MY CURRENT THOUGHTS ON COEFFICIENT ALPHA AND SUCCESSOR PROCEDURES

LEE J. CRONBACH
Stanford University

Editorial Assistance by
RICHARD J. SHAVELSON
Stanford University

In 1997, noting that the 50th anniversary of the publication of "Coefficient Alpha and the Internal Structure of Tests" was fast approaching, Lee Cronbach planned what have become the notes published here. His aim was to point out the ways in which his views on coefficient alpha had evolved, doubting now that the coefficient was the best way of judging the reliability of an instrument to which it was applied. Tracing in these notes, in vintage Cronbach style, his thinking before, during, and after the publication of the alpha paper, his "current thoughts" on coefficient alpha are that alpha covers only a small perspective of the range of measurement uses for which reliability information is needed and that it should be viewed within a much larger system of reliability analysis, generalizability theory.

Keywords: *coefficient alpha; reliability; internal consistency; generalizability theory; variance components*

Where the accuracy of a measurement is important, whether for scientific or practical purposes, the investigator should evaluate how much random error affects the measurement. New research may not be necessary when a

The project could not have been started without the assistance of Martin Romeo Shim, who helped me not only with a reexamination of the 1951 paper but with various library activities needed to support some of the statements in these notes. My debt is even greater to Shavelson for his willingness to check my notes for misstatements and outright errors of thinking, but it was understood that he was not to do a major editing. He supported my activity, both psychologically and concretely, and I thank him.

Educational and Psychological Measurement, Vol. 64 No. 3, June 2004 391-418
DOI: 10.1177/0013164404266386

procedure has been studied enough to establish how much error it involves. But with new measures, or measures being transferred to unusual conditions, a fresh study is in order. Sciences other than psychology have typically summarized such research by describing a margin of error; a measure will be reported, followed by a plus or minus sign and a numeral that is almost always the standard error of measurement (which will be explained later).

The alpha formula is one of several analyses that may be used to gauge the reliability (i.e., accuracy) of psychological and educational measurements. This formula was designed to be applied to a two-way table of data where rows represent persons (*p*) and columns represent scores assigned to the person under two or more conditions (*i*). *Condition* is a general term often used where each column represents the score on a single item within a test. But it may also be used, for example, for different scorers when more than one person judges each article and any scorer treats all persons in the sample. Because the analysis examines the consistency of scores from one condition to another, procedures like alpha are known as internal consistency analyses.

Origin and Purpose of These Notes

My 1951 Article and Its Reception

In 1951, I published an article entitled, "Coefficient Alpha and the Internal Structure of Tests." The article was a great success and was cited frequently [no less than 5,590 times].[1] Even in recent years, there have been approximately 325 social science citations per year.[2]

The numerous citations to my article by no means indicate that the person who cited it had read it, and does not even demonstrate that he or she had looked at it. I envision the typical activity leading to the typical citation as beginning with a student laying out his research plans for a professor or submitting a draft report, and it would be the professor's routine practice to say, wherever a measuring instrument was used, that the student ought to check the reliability of the instrument. To the question, "How do I do that?" the professor would suggest using the alpha formula because the computations are well within the reach of almost all students undertaking research and because the calculation can be performed on data the student will routinely collect. The professor might write out the formula or simply say, "You can look it up." The student would find the formula in many textbooks that would be likely to give the 1951 article as a reference, so the student would copy that reference and add one to the citation count. There would be no point for him or her to try to read the 1951 article, which was directed to a specialist audience. And the professor who recommended the formula may have been born well after 1951 and not only be unacquainted with the article but uninterested in the

debates about 1951 conceptions that had been given much space in it. (The citations are not all from nonreaders; throughout the years, there has been a trickle of articles discussing alpha from a theoretical point of view and sometimes suggesting interpretations substantially different from mine. These articles did little to influence my thinking.)

Other signs of success: There were very few later articles by others criticizing parts of my argument. The proposals or hypotheses of others that I had criticized in my article generally dropped out of the professional literature.

A 50th Anniversary

In 1997, noting that the 50th anniversary of the publication was fast approaching, I began to plan what has now become these notes. If it had developed into a publishable article, the article would clearly have been self-congratulatory. But I intended to devote most of the space to pointing out the ways my own views had evolved; I doubt whether coefficient alpha is the best way of judging the reliability of the instrument to which it is applied.

My plan was derailed when various loyalties impelled me to become the head of the team of qualified and mostly quite experienced investigators who agreed on the desirability of producing a volume (Cronbach, 2002) to recognize the work of R. E. Snow, who had died at the end of 1997.

