EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1977,37

LIMITATIONS OF COEFFICIENT ALPHA AS AN INDEX OF TEST UNIDIMENSIONALITY[1]

SAMUEL B. GREEN[2,3]
Auburn University
ROBERT W. LISSITZ
University of Georgia
STANLEN A. MULAIK
Georgia Institute of Technology

Confusion in the literature between the concepts of internal consistency and homogeneity has led to a misuse of coefficient alpha as an index of item homogeneity. Coefficient alpha is actually a complexly determined test statistic, item homogeneity only being one influence on its magnitude. The related statistic, the average intercorrelation, has similar difficulties. Several indices of item homogeneity derived from the model of common factor analysis are offered as alternatives.

IN this paper we intend to show how a confusion in the literature between the concepts of internal consistency and homogeneity has led to a misuse of coefficient alpha as an index of item homogeneity in test construction. We offer numerical counterexamples to show how coefficient alpha and the item-total-score correlations can be high when the component items are not homogeneous. We also offer several indices of item homogeneity derived from the model of common factor

[1] This research was supported in part by the Auburn University Research Council.
[2] Authors' names appear in alphabetical order. Green and Lissitz contributed the initial conceptualization of the project. Green was primarily responsible for computer programming and analysis of data generated. Mulaik contributed much of the literature review and mathematical analysis. All work was the result of discussions among the authors.
[3] Reprints may be obtained from Samual B. Green at the Department of Psychology, Auburn University, Auburn, Alabama 36830.

analysis. We conclude that measures of internal consistency and reliability should not be used to indicate homogeneity.

Originally coefficient alpha (Cronbach, 1951) never was intended to be an index of item homogeneity in the construction of composite tests and attitude scales. Coefficient alpha is an internal consistency estimate of composite test reliability. But it is our impression that in recent years coefficient alpha has come to be regarded by some test and scale constructors as of central importance in the construction of homogeneous composite tests. For example, Armor (1974) designates a common method of constructing homogeneous composite scales as *covariance scaling*. He defines covariance scaling as "the technique of maximizing the alpha reliability of a composite under various constraints." And he adds,

> Other terms used to describe this scaling process include *item analysis* (Upshaw, 1968), test construction (Cronbach, 1960), summated ratings (Edwards, 1957), and *Likert scaling* after the psychologist who invented the agree-disagree item scoring scheme. Sociologists also often use the term *index construction* to denote a similar procedure. The term covariance scaling can be applied to all these methods since they are all based on the same basic and self-evident assumption: if a set of items is measuring the same or similar properties and the property comprises a single continuum or dimension, the items should all covary to some extent. For a fixed number of items, the greater and more consistent the inter-item correlations the more reliable the composite (Armor, 1974, p. 23).

Armor (1974) furthermore cites 5 steps usually followed in the process that he calls covariance scaling:

1. Inspection of item face content to ensure clear and consistent meanings
2. Calculation and inspection of item-to-scale correlations to pinpoint items that are not contributing to the scale as a whole
3. Calculation of a T-test for each item between the highest 25 percent and the lowest 25 percent of the composite scale scores (usually a substitute for step 2)
4. Examination of all inter-item correlations for patterns of lower or negative correlations
5. Recalculation of alpha reliability after eliminating items according to information obtained in steps 1 to 4

Similar steps for the construction of homogeneous composite tests and scales have been cited by Nunnally (1970, pp. 212–217) and Crano and Brewer (1973, pp. 228–242).

Armor (1974) notes that not all scale constructors perform each of

these five steps. Many, he observes, do not perform step 4, which he claims is critical for the construction of homogeneous scales. We concur in these observations for we commonly observe test constructors emphasizing just obtaining item-total-score correlations to identify irrelevant items and using coefficient alpha to evaluate the overall success of picking a homogeneous set of items.

