Does item homogeneity indicate internal consistency or item redundancy in psychometric scales?

By Gregory J. Boyle

Department of Psychology, University of Queensland, St Lucia 4067, Queensland, Australia

Abstract

The term ‘internal consistency’ has been used extensively in classical psychometrics to refer to the reliability of a scale based on the degree of within-scale item intercorrelation, as measured by say the split-half method, or more adequately by Cronbach's (1951) (Psychometrika, 16, 297–334) alpha, as well as the KR20 and KR31 coefficients. This term is a misnomer, as a high estimate of internal item consistency/item homogeneity may also suggest a high level of item redundancy, wherein essentially the same item is rephrased in several different ways.

Internal consistency or item homogeneity is often used for estimating intra-scale reliability, in terms of the item variances and covariances derived from a single occasion of measurement. While it is desirable that items in a psychometric scale measure something in common (i.e. exhibit uni-dimensionality), Hattie (1985) has indicated that there is still no satisfactory index. As Hattie (pp. 157-158) pointed out, a uni-dimensional scale (having an underlying latent trait), is not necessarily reliable, internally consistent or homogeneous. Hattie concluded that the frequent use of Cronbach’s alpha coefficient as a measure of uni-dimensionality is not justified. Hattie further stated that,

alpha can be high even if there is no general factor, since (1) it is influenced by the number of items and parallel repetitions of items, (2) it increases as the number of factors pertaining to each item increases, and (3) it decreases moderately as the item communalities increase.

The subsequent assertion by Ray (1988) that internal consistency of a psychometric scale should be maximised, represents a further restatement of classical itemetric theory, and ignores the previous work of Hattie (1985), and many others, as outlined below. There is an optimal range of internal consistency/item homogeneity, if significant item redundancy is to be avoided (Boyle, 1983, 1985. 1986). According to Kline (1979, p. 3), with item inter-correlations which are lower than about 0.3,

each part of the test must be measuring something different…A higher correlation than (0.7), on the other hand suggests that the test is too narrow and too specific…if one constructs items that are virtually paraphrases of each other, the results would be high internal consistency and very low validity.

Furthermore, according to Kline (1986, p. 3),

maximum validity…is obtained where test items do not all correlate with each other, but where each correlates positively with the criterion. Such a test would have only low internal-consistency reliability.
As Cattell (1978) pointed out, a scale comprised of many items which are essentially repetitions of each other can appear in factor analysis as a ‘bloated specific’ (as in Guilford’s S-O-I model of intellectual structure, cf. Brody & Brody. 1976). Kline (1986, pp. 118-119) further remarked that high internal consistency can be…antithetical to high validity…the importance of internal-consistency reliability has been exaggerated in psychometry (i.e. I agree with Cattell)…

According to Hayes. Nelson and Jarrett (1987, p. 972). “a measure could readily have treatment utility without internal consistency… high internal consistency should not necessarily be expected.” Likewise, as Allen and Potkay (1983. p. 1088). Lachar and Wirt (1981. p. 616) and McDonald (1981) have all shown, either high or low item homogeneity can be associated with either high or low reliability, despite classical itemetric opinion.

According to McDonald (p. 113). “Coefficient alpha cannot be used as a reliability coefficient…” McDonald (p. 100) has refuted on mathematical grounds, the commonly held belief that the alpha coefficient measures ‘internal consistency’ or item homogeneity of a scale. McDonald (p. 110) stated that, it has never been made clear what is meant by internal consistency or why KR-20 or coefficient alpha can be deemed to measure it…confusion pertaining to coefficient alpha has a long history…reviewed by Green, Lissitz & Mulaik (1977).

Furthermore, McDonald (p. 111) concluded that, alpha has not been shown to be a quantitative measure of any intelligible and useful psychometric concept, except when computed from items with equal covariances.