When the team manuscript had been sent off for publication as a book, I might have returned to alpha. Almost immediately, however, I was struck by a health problem that removed most of my strength, and a year later, when I was just beginning to get back to normal strength, an unrelated physical disorder removed virtually all my near vision. I could no longer read professional writings and would have been foolish to try to write an article of publishable quality. In 2001, however, Rich Shavelson urged me to try to put the thoughts that might have gone into the undeveloped article on alpha into a dictated memorandum, and this set of notes is the result. Obviously, it is not the scholarly review of uses that have been made of alpha and of discussions in the literature about its interpretation that I intended. It may nonetheless pull together some ideas that have been lost from view. I have tried to present my thoughts here in a nontechnical manner with a bare minimum of algebraic statements, and I hope that the material will be useful to the kind of student who in the past was using the alpha formula and citing my 1951 article.

My Subsequent Thinking

Only one event in the early 1950s influenced my thinking: Frederick Lord's (1955) article in which he introduced the concept of randomly parallel tests. The use I made of the concept is already hinted at in the preceding section.

A team started working with me on the reliability problem in the latter half of the decade, and we developed an analysis of the data far more complex than the two-way table from which alpha is formed. The summary of that thinking was published in 1963, but is beyond the scope of these notes. The lasting influence on me was the appreciation we developed for the approach to reliability through variance components, which I shall discuss later.[3]

From 1970 to 1995, I had much exposure to the increasingly prominent, statewide assessments and innovative instruments using samples of student performance. This led me to what is surely the main message to be developed here. Coefficients are a crude device that does not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement, as I discuss later.

Conceptions of Reliability

The Correlational Stream

Emphasis on individual differences. Much early psychological research, particularly in England, was strongly influenced by the ideas on inheritance suggested by Darwin's theory of Natural Selection. The research of psychologists focused on measures of differences between persons. Educational measurement was inspired by the early studies in this vein and it, too, has given priority to the study of individual differences, that is, this research has focused on person differences.

When differences were being measured, the accuracy of measurement was usually examined. The report has almost always been in the form of a reliability coefficient. The coefficient is a kind of correlation with a possible range from 0 to 1.00. Coefficient alpha was such a reliability coefficient.

Reliability seen as consistency among measurements. Just what is to be meant by reliability was a perennial source of dispute. Everyone knew that the concern was with consistency from one measurement to another, and the conception favored by some authors saw reliability as the correlation of an instrument with itself. That is, if, hypothetically, we could apply the instrument twice and on the second occasion have the person unchanged and without memory of his first experience, then the consistency of the two identical measurements would indicate the uncertainty due to measurement error, for example, a different guess on the second presentation of a hard item. There were definitions that referred not to the self-correlation but to the correlation of parallel tests, and parallel could be defined in many ways (a topic to which I shall return). Whatever the derivation, any calculation that did not directly fit the definition was considered no better than an approximation. As no for-

mal definition of reliability had considered the internal consistency of an instrument as equivalent to reliability, all internal consistency formulas were suspect. I did not fully resolve this problem; I shall later speak of developments after 1951 that give a constructive answer. I did, in 1951, reject the idealistic concept of a self-correlation, which at best is unobservable; parallel measurements were seen as an approximation.

The split-half technique. Charles Spearman, just after the start of the 20th century, realized that psychologists needed to evaluate the accuracy of any measuring instrument that they used. Accuracy would be naively translated as the agreement among successive measures of the same thing by the same technique. But repeated measurement is suspect because participants learn on the first trial of an instrument and, in an ability test, are likely to earn better scores on later trials.

Spearman, for purposes of his own research, invented the split-half procedure in which two scores are obtained from a single testing by scoring separately the odd-numbered items and the even-numbered items.[4] This is the first of the internal consistency procedures, of which coefficient alpha is a modern exemplar. Thus, with a 40-item test, Spearman would obtain total scores for two 20-item half-tests, and correlate the two columns of scores. He then proposed a formula for estimating the correlation expected from two 40-item tests.

In the test theory that was developed to provide a mathematical basis for formulas like Spearman's, the concept of true score was central. Roughly speaking, the person's true score is the average score he or she would obtain on a great number of independent applications of the measuring instrument.

The problem of multiple splits. Over the years, many investigators proposed alternative calculation routines, but these either gave Spearman's result or a second result that differed little from that of Spearman; we need not pursue the reason for this discrepancy.

In the 1930s, investigators became increasingly uncomfortable with the fact that comparing the total score from Items 1, 3, 5, and so on with the total on Items 2, 4, 6, and so on gave one coefficient, but that contrasting the sum of scores on Items 1, 4, 5, 8, 9, and so on with the total on Items 2, 3, 6, 7, 10 and so on would give a different numerical result. Indeed, there were a vast number of such possible splits of a test, and therefore any split-half coefficient was, to some degree, incorrect.