Now, we claim that coefficient alpha is not a good index of test homogeneity regardless of how well it serves the purpose of indicating reliability. But we think that the tendency to use coefficient alpha as an index of homogeneity is based on a confusion between the concept of "internal consistency" and the concept of "homogeneity" that occurs in the literature. For example, Nunnally (1970, p. 125) seems to treat these two terms as referring to the same concept: "The individual who works in the field of testing will encounter special methods of reliability estimation based on homogeneity, or amount of correlation between item responses within one set. . . . What the equations concerning homogeneity (or internal consistency) do is to estimate the correlation between an existing test and a hypothetical equivalent form, one that may never actually be constructed." But in this case Nunnally uses the term "homogeneity" for what most other authors (including ourselves) normally refer to by "internal consistency"; for example, Crano and Brewer (1973, p. 229) write, "The term *internal consistency* best describes the condition in which there is a high degree of interrelatedness among items . . ." On the other hand, many other authors use the term "homogeneity" to refer to the case where a set of items all measure a single common dimension. For Lord and Novick (1968, p. 95) homogeneous tests are composite tests whose components are all essentially tau-equivalent: "Thus a homogeneous test is one whose components all 'measure the same thing' in their true-score components." A slightly weaker definition for a homogeneous test would be that the test consists of components whose true-score parts are all linear functions of a single common factor, that is, a set of congeneric components. Gulliksen (1950, p. 220) is in accord with this definition when he writes, "If there is only one common factor (among some items), the items are homogeneous." The factor analytic concept of homogeneity used by Gulliksen is, however, not the last word on this matter. Lord and Novick (1968, p. 95) suggest that in a broader context all items whose true scores can be shown to be monotonic increasing function of some single variable could be regarded as homogeneous. But such items' true scores could be related to one another in a nonlinear way so that their correlation matrix would not necessarily produce a single common factor. Unidimensionality might be discovered only with some nonmetric multidimensional scaling

procedure. While we recognize this possibility, we will nevertheless confine our discussion of homogeneity in this article to the classic test-theory idea that homogeneous items have but a single common factor among them and are related to the underlying factor of ability or attitude in a linear manner. Our remarks will apply then only in an approximate manner to items scored dichotomously or in grading fashion where the regressions of observed scores onto the latent ability or attitude dimension are monotonically increasing but nonlinear in nature.

Now, while it may seem that the distinction between internal consistency (which implies interrelatedness but not necessarily unidimensionality among a set of items) and homogeneity (which implies unidimensionality among a set of items) can be made clear, consider that many authors confuse this distinction when they come to discuss the use of coefficient alpha in test construction. Gulliksen may have generated some of this confusion when three pages after giving his factor analytic definition of a homogeneous test, he cited a formula for an index of item homogeneity that is mathematically equivalent to coefficient alpha (Gulliksen, 1950, equation 10, p. 223). On the other hand, Crano and Brewer (1973, pp. 229–232) define internal consistency as interrelatedness but then go on to suggest that they place the "emphasis on internal consistency to ensure that an attitude scale measures a single attitudinal disposition. . . ." And they claim that " 'Coefficient alpha,' the average interitem correlation (sic) of all items constituting a scale, represents probably the best estimate of internal consistency." And they furthermore state, "If a scale cannot satisfy the criterion of internal consistency (i.e., if coefficient alpha is not at least in the .80's) the investigator must attempt to determine the probable reason for this failure." The confusion between internal consistency and homogeneity is also fostered by statements in the semiofficial *Standards for Educational and Psychological Tests and Manuals* (1966) such as, "If the test manual suggests that a score is a measure of a generalized homogeneous trait, evidence of internal consistency should be reported." "Estimates of internal consistency should be determined by the split-half method or methods of the Kuder-Richardson type (coefficient alpha)."

Much of the emphasis on the use of coefficient alpha in the evaluation of homogeneous scales and tests can be traced to Cronbach's (1951) demonstration that coefficient alpha is high when a test is homogeneous. Cronbach relates coefficient alpha to the internal structure of the test items as formulated by a common-factor-analysis model (which we feel has sufficient generality for these purposes). Now, Cronbach (1951, p. 320) maintains, "If . . . a test is composed of a group of items, each measuring a different factor, it is uncertain

which factor to invoke to explain the meaning of a single score. For a test to be interpretable, however, it is not essential that all items be factorially similar. What is required is that a large proportion of the test variance be attributable to the principal factor running through the test." He shows in his paper how each item could be regarded as being determined by a general factor, several group factors and specific and error factors. The weight assigned the general factor need not be high for any particular item. But with a large number of items so composed he shows that the proportion of the total-score variance on the test due to the general factor among the items will be large. He then shows how coefficient alpha is a lower bound to the proportion of the total-score variance due to the common factors and an upper bound to the proportion of the total-score variance due to the first common factor (when the partial correlations among the items with the first common factor removed are all zero or positive). For Cronbach the first common factor can be identified with a general factor. But the upper-bound property for coefficient alpha holds for any common factor taken as the first common factor, even in situations where there is no general factor.

