition	0.59	0.92	0.10	0.46	0.48	0.73
Independence	0.79	0.90	0.18	0.37	0.42	0.69
Benevolence	0.65	0.92	0.11	0.42	0.44	0.64
Leadership	0.74	0.95	0.16	0.57	0.35	0.76
HBS5						
Openness	0.57	0.73	0.10	0.21	-	-
Agreeableness	0.51	0.60	0.09	0.10	-	-
Extraversion	0.58	0.66	0.12	0.13	-	-
Anxiety	0.59	0.74	0.12	0.21	-	-
Conscience	0.43	0.72	0.07	0.20	-	-

The immediate finding from the table above is that the alpha reliabilities are consistently lower in the ipsative case than in the normative. A similar pattern emerged in the studies by Knapp (1964) and Sweet (1989), and it is compatible with the fact that there is a mathematical constraint placed upon the maximum inter-item correlations (the **inter. r** column in the table above) of the ipsative items. A more

serious concern is whether these alpha indices are at all meaningful, since the basic classical test model is not definable where random error variance cannot exist, as is the case in ipsative tests. A more heuristic strategy for assessing reliability in such a case may be to examine the stability of test scores across time. In the case of the SIV data reported above, these test-retest correlations for the ipsative form are substantially lower than those obtained for the normative form. Indeed, little indication of stability over the 3 month period is evident for the ipsative form, while the normative form manifests a high degree of test-retest reliability. The results for the HBS5 support those presented for the SIV. Although its normative psychometric structure is weaker than the SIV, the weakness of the ipsative format is nevertheless also apparent.

A principal component analysis of the SIV was also undertaken, contrasting the Ipsative and Normative component factor structures. Due to the collinearity introduced by ipsativity, no method of common factor extraction was possible, hence the adoption of principal component decomposition. A 6 factor solution was obtained and rotated to simple structure using an oblique procrustes rotation toward a hypothesised target matrix (Hammond, 1988). This matrix was made up from ones and zeroes, corresponding to the scoring key for the items. The Procrustes procedure was used in order to maximise the congruence between the two solution structures. Congruence coefficients, computed between the factors from the two solutions, are reported below in Table 2.

Table 2: Factor Congruence Between Ipsative and Normative forms of the SIV

		Normative					
		S	**C**	**R**	**I**	**B**	**L**
Ipsative	**S**	***0.63***	0.02	0.04	0.08	0.01	0.12
	C	0.06	***0.75***	0.06	0.02	0.03	0.00
	R	0.01	0.11	***0.72***	0.03	0.03	0.01
	I	0.20	0.10	0.14	***0.64***	0.09	0.02
	B	0.11	0.06	0.07	0.19	***0.84***	0.04
	L	0.03	0.07	0.11	0.02	0.10	***0.72***

where: **S** = Support, **C** = Conformity, **R** = Recognition, **I** = Independence, **B** = Benevolence, and **L** = Leadership

From table 2, it is apparent that the congruence between the two forms was not large. This is a result similar to that found when comparing the two factor patterns found from the correlation matrices presented for the Concept Model OPQ by Saville and Willson. There are significant departures from similarity.

2) Problems of Equivalence

One of the main claims made by Saville and Willson (1991) was that there is a high degree of equivalence between the ipsative and normative forms of the same test. This claim was tested by correlating the SIV scale scores provided by 208 university students with their FIRO-B scale scores. The FIRO was used in this context because it had been used by Gordon in his initial validation of the SIV. As expected, the correlations of the ipsative form were much lower than those for the normative version. However, what was less expected was that the patterns of correlations between the two forms and the target scales were not consistent. Most notable was the finding that the Independence scale on the ipsative form was most highly correlated with 4 out of the 6 FIRO scales, while in the normative form, it manifested the lowest of the SIV correlations across all FIRO scales. In addition, some of the correlations between the ipsative form of the SIV and the FIRO simply did not seem to make psychological sense. Although our analysis was similar to that published by Gordon, he failed to comment on the problem of the dubious psychological meaning of some of the scale intercorrelations.

Delving further into this issue, a multi-trait-multi-method (MTMM) analysis was undertaken on the normative and ipsative SIV. The average correlation in the validity diagonal of the MTMM matrix was 0.10. Although more sophisticated methods of analysis exist for the analysis of such MTMM matrices, none could be used here because of the ipsativity constraints . However, a new method of analysis of MTMM matrices that compare ipsative and normative tests was developed (Hammond, Barrett, and Wilson, 1993) that involves the use of nonmetric multidimensional scaling (MDS) procedures. These methods do not require the use of product-moment correlation and are insensitive to the ipsativity measurement constraints. Essentially, nonmetric MDS attempts to reconstruct the relative rank order of intervariable similarities, however measured. As with other forms of confirmatory analysis, a target matrix can be specified against which the input matrix of similarities may be fitted. The outcome of this analysis again demonstrated that the trait measures from both forms of the SIV do not correspond to one another.