In the period from the 1930s to the late 1940s, quite a number of technical specialists had capitalized on new statistical theory being developed in England by R. A. Fisher and others, and these authors generally presented a formula whose results were the same as those from the alpha formula. Independent of these advances, which were almost completely unnoticed by persons using measurement in the United States, Kuder and Richardson developed a

set of internal consistency formulas that attempted to cut through the confusion caused by the multiplicity of possible splits. They included what became known as K-R 20, which was mathematically a special case of alpha that applied only to items scored one and zero. Their formula was widely used, but there were many articles questioning its assumptions.

Evaluation of the 1951 article. My article was designed for the most technical of publications on psychological and educational measurement, *Psychometrika*. I wrote a somewhat encyclopedic article in which I not only presented the material summarized above but reacted to a number of publications by others that had suggested alternative formulas based on a logic other than that of alpha, or commenting on the nature of internal consistency. This practice of loading an article with a large number of thoughts related to a central topic was normal practice and preferable to writing half a dozen articles on each of the topics included in the alpha article. In retrospect, it would have been desirable for me to write a simple article laying out the formula, the rationale and limitations of internal consistency methods, and the interpretation of the coefficients the formula yielded. I was not aware for some time that the 1951 article was being widely cited as a source, and I had moved on once the article was published to other lines of investigation.

One of the bits of new knowledge I was able to offer in my 1951 article was a proof that coefficient alpha gave a result identical with the average coefficient that would be obtained if every possible split of a test were made and a coefficient calculated for every split. Moreover, my formula was identical to K-R 20 when it was applied to items scored one and zero. This, then, made alpha seem preeminent among internal consistency techniques.

I also wrote an alpha formula that may or may not have appeared in some writing by a previous author, but it was not well known. I proposed to calculate alpha as

$$\left(\frac{k}{k-1}\right)\left(1-\frac{\sum s_i^2}{s_t^2}\right).$$

Here, k stands for the number of conditions contributing to a total score, and s is the standard deviation, which students have learned to calculate and interpret early in the most elementary statistics course. There is an s_i for every column of a $p \times i$ layout (see Table 1a) and an s_t for the column of total scores (usually test scores). The formula was something that students having an absolute minimum of technical knowledge could make use of.

Not only had equivalent formulas been presented numerous times in the psychological literature, as I documented carefully in the 1951 article, but the fundamental idea goes far back. Alpha is a special application of what is called the intraclass correlation, which originated in research on marine pop-

Table 1a
Person × Item Score (X_{pi}) Sample Matrix

	Item						
Person	1	2	. . .	*i*	. . .	*k*	Sum or Total
1	X_{11}	X_{12}	. . .	X_{1i}	. . .	X_{1k}	X_1
2	X_{21}	X_{22}	. . .	X_{2i}	. . .	X_{2k}	X_2
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
p	X_{p1}	X_{p2}	. . .	X_{pi}	. . .	X_{pk}	X_p
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
n	X_{n1}	X_{n2}	. . .	X_{ni}	. . .	X_{nk}	X_n

Note. Table added by the editor.

ulations where statistics were being used to make inferences about the laws of heredity.[5] R. A. Fisher did a great deal to explicate the intraclass correlation and moved forward into what became known as the analysis of variance. The various investigators who applied Fisher's ideas to psychological measurement were all relying on aspects of analysis of variance, which did not begin to command attention in the United States until about 1946.[6] Even so, to make so much use of an easily calculated translation of a well-established formula scarcely justifies the fame it has brought me. It is an embarrassment to me that the formula became conventionally known as Cronbach's α.

The label alpha, which I applied, is also an embarrassment. It bespeaks my conviction that one could set up a variety of calculations that would assess properties of test scores other than reliability, and alpha was only the beginning. For example, I thought one could examine the consistency among rows of the matrix mentioned above (see Table 1a) to look at the similarity of people in the domain of the instrument. This idea produced a number of provocative ideas, but the idea of a coefficient analogous to alpha proved to be unsound (Cronbach & Gleser, 1953).

My article had the virtue of blowing away a great deal of dust that had grown up out of attempts to think more clearly about K-R 20. So many articles tried to offer sets of assumptions that would lead to the result that there was a joke that "deriving K-R 20 in new ways is the second favorite indoor sport of psychometricians." Those articles served no function once the general applicability of alpha was recognized. I particularly cleared the air by getting rid of the assumption that the items of a test were unidimensional, in the sense that each of them measured the same common type of individual difference, along with, of course, individual differences with respect to the specific content of items. This made it reasonable to apply alpha to the typical

tests of mathematical reasoning, for example, where many different mental processes would be used in various combinations from item to item. There would be groupings in such a set of items, but not enough to warrant formally recognizing the groups in subscores.

Alpha, then, fulfilled a function that psychologists had wanted fulfilled since the days of Spearman. The 1951 article and its formula thus served as a climax for nearly 50 years of work with these correlational conceptions.