The fallacy of relying on Cronbach's results as justification for the use of coefficient alpha as an index of test homogeneity lies in mistaking necessary properties of homogeneity for sufficient properties of homogeneity. Certainly high internal consistency as indicated by a high coefficient alpha will result when a general common factor runs through the items of the test. But this does not rule out obtaining high internal consistency as measured by coefficient alpha when there is no general factor running through the test items. Since coefficient alpha is a lower bound to the proportion of total-score variance due to common factors running through the items, one can establish high values for coefficient alpha when most of the item variances are determined by several common factors. In other words, while homogeneity implies high internal consistency, high internal consistency need not imply homogeneity.

For example, assume that 10 items in a test occupy a five-dimensional common-factor space. Furthermore let the common-factor space be spanned by five orthogonal factors and let each item load equally $\sqrt{.45}$ on two factors in such a way that no two items load on the same pair of common factors. The resulting factor-pattern matrix looks like that in Figure 1. The factor-pattern matrix in Figure 1 satisfies requirements for simple structure since no item requires all common factors to account for its common-factor variance. The factors are also well-determined with four items determining each factor. Each item has a communality of .90.

The reduced correlation matrix $\mathbf{R} - \mathbf{U}^2 = \mathbf{FF}'$ has for its elements

		I	II	III	IV	V
	1	$\sqrt{.45}$	$\sqrt{.45}$			
	2	$\sqrt{.45}$		$\sqrt{.45}$		
	3	$\sqrt{.45}$			$\sqrt{.45}$	
	4	$\sqrt{.45}$				$\sqrt{.45}$
	5		$\sqrt{.45}$	$\sqrt{.45}$		
F =	6		$\sqrt{.45}$		$\sqrt{.45}$	
	7		$\sqrt{.45}$			$\sqrt{.45}$
	8			$\sqrt{.45}$	$\sqrt{.45}$	
	9			$\sqrt{.45}$		$\sqrt{.45}$
	10				$\sqrt{.45}$	$\sqrt{.45}$

Figure 1. Hypothetical factor pattern matrix for a model in which coefficient alpha equals .811.

the sum of cross-products for different pairs of rows in the factor-pattern matrix **F**. In our example when the correlation between any two different items is not equal to zero, the correlation between them will equal .45. There are 30 nonzero above-diagonal elements in the correlation matrix out of a total of 45 above-diagonal elements. Now, one form of coefficient alpha is given by

$$\alpha = \frac{n}{(n-1)} \frac{\text{sum of off-diagonal elements of } \mathbf{R}}{\text{sum of diagonal and off-diagonal elements of } \mathbf{R}}.$$

In our present example $\alpha = .811$. Furthermore the item-total-score correlation for any item in the example equals .608. Thus both coefficient alpha and the item-total-score correlations suggest that the 10 items in the example would form a homogeneous composite test by commonly accepted criteria. But the example by no means represents a unidimensional situation. The example also points out the dangers of omitting Armor's step 4 already mentioned. Fifteen of the 45 distinct interitem correlations in our example are equal to zero. That alone would be sufficient to alert the test constructor that the items are not

homogeneous. But some test constructors do not look at the correlations among the items, believing that item-total-score correlations convey all the information they need to discover irrelevent items.

To generalize the results of this example, we artifically constructed other models and obtained values of coefficient alpha for each. The models were constructed so that no item in a composite was determined by all factors present among them, and so the items satisfied the principle of simple structure. Within each model we required that every item load on the same number of factors, have equal-vlaued nonzero loadings, and have equal communalities. (These requirements made generating quantitative aspects of these models simple). Then across models we varied the number r of common factors among the items, the number L of factors relating to a single item, the communalities h_i^2 (required to be the same for every variable in a model), and the number of repetitions I (an integer greater than or equal to 1) of a basic set of items (a set of items representing the distinct combinations of r factors taken L at a time) in the model. The number of repetitions I was also limited by the number of items in a basic set (determined by the size of r and L) because just a few repetitions of a basic set involving a large number of items could make the model too unwieldy to handle. The number of repetitions I was also varied while holding all other influences constant. The total number of items included in any model was designated by n, which was equal to the number of combinations of r things taken L at a time multiplied times I.