This conclusion was based on the original use of item homogeneity as an estimate of scale reliability by Gulliksen (1950). which was shown by Lord and Novick (1968) to be valid only when items are tau equivalent. Accordingly, it may often be more appropriate to regard estimates such as the alpha coefficient as indicators of item redundancy and narrowness of a scale (cf. Boyle. 1985).

Items should be selected which are loaded maximally by the factor representing that scale, but which exhibit moderate to low item inter-correlations in order to maximise the breadth of measurement of the given factor. Merely adding additional items to a scale as classical itemetrics has advocated in accord with the Spearman Brown formula, ignores the error variance associated with each item , and must be regarded by any contemporary and objective assessment (such as demonstrated with LISREL congeneric factor analysis-Joreskog & Sorbom, 1989), as being a rather unsophisticated method of increasing scale reliability. Ray (1988) uncritically cited Nunnally (1967) – not (1978)–as well as Cronbach (1951) in restating classical reliability theory. However, Pedhazur (1982, p. 636) has indicated that Nunnally’s classical approach to reliability failed to acknowledge that measurement errors are
often systematic and non-random. Ray’s comments arc therefore founded on psychometric views published over 20 years ago!

Ray (1988) claimed that broad validity of a scale is facilitated by the use of subscales. However, this ignores the fact that in many multidimensional psychometric instruments (such as the EPI, EPQ, JEPI, 16PF, CAQ, 8SQ, POMS, DES-IV, MAACL, etc.) each subscale actually measures a discrete factor analytic dimension. Despite Ray’s dogmatic assertions, semantic overlap of items is only one possible influence on observed item inter-correlations, as indicated above in relation to Hattie’s (1985) work. As well, Ray made no distinction between state vs trait scales (cf. Boyle, 1983, 1985, 1986, 1987). While a reliable trait scale should exhibit high test-retest correlations for both immediate retest (dependability) and for longer term retest (stability), a reliable state scale should exhibit only a high dependability coefficient, if the scale is truly sensitive to situational variability in mood.

Ray and Pedersen (1986) asserted on the basis of a highly biased, unrepresentative and very restricted sample of the U.S.A. population, that Eysenck’s Psychoticism scale in the EPQ was “a failed experiment”, not on the grounds of inadequate validity, but again merely on the basis of dated classical itemetric references. Ray objected to Eysenck’s Psychoticism scale because he found that the mean item inter-correlations were only moderate. Yet, Ray’s results with the EPQ were clearly biased due to severe restriction of variance in his data. Ray (1988) subsequently criticised Smedslund (1987) for not appreciating the virtues of the EPQ, despite denigrating it in the Ray and Pedersen note (cf. Smedslund, 1988). This amounts to little more than “the pot calling the kettle black”.

Ray (1988) recommended Comrey’s FHID approach to scale construction with the aim of increasing scale reliability. However, he was mistaken as to the actual composition of the item parcels in the CPS (four items counterbalanced for direction of scoring, not three as stated). While it is undoubtedly true that such item-parcel variables are more reliable than items as such, nevertheless, for a specified number of items in a scale, less of the pertinent construct is actually measured. Moreover, Cattell (1973, p. 360) has indicated that, “The high homogeneity in the FHIDs is carried over with the second-order factor scales, leaving them excessively homogeneous.” Hence, Ray’s assertions concerning scale construction with item-parcel variables would seem quite inadequate.

