Statistics - Reliability, Validity, Consistency

"A Nonparametric Coefficient of Internal Consistency", Robert R. Trippi and Robert B. Settle, *Multivariate Behavioral Research*, Vol. 4, No. 11, (October 1976) pp 419-424

Alternative to Cronbach's Alpha

A NONPARAMETRIC COEFFICIENT OF INTERNAL CONSISTENCY

ROBERT R. TRIPPI and ROBERT B. SETTLE
San Diego State University

ABSTRACT

This paper discusses a frequency based, nonparametric measure of internal test consistency, referred to herein as coefficient alpha$_T$, which allows facile measurement of the significance of differences in internal consistency between tests, administrations, or scoring methods. It also permits analysis of psychological tests containing items with discrete categories of response, yielding nominal scale data. Use of alpha$_T$ encourages flexibility in test construction, since multiple dimensions can be incorporated into individual test items.

The traditional measure of the internal consistency reliability of a test is defined as the ratio of the variance of true test scores to the variance of the observed test scores. The internal consistencies of many psychological tests have been computed by the estimation formulae known as coefficient alpha (Cronbach, 1951), which for dichotomous choice items reduces to the well-known formula, KR-20 (Kuder & Richardson, 1937). The literature on these measures, their equivalences and properties is voluminous (Cureton, 1958; Dressel, 1940; Horn, 1971; Lord, 1955; Lyerly, 1958; Osborn, 1969). Although some work has appeared on the testing of differences between internal consistency reliability coefficients, the methods are either not readily available or difficult to use (Feldt, 1965; Kristoff, 1964).

Since coefficient alpha includes the variance parameter both in its definition and computational formulae, it is, strictly speaking, a parametric statistic. It is not invariant to nonlinear monotonic transformations on the item scales. We propose here a new definition of internal consistency: one which is nonparametric, based solely on deviations of observed item choices from those to be expected from random choice. Since this definition of internal consistency is quite different from that of coefficient alpha, we should expect there to be no precise mathematical relationship between the two measures.

Definition of Coefficient Alpha$_T$

For a test having a total of k items i, each of which has c_i choices, administered to n subjects j, we first compute the total observed number of items for which subject j chooses alternative q. Let this be represented as o_{jq}. Next, the expected number of times

under random choice that an individual would have chosen q, e_q, is computed. For the most common case of a test in which each item has the same number of choices m, the maximum number, we would thus use $e_q = k/m$ for all q.

Next, a χ^2 value is computed over all of the subjects in the sample:

$$\chi^2 = \sum_j \sum_q (o_{jq} - e_q)^2/e_q \ ,$$

and the maximum possible value of χ^2 (for a perfectly consistent test) is computed as

$$\chi^2_{\max} = n((k - e_1)^2/e_1 + \sum_{q=2}^{m} e_q) \ .$$

That is, the maximum χ^2 value for each subject will occur when he consistently chooses alternative 1 for every item. For a test having the same number of choices for each item, the pivotal index in the formula for $\chi^2_{\max}$ would obviously be immaterial; the identical value would result from using 2, 3, etc. as the first subscript in the formula since the e_q's are all equal.[1] The summation term would then simply exclude this subscript.

The nonparametric coefficient of internal consistency, alpha$_T$, is defined as the ratio of computed to maximum possible χ^2:

$$\text{alpha}_T = \chi^2/\chi^2_{\max} \ .$$

Alpha$_T$ shares many of the operational characteristics of the parametric consistency coefficient. Examination of the above definition will reveal that its range is from zero to one. If choice by all subjects on all items is totally random, χ^2 will be zero, and hence the minimum internal consistency of zero will result. If, on the other hand, on a given test subject 1 chooses, for example, alternative x on every item, subject 2 chooses alternative y on every item, subject 3 chooses alternative z on every item, and so on, the maximum possible value of χ^2 cumulated over subjects as above will be attained and the test will be a perfectly consistent one with

1. This simplified formula derives from the more general χ^2 formula for identically distributed random variables: $\sum(o-e)^2/e$. With alternative i chosen k times, its contribution is $(k-e_i)^2/e_i$, while the contribution of the remaining unchosen alternatives is $\sum_{q\neq i}(o-e_q)^2/e_q = \sum_{q\neq i} e_q$. Since there are n subjects taking the identical test, $\chi^2_{\max}$ includes this factor.

alpha$_T$ = 1. Unlike the parametric coefficient alpha$_T$ is extremely insensitive to both sample size and test length. However, the outcomes of tests of significance will obviously depend upon sample size and test length, since these determine the degrees of freedom. More importantly, alpha$_T$, unlike the parametric internal consistency statistic, can be used even though no progression exists in choice alternatives. Thus, we do not even require an ordinal scale, and choice alternatives for a given item could be, for example, *1. Always, 2. Never, 3. Sometimes, 4. Don't know.*

In the typical use of χ^2 one assumes independence of observations, and determines whether this assumption can be disproved. The degree of departure from independence is specifically what is meant here by internal consistency, and measured by alpha$_T$. Perfect independence, and resultant value of alpha$_T$ of zero, implies total lack of consistency among the items of the test.

There is no way of directly comparing values of alpha to those of alpha$_T$. Both are sensitive to differences in test construction and even scoring methods, as will be evident in the later examples. Alpha$_T$ measures the strength of the interitem concurrence of choice for each subject, and its cumulated value over subjects. Therefore, it is assumed that ordinal scaled items will be reflected beforehand where necessary, or that choice alternatives for multiple dimensioned items will be reordered prior to scoring to provide coincidence of construct with numerical choice.