For each model we computed the following: $\bar{r}_{ij}$, the average intercorrelation; α, coefficient alpha; μ, an index of unidimensionality; and $\hat{\mu}$, an estimate of μ (both μ and $\hat{\mu}$ will be discussed later). While we investigated a larger number of models, we include the results of only a representative few in Table 1. Each row gives the parameters r, L, h_i^2, I and n of a model along with the resulting values for $\bar{r}_{ij}$, coefficient α, μ, and $\hat{\mu}$ derived for that model. In this table we give the results for models with r's (number of factors) from 1 to 6, with L's (number of factors per item) from 1 to $r - 2$, h_i^2's of .3 and .9, and with varying n's as indicated.

After examining our results we can offer the following observations on the effect of varying these various influences with other influences held constant on coefficient alpha. (1) Alpha increases as n increases. (2) Alpha increases rapidly as the number of parallel repetitions of each type of item increases. (3) Alpha increases as the number of factors pertaining to each item increases. (4) Alpha readily approaches and exceeds .80 when the number of factors pertaining to each item is two or greater and n is moderately large. (5) Alpha decreases moderately as the items' communalities decrease.

The chief defect of coefficient alpha as an index of homogeneity is its susceptibility to increase with increases in the number of items n. Coefficient alpha can thus be large when the underlying dimensionality is high. In summary, it is apparent that alpha is too complexly determined to be a suitable index of homogeneity.

In all fairness to Cronbach (1951) we should point out that he realized the extent to which coefficient alpha is influenced by the number of items in the test. As an alternative index of internal consistency he recommended $\bar{r}_{ij}$, which can be computed, given coefficient alpha, from the formula

$$\bar{r}_{ij} = \frac{\alpha}{n - (n-1)\,\alpha}$$

which is the average correlation among different items in the test.

In our investigation we also computed $\bar{r}_{ij}$ for each model, and these values also appear in the designated column in Table 1. While it is true that $\bar{r}_{ij}$ is not affected by n and decreases as the number of factors increases, $\bar{r}_{ij}$ is unduly influenced by the communalities of the items. When there is but one common factor among the items $\bar{r}_{ij}$ can still be low if the communalities of the items are low. Furthermore, $\bar{r}_{ij}$ can be unduly influenced by negative inter-correlations among the items.

To overcome these disadvantages of $\bar{r}_{ij}$ we developed a new index of item homogeneity μ defined by the following formula:

$$\mu = \frac{\sum\limits_{i \neq j} \sum |r_{ij}|}{\sum\limits_{i \neq j} \sum \sqrt{h_i^2 h_j^2}}$$

The rationale behind this formula is that when there is but a single common factor among the items, their loadings on this factor equal the square root of their respective communalities. Also in this case the correlation between any two items equals the product of their respective factor loadings or the square root of the product of their respective communalities. For any particular pair of items i and j the following inequality holds: $|r_{ij}| \leq \sqrt{h_i^2 h_j^2}$. The equality holds for items occupying a single common-factor space. The inequality holds for items occupying more than one dimension. Thus the sum in the numerator for μ is always less than or equal to the sum in the denominator. When there is but one dimension among the items, μ takes on a value of 1.00. When there are more dimensions among the items, μ takes on values less than 1.00 and has a lower limit somewhere above 0.00, depending on the situation.