A final attempt was made to ascertain the equivalence of the two forms using a categorical correspondence/dual-scaling analysis procedure. Interestingly, Cornwell and Dunlap (1994) also chose to limit the interpretation of ipsative measures using a purely categorical level of analysis). The basic idea involved the generation of a contingency table with the SIV scales represented as columns and the variables against which they were to be validated as rows. e.g. Normative scales as columns, Ipsative as rows. The entries in each row of this table were simply the frequencies of respondents who scored above the median on both the row and column variables in question. The main diagonal of this matrix should have yielded frequencies consistently higher than the off-diagonal entries. However, the results of this analysis indicated that the pattern of frequencies in this table was essentially random.

3) Problems of Validation

A number of early studies suggested that, in certain circumstances, ipsative measurement might be more valid than equivalent normative methods. Traditionally, validity assessment is usually implemented by calculating the correlation between some measure and a criterion variable. In the case of a multi-trait measure, this usually

involves the use of a multiple regression procedure. However, as has been demonstrated elsewhere, the correlations of ipsative scales with an external criterion are mathematically constrained to sum to zero. It is hard to see, therefore, the value of a single correlation coefficient as being of much use. It is affected as much by the relationship between the scales in an ipsative test as by the relationship of a particular scale with a criterion. Nevertheless, there are a number of ipsative personality tests that are widely used in occupational selection and clinical decision support for which strong claims as to predictive validity are a necessary factor in their continued use.

One common attempt to mitigate the obvious distortion of ipsative measures in validation studies is to exclude one of the ipsative scales. Since the main problem can be viewed as the fact that all the scales from an ipsative measure will add up to a constant for all respondents, the exclusion of one scale allows the sum of scales to vary. This is a necessary condition for multiple regression since a complete set of ipsative scores produces a condition of total multicollinearity (every variable can be perfectly predicted from every other variable in a matrix of variable scores). In this situation, the regression equations cannot be solved and even reduced variance methods such as ridge regression are unable to proceed.

It is known that if one scale is deleted from a full set of ipsative scales, the multiple correlation resulting from a multiple regression on an external criterion will remain the same, *regardless of the scale that is deleted.* Thus, it is argued, ipsative tests can be used to predict with an observed degree of certainty the variation in an external criterion variable, and furthermore, this prediction is consistent irrespective of the scale that is deleted. Saville and Willson (1991), quoting correspondence from Lee Cronbach, argue that this fact may also be relevant in factor analysis, where the removal of one scale frees up the factor analysis to produce an unrestricted solution.

In order to test these assertions, a series of least squares multiple regression analyses were implemented using the SIV, HBS5, and a number of simulated ipsative scores, using a variety of external criteria. Applying the principle of removing one scale for each analysis, it was possible to demonstrate the consistency of the multiple correlation irrespective of the scale deleted. However, there were large variations in the regression weights for the same variables within each analysis. An example of these analysis is provided in Table 3 below, in which the ipsative SIV scores were used to predict the OPP scale score of Assertiveness. The analysis summarises the 6 regression analyses computed in which one of the SIV scales was excluded from the prediction equation. The unstandardised regression beta weight for each variable was observed for each analysis, enabling a mean beta weight and range of values to be computed. From this table of data, it can be seen that the weights varied considerably depending upon the combination of scales that were in the analysis. Thus the benevolence scale had a minimum beta weight of –0.19 and a maximum of 0.43, despite the fact that it was predicting the same criterion in each analysis. As expected, the multiple correlation coefficient remained the same (0.45) for all six analyses.

This finding completely invalidates Saville and Willson's (1991) and, by extension, Cronbach's contention that a factor analysis can be reasonably implemented on ipsative data by simply dropping one score. The interpretation of factor analysis depends

Table 3: Summary of SIV regressions on the Assertiveness scale of the OPP

SIV Scale	Maximum Beta	Minimum Beta	Average Beta
Support	0.62	0.11	0.39
Conformity	0.51	–0.11	0.25
Recognition	0.02	–0.51	–0.22
Independence	0.09	–0.53	–0.24
Benevolence	0.43	–0.19	0.17
Leadership	–0.09	–0.62	–0.35

entirely on the weights of the variables after regression onto a number of underlying traits. Thus, unless the focus of a factor analysis was simply to determine the amount of variance accounted for by each factor, this procedure is quite insupportable. The choice of which scale to drop will dramatically affect the interpretation of the factor solution.