Cattell (1973, pp. 357-379; 1978, pp. 289-293; 1982) has argued that generally there is an optimally low level of item homogeneity. Cattell provided a conceptual demonstration of high item validity in the context of zero item homogeneity. Since a scale which is valid must also be reliable, it follows that it is theoretically possible for a scale to be reliable even though the ‘internal consistency’ is zero. On the other hand, it is well known that even a highly reliable scale is not necessarily valid. Any number of invalid scales can be made more reliable simply by adding further invalid items in accord with the Spearman-Brown prophecy formula, and/or by adding further items which are essentially mere repetitions of the items already included in the scale. Ray’s (1988) recommendations, if followed, can only result in significant item redundancy and likely contamination of the factor purity of psychometric scales.
The advantage of moderate to low item homogeneity is seen in multiple regression analysis, wherein a higher multiple R is produced from predictor variables (items) with only moderate item inter-correlations. Cattell’s “behavioural dispersion principal” suggests that only when there is considerable item diversity, enabling sampling of behaviours across a wide spectrum of life expressions, can individuals be advantaged equally in responding to the items in a particular psychometric scale. As well, reduced item homogeneity facilitates the maintenance of validity across different cultures. A given item may elicit discrepant responses in different cultural settings. If there is high item homogeneity and most of the items are similar (i.e. there is significant item redundancy—cf. Boyle, 1985), measurement error due to cultural distortions probably will be evident. This problem can be minimised by including a wide diversity of items (i.e. maximising breadth of measurement) in psychometric scale construction.

Cattell indicated that a scale which has high ‘internal consistency’ is probably contaminated on the one hand, by a bloated specific factor (such as in Guilford’s S-O-I model), wherein over-inclusion of particular items pertaining to a specific dimension, gives the impression of a substantive factor, despite its lack of practical significance and evident triviality. On the other hand, psychometric scale contamination occurs by inclusion of several items predictive of an unwanted common factor.

Cattell (1978. p. 289) demonstrated that,

> a very narrow specific can be blown up to the apparent status of a common factor in any given matrix by entering the experiment with several items that are close variants on the specific variable.

In this instance, item homogeneity (internal consistency) is increased by confounding the true factor with a bloated specific. Selection of items with high homogeneity/internal consistency, undoubtedly often results in a scale with a contaminated factor structure.

To minimise these distorting influences, it is desirable to invoke suppressor action by including items that are loaded positively and negatively on the unwanted dimensions, which also are loaded significantly on the relevant common factor. In contrast to Ray’s (1988) restatement of classical itemetric opinion, Cattell (1973, p. 359) asserted that. “In practice… the random tendency to opposite loadings on these other factors will reduce the item homogeneity virtually to zero….” Item diversity therefore, results in reduced item homogeneity and concomitantly, reduced item inter-correlations, but maximises breadth of measurement of a given construct. However, Cattell cautioned that,

> since low homogeneity means different specific factors and suppressor action by opposite loadings on unwanted common factors… a test which (misguidedly) advertises high homogeneity is contaminated either with a bloated specific or by items sharing a common unwanted factor.

In summary, high internal consistency/item homogeneity results spuriously from the inadvertent inclusion of essentially similar items in a psychometric scale.
Determination of what should be considered appropriate item homogeneity for a scale is, according to Cattell (1973. pp. 361-362)

far more complex than is commonly considered in classical itemetrics... The complexity is generated on the one side by... the natural history of the domain... and on the other by the unusual complexity of the purely statistical psychometric laws involved.

According to Cattell, to obtain a broad but valid, behaviourally based rather than semantically based scale, “test constructors will need to sift by factor analysis hundreds of items to get those having validity despite high diversity”. In this regard, the newer congeneric factor analytic methods using programs such as LISREL (Joreskog & Sorbom, 1989) will undoubtedly minimise the amount of ‘noise’ which is so prevalent among the items of many existing psychometric scales, designed along classical psychometric lines, wherein ‘internal consistency’ has been maximised. This traditional itemetric view of intra-class correlation still persists in the contemporary psychometric literature [e.g. Crocker & Algina, 1986, pp. 119-122: Cronbach, 1990, pp. 202-204; Ferguson, 1981, pp. 438-439: also see Boyle, 1987, for a discussion of the limitations of the (1985) AERA/APA/NCME Standards in this regard]. However, especially in the non-ability areas of motivation, personality and mood states, moderate to low item homogeneity is actually preferred if one is to ensure a broad coverage of the particular constructs being measured.

References