We also computed values of μ for each model in our investigation. We observe in Table 1 that μ is relatively independent of both n and the communality of the items. It is true that μ increases somewhat as

TABLE 1

Values of Average Interitem Correlation $\bar{r}_{ij}$, Coefficient α, Index of Homogeneity μ and Its Estimate $\hat{\mu}$ for Various Models of Items. Models Vary by Number of Factors r Among Items, Number L of Factors Determining a Variable, Communality h_i^2 Each Variable, Number of Replications I if a Basic Set of Items in the Model, and the Number n of Items

Model Parameters						Statistics			
r	L	h_i^2	$\frac{r}{L}$	I	n	$\bar{r}_{ij}$	α	μ	$\hat{\mu}$
1	1	.3	1	19	19	.30	.89	1.00	1.13
1	1	.3	1	37	37	.30	.94	1.00	1.06
1	1	.9	1	19	19	.90	.99	1.00	1.01
1	1	.9	1	37	37	.90	1.00	1.00	1.00
2	1	.3	2	19	38	.15	.87	.49	.55
2	1	.3	2	37	74	.15	.93	.49	.53
2	1	.9	2	19	38	.45	.97	.49	.49
2	1	.9	2	37	74	.45	.93	.49	.49
3	1	.3	3	9	27	.09	.73	.31	.40
3	1	.3	3	17	51	.10	.84	.32	.37
3	1	.9	3	9	27	.28	.91	.31	.31
3	1	.9	3	17	51	.29	.95	.32	.32
4	1	.3	4	9	36	.07	.73	.23	.30
4	1	.3	4	17	68	.07	.84	.24	.27
4	1	.9	4	9	36	.21	.90	.23	.23
4	1	.9	4	17	68	.21	.95	.24	.24
4	2	.3	6	5	30	.14	.84	.48	.64
4	2	.3	6	9	54	.15	.90	.49	.58
4	2	.9	6	5	30	.43	.96	.48	.49
4	2	.9	6	9	54	.44	.98	.49	.49
5	1	.3	5	5	25	.05	.57	.17	.26
5	1	.3	5	9	45	.05	.72	.18	.23
5	1	.9	5	5	25	.15	.82	.17	.17
5	1	.9	5	9	45	.16	.90	.18	.18
5	2	.3	10	3	30	.11	.79	.38	.55
5	2	.3	10	5	50	.12	.87	.39	.48
5	2	.9	10	3	30	.34	.94	.38	.39
5	2	.9	10	5	50	.35	.96	.39	.39
5	3	.3	10	3	30	.18	.86	.59	.81
5	3	.3	10	5	50	.18	.92	.59	.73
5	3	.9	10	3	30	.53	.97	.59	.60
5	3	.9	10	5	50	.53	.98	.59	.60
6	1	.3	6	5	30	.04	.56	.14	.22
6	1	.3	6	9	54	.05	.72	.15	.19
6	1	.9	6	5	30	.12	.81	.14	.14
6	1	.9	6	9	54	.14	.89	.15	.15
6	2	.3	15	3	45	.10	.83	.32	.43
6	2	.3	15	5	75	.10	.89	.32	.39
6	2	.9	15	3	45	.29	.95	.32	.32
6	2	.9	15	5	75	.29	.97	.32	.33
6	3	.3	20	2	40	.15	.87	.49	.66
6	3	.3	20	3	60	.15	.91	.49	.61
6	3	.9	20	2	40	.44	.97	.49	.50
6	3	.9	20	3	60	.44	.98	.49	.50
6	4	.3	15	3	45	.20	.92	.66	.89
6	4	.3	15	5	75	.20	.95	.66	.77
6	4	.9	15	3	45	.59	.98	.66	.67
6	4	.9	15	5	75	.60	.99	.66	.67

the number of factors loading on an item increases, and in this μ reflects the high degree of overlap (but not of unidimensionality) among the items. In this case, however, μ does not appear to be as strongly influenced as coefficient alpha, i.e., yielding large values.

Unfortunately because computing μ depends upon knowledge of the communalities of the items, μ cannot be applied in practical situations where these communalities are unknown. But it is possible to obtain an estimate of μ using estimates of the communalities in place of the true communalities in the formula for μ. In our own data analysis we estimated the communality of an item by using the squared multiple correlation R_i^2 for predicting an item in the test from all the other items in the test. This is not a difficult task if one has a computer program to compute the inverse of the correlation matrix for the items. If S_i^2 is the reciprocal of the ith diagonal element of the inverse of the correlation matrix, then $R_i^2 = 1 - S_i^2$. In computing such estimates one must make certain that no items that are exact linear combinations of the other items are included in the test, otherwise the matrix inversion will not be possible and the estimates not obtainable. Since R_i^2 is a lower-bound estimator of the communality, $\hat{\mu}$ computed using R_1^2 as an estimate of the ith item's communality will tend to overestimate μ. Consequently, in some instances $\hat{\mu}$ can take on values greater than 1.00. In comparing $\hat{\mu}$ with μ in our data we observed that the estimate of μ by $\hat{\mu}$ improved as the communality of the items increases. Although we did not investigate them, other estimates of μ based on more exact estimates of the communalities (given, say, by a factor analysis) may be more accurate.