Significance of Intertest Differences in Internal Consistency

Since it is defined in terms of the χ^2 distribution, any computed value of alpha$_T$ can be tested for significantly differing from zero (a perfectly inconsistent test) or any specified minimum consistency level, by reference to the χ^2 table, using $n(m-1)$ degrees of freedom. For large values of $n(m-1)$ the standard normal approximation

$$z = \sqrt{2\chi^2} - \sqrt{2\alpha\chi^2_{max}} - \sqrt{2n(m-1)-1}$$

may be used for any hypothesized consistency level, α.[2]

2. This is the well-known normal approximation for large degrees of freedom to the χ^2 distribution with mean shifted by the hypothesized consistency level α [Kendall, 1963, p. 371].

In addition, by use of the measure, the internal consistencies of two tests may be tested for significant intertest differences. For each test, the standardized χ^2 factors $R = \chi^2/n(m-1)$ are first computed. The ratios of pairs of such factors can be tested for level of significance by reference to the F table, with the two appropriate degrees of freedom. Thus, it is not only a simple matter to determine whether their internal consistency measures differ significantly from any predetermined value for a particular test or administration, but significant intertest differences in consistency can also be detected with relative ease.

Examples of Applications

Two paper and pencil tests of need for achievement were jointly administered to a group of 180 undergraduate business administration students. Within each test, items were randomly ordered. The Mehrabian test is a frequently used instrument (Mehrabian, 1968). The Hermans test is relatively new, without a large body of interpretive data outside of the standardization sample reported by the author (Hermans, 1970).

The Mehrabian test is assumed to yield interval scaled data, while the Hermans test requires multiple choice responses that usually would be regarded as either nominal or ordinal scaled, depending on one's interpretation. By the most rigorous point of view, however, the Hermans would be considered a test yielding nominal data, due to the lack of strict transitivity or unidimensionality among many of its alternatives. The author's suggested scoring method is a dichotomous one pivoting on the modal choice, although four or five choices are actually available for each item. In the study, the Hermans test was scored using both dichotomous and full choice methods.

The results of the test analyses are summarized in Table 1. Note that although the Hermans dichotomous scoring method and the Mehrabian test yielded virtually the same parametric coefficient alpha values of .72, they differed quite markedly in their values of coefficient alpha$_T$, with a value of .22 for Hermans and .11 for the Mehrabian. The difference is significant at the .05 level with $F = 6.47/2.78$. Interestingly, the Hermans test yielded a greater value of alpha for the full choice than for the dichotomous scoring method, with coefficients alpha of .82 and .72, respectively. On the other hand, the corresponding values of coefficient alpha$_T$ were

Table 1
Achievement Test Results

	Hermans continuous	Hermans discrete	Mehrabian
Mean score	96.12	37.69	140.19
Standard deviation	9.89	3.63	19.03
Variance	97.82	13.16	361.99
Number of subjects	180	180	180
Number of items	29	29	26
Maximum number of alternatives	5	2	9
Coefficient alpha	0.82	0.72	0.72
Coefficient alpha$_T$	0.16	0.22	0.11
Chi-square value	2286	1165	4012
Degrees of freedom	720	180	1440
Chi-square/d.f.	4.00	6.47	2.78
Z value	38.04	29.32	35.92

.16 and .22. The example demonstrates that the two internal consistency coefficients are not even generally comparable in absolute magnitudes, though both can range from zero to one. Values of .2 to .4 appear to be relatively strong for alpha$_T$, while normally values of .5 to .9 are considered strong for parametric coefficient alpha.

Conclusions

Previously the test builder who desired a measure of internal consistency was confined to unidimensional items with interval scales. With the use of coefficient alpha$_T$, a frequency based rather than a variance based measure, a measure of internal consistency is available even though the test contains multiple choice alternatives that are discrete. In fact, there is no necessity that the alternatives for any given item be ordinal or even contained on a single dimension. Thus, the test constructor might use a series of items whose alternatives indicate, for example, compliance, aggressiveness, or detachment. Use of the conventional, parametric measure of internal consistency would require that items be confined to one dimension, with three subscales for the different modes of response. Coefficient alpha$_T$ permits the three modes to be opposed to one another within each item, over various situations. Consequently, the nonparametric method permits greater flexibility in test construction, and a closer approach to the real-world choice situation.

REFERENCES

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-333.

Cureton, E. E. The definition and estimation of test reliability. *Educational and Psychological Measurement*, 1958, *18*, 715-738.

Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. *Psychometrika*, 1940, *5*, 305-310.

Feldt, L. S. The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika*, 1965, *30*, 357-370.

Hermans, H. A questionnaire measure of achievement motivation. *Journal of Applied Psychology*, 1970, *54*, 353-363.

Horn, J. L. Integration of concepts of reliability and standard error of measurement. *Educational and Psychological Measurement*, 1971, *31*, 57-74.

Kendall, M. G. *The Advanced Theory of Statistics* (Vol. 1). London: Chas. Griffin & Co., 1963.

Kristoff, W. Testing differences between reliability coefficients. *The British Journal of Mathematical and Statistical Psychology*, 1964, *17*, 105-111.

Kuder, G. F., & Richardson, M. W. The theory of the estimation of test reliability. *Psychometrika*, 1937, *2*, 151-160.

Lord, F. M. Estimating test reliability. *Educational and Psychological Measurement*, 1955, *15*, 325-336.

Lyerly, S. B. The Kuder-Richardson formula (21) as a split-half coefficient and some remarks on its basic assumption. *Psychometrika*, 1958, *23*, 267-270.

Mehrabian, A. Male and female scales of the tendency to achieve. *Educational and Psychological Measurement*, 1968, *28*, 493-502.

Osborn, H. G. The effect of item stratification on errors of measurement. *Educational and Psychological Measurement*, 1969, *29*, 295-301.