4) Problems with Response Bias

The main reason for developing ipsative tests in the first place was to reduce the effect of social desirability. This approach only solves the problem if the average affectivities of the items that are juxtaposed against each other in the forced choice format are equal. If this is not the case, then forced choice procedures produce even worse artefactual distortion by building response bias into the test directly. In a simple summated rating format, the item parameters may be examined to identify the presence of any skew or distortion that might indicate response bias. However, in the case of forced choice tests, the identification of bias is a more complex affair. In the classical approach, item affectivity can be reliably estimated by examining the mean and standard deviation of item responses. Unfortunately, due to the mutual dependency of ipsative items, this is not an easy procedure with ipsative tests. In order to assess response bias, a configural frequency analysis (CFA: Lienert, 1986; von Eye, 1990) on the response profiles for each item was undertaken. Essentially, CFA as used here, assesses the probability that the response profiles for an item group (^nP_r profiles, where ^nP_r = the permutation of r responses from n items in the item group) differ significantly from one another. Thus, with respect to the ipsative SIV triads, there were six possible response profiles. The frequency of responses for each profile, for each item triad, were observed. The CFA analysis of these frequencies indicated that there was no evidence to suggest that that any of the item triads contained equally balanced item responses. In other words, response bias is a problem with the ipsative version of the SIV, which runs counter to the claims regarding "balance" made about ipsative tests by others.

Conclusions

The analyses above represent a comprehensive examination of various properties of ipsative format tests. It is apparent from the raft of results that classical methods of item, scale, and test analysis are of little value in interpreting the data from the ipsative tests analysed in this study. Given the results from the MTMM matrix analysis, the two test formats of the SIV produce quite divergent sets of test scores for the same respondents. With the categorical analysis of the same data, we were forced to

conclude that the agreement between the two test formats was at a random, chance level. The final analysis demonstrated that the supposed balance of response, a consequence of the forced-choice format, was missing within the entire set of 30 item triads within the SIV. With relevance to the Saville and Willson (1991) results using the 30-scale concept model ipsative OPQ, it may be that the massively increased number of scales does mitigate against the drawbacks of using ipsative format items for a few scales. However, our results indicate that attempting to collapse back these scales to higher order or more global interpretive scales will introduce significant and psychologically, quite misleading results. This is in agreement with Saville and Willson's own analysis. On balance, given our datasets and analyses, we would conclude that there seems little point in continuing with the development or use of ipsative tests. At best, they approximate normative data, at worst, they distort and change completely the psychological import of trait scores and their interpretation.

References

Cattell, R.B. (1944) Psychological Measurement: ipsative, normative, and interactive. *Psychological Review*, 51, 292-303.

Clemans, W.V. (1956) An analytical and empirical examination of some properties of ipsative measures. Unpublished doctoral dissertation, University of Washington, Seattle.

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

Closs, S.J. (1976) Ipsative vs normative interpretation of test scores or "What do you mean by like?". *Bulletin of the British Psychological Society*, 29, 228-299

Cornwell, J.M. and Dunlap, W.P. (1994) On the questionable soundness of factoring ipsative data: a response to Saville and Willson. *Journal of Occupational and Organisational Psychology*, 67, 89-100.

Gordon, L.V. (1951) Validation of the forced choice and questionnaire methods of personality assessment. *Journal of Applied Psychology*, 25, 407-412

Gordon, L.V. (1953) *Survey of Interpersonal Values (2nd. Ed.)* Chicago: Science Research Associates.

Guilford, J.P. (1954) *Psychometric Methods*. London: McGraw-Hill

Hammond, S.H. (1988) *The Psychometric Analysis Package*. Dept. of Psychology, University of Surrey.

Hammond, S.H., Barrett, P.T., and Wilson, S. (1993) A Facet Theory approach to troublesome MTMM matrix analysis. *International Conference of Facet Theory*, Prague.

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

Horn, J. (1966) Motivation and Dynamic Calculus concepts from multivariate experiment. In the *Handbook of Multivariate Experimental Psychology*, Cattell, R.B. (ed.) Chicago: Rand McNally and Co.

Johnson, C. E., Wood, R., and Blinkhorn, S. F. (1988) Spurioser and Spurioser: the use of ipsative personality tests. *Journal of Occupational Psychology*, 61, 153-162.

Knapp, R.R. (1964) An empirical investigation of the concurrent and observational validity of an ipsative vs a normative measure of six interpersonal values. *Educational and Psychological Measurement*, 24, 1, 65-73

Lienert, G. A. (1986) *Angewandte Konfigurationsfrequenzanalyse*. Frankfurt: Athaneum Press.

Paltiel, L. (1986) The Occupational Personality Profile. Letchworth, UK: Psytech International Ltd.

Radcliffe, J. (1963) Some properties of ipsative score matrices. *Australian Journal of Psychology*, 6, 1-10.

Roberts, S.E. (1985) *A flexible format for the volatile value. An evaluation of the psychometric properties of a Likert scaled alternative to the forced choice ipsative format of Gordon's SIV*. Dissertation submitted to the University of Surrey, UK.

Saville, P. and Willson, E. (1991) The reliability and validity of normative and ipsative approaches. *Journal of Occupational Psychology*, 64, 219-238.

Sweet, M.R. (1991*) A comparative study of Ipsative and Normative Measurements of Personality*. Dissertation submitted to the University of Surrey.

Travers, R.M. (1951) A critical review of the validity and rationale of the forced choice technique. *Psychological Bulletin*, 48, 62-70.

Von Eye, A.(1990) Introduction to Configural Frequency Analysis. New York: Cambridge University Press.