Although we did not study it in our data analysis because it came to us as an afterthought, another index similar to μ can be computed as an index of item homogeneity. Recall that $\bar{r}_{ij}$ is affected by the communalities of the items. If the correlations $|r_{ij}|$ are first corrected for communality (by dividing them by the product of the square roots of their respective item communalities much as you would perform the correction for attenuation) and then the resulting values averaged, one gets an index of homogeneity not affected by the communalities. Such an index also takes on values between .00 and 1.00 with 1.00 indicating unidimensionality.

Conclusions

The process of constructing homogeneous tests by examining item-total-score correlations and the value of coefficient alpha computed for the composite test score is not a fool-proof method. Both high item-total-score correlations and high coefficient alpha can be ob-

tained in situations involving heterogeneous items. Perhaps the test constructor would be better advised to look both at the raw correlations among his items and perform a factor analysis of these correlations to see how tenable the notion of homogeneity is for his items. While high reliability for a composite score is a desired goal, reporting the values of coefficient alpha for the composite should not be regarded as offering sufficient proof for the homogeneity of the test items. If appropriate, an index of test homogeneity such as $\hat{\mu}$ should also be reported to establish the homogeneity of the composite test.

We also wish to point out that the traditional method of constructing homogeneous tests using item-total-score correlations and coefficient alpha may be giving way to procedures that rely more and more upon factor analysis. Armor (1974), for example, recommends performing principal components analysis and then rotating to simple structure to isolate more homogeneous subscales. (This will work if the items fall out in independent clusters, but not if items load substantially on two or more common factors, which can occur in simple structure as our examples show). Armor also suggests using the factor scores of rotated factors to represent the underlying dimensions measured, a procedure that would not depend upon the items' being homogeneous. Allen (1974) recommends finding factor scores for the first canonical factor for the items. His emphasis, however, is on scaling to maximize reliability and not necessarily homogeneity. Principal components, weighted or unweighted, may not represent anything theoretically meaningful or homogeneous. But in any case we believe a greater use of factor analysis in the construction of tests would be of benefit in the development of homogeneous measures.

Recent developments in the technique of confirmatory factor analysis would allow the test constructor to test hypotheses about the unidimensionality of a set of items as well as hypotheses about the presence of various prespecified factors among the items. For a review of the technique of confirmatory factor analysis along with references to related articles in the literature the reader is referred to Mulaik (1975).

REFERENCES

Allen, M. P. Construction of composite measures by the canonical factor-regression method. In H. L. Costner (Ed.), *Sociological methodology* 1973–1974. San Francisco: Jossey-Bass, 1974.

Armor, D. J. Theta reliability and factor scaling. In H. L. Costner (Ed.), *Sociological methodology* 1973–1974. San Francisco: Jossey-Bass, 1974.

Crano, W. D. and Marilynn B. Brewer. *Principles of research in social psychology*. New York: McGraw-Hill, 1973.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297–334.

Cronbach, L. J. *Essentials of psychological testing* (2nd ed.). New York: Harper and Row, 1960.

Edwards, A. L. *Techniques of attitude scale construction.* New York: Appleton, Century, Crofts, 1957.

Gulliksen, H. *Theory of mental tests.* New York: Wiley, 1950.

Lord, F. M. and M. R. Novick. *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley, 1968.

Nunnally, J. C. *Introduction to psychological measurement.* New York: McGraw-Hill, 1970.

Standards for educational and psychological tests and manuals. Washington, D. C.: American Psychological Association, 1966.

Upshaw, H. S. Attitude measurement. In H. M. Blalock, Jr. and A. B. Blalock (Eds.), *Methodology in social research.* New York: McGraw-Hill, 1968.