It would be wrong to say that there were no assumptions behind the alpha formula (e.g., independence), but the calculation could be made whenever an investigator had a two-way layout of scores with persons as rows and columns for each successive independent measurement.[7] This meant that the formula could be applied not only to the consistency among items in a test but also to agreement among scorers of a performance test and the stability of performance of scores on multiple trials of the same procedure, with somewhat more trust than was generally defensible.

The Variance-Components Model

Working as a statistician in an agricultural research project station, R. A. Fisher designed elaborate experiments to assess the effects on growth and yield of variations in soil, fertilizer, and the like. He devised the analysis of variance as a way to identify which conditions obtained superior effects. This analysis gradually filtered into American experimental psychology, where Fisher's *F* test enters most reports of conclusions. A few persons in England and Scotland, who were interested in measurement, did connect Fisher's method with questions about reliability of measures, but this work had no lasting influence. Around 1945, an alternative to analysis of variance was introduced, and this did have an influence on psychometrics.

In the middle 1940s, a few mathematical statisticians suggested a major extension of Fisherian thinking into new territory. Fisher had started with agricultural research and thought of environmental conditions as discrete choices. A study might deal with two varieties of oats, or with several kinds of fertilizer, which could not be considered a random sample from a greater array of varieties. Fisher did consider plots to be sampled from an array of possible plots. That is, he would combine Species A with Fertilizer 1 and measure the results in some number of scattered areas. Similar samples of plots were used for each of the other combinations of species and fertilizer.

In the postwar literature, it was suggested that one or both factors in a two-way design might be considered random. This opened the way for a method that reached beyond what Fisher's interpretation offered. I have already mentioned the sampling of persons and the sampling of items or tasks, which can be analyzed with the new components-of-variance model, as will be seen.

Burt, working in London and subject to the influence of Fisher, had carried the variance approach in the direction that became generalizability (G) theory, with alpha as a simplified case (Cronbach, Gleser, Nanda, & Rajaratnam, 1972).[8] His notes for students in the 1930s were lost during World War II, and his ideas only gradually became available to Americans in articles where students had applied his methods. In 1951, Burt's work was unknown to U.S. psychometricians.

Basics of Alpha

We obtain a score X_{pi} for person p by observing him in condition i. The term *condition* is highly general, but most often in the alpha literature it refers either to tests or to items, and I shall use the symbol i. The conditions, however, might be a great variety of social circumstances, and it would very often be raters of performance or scorers of responses. If the persons are all observed under the same condition, then the scores can be laid out in a column with persons functioning as rows; and when scores are obtained for two or more conditions, adding the columns for those conditions gives the score matrix (see Table 1a).[9]

We usually think of a set of conditions i with every person having a score on the first condition, on the second condition, and so on, although if there is an omission we will generally enter a score of 0 or, in the case of the scorer failing to mark the article, we will have to treat this as a case of missing data. The alternative, however, is where each person is observed under a different series of conditions. The obvious example is where person p is evaluated on some personality trait by acquaintances, and the set of acquaintances varies from person to person, possibly with no overlap. Then there is no rational basis for assigning scores on the two persons to the same column. Formally, the situation where scores are clearly identified with the same condition i is called a crossed matrix because conditions are crossed with persons. In the second situation, there is a different set of conditions for each person; therefore, we may speak of this as a nested design because raters are nested within the person. Virtually all the literature leading down to the alpha article has assumed a crossed design, although occasional side remarks will recognize the possibility of nesting. Note that we also have a nested design when different questions are set for different persons, which can easily happen in an oral examination and may happen in connection with a portfolio.

Second, a distinction is to be made between the sample matrix of actual observations (see Table 1a) and the infinite matrix (see Table 1b) about which one wishes to draw conclusions. (I use the term *infinite* because it is likely to be more familiar to readers than the technical terms preferred in mathematical discourse.) We may speak of the population-universe matrix for a concep-

tion where an infinite number of persons all in some sense of the same type respond to an infinite universe of conditions, again of the same type.[10] The matrix of actual data could be described as representing a sample of persons crossed with a sample of conditions, but it will suffice to speak of the sample matrix. The alpha literature and most other literature prior to 1951 assumed that the sample matrix and the population matrix were crossed. Mathematically, it is easy enough to substitute scores from a nested sample matrix by simply taking the score listed first for each as belonging in column 1, but this is not the appropriate analysis.

All psychometric theory of reliability pivots on the concept of true score. (In G Theory, this is renamed "Universe Score", but we need not consider the reasons here.) The true score is conceptualized as the average score the person would reach if measured an indefinitely large number of times, all measurements being independent, with the same or equivalent procedures [average over $k \rightarrow \infty$; see Table 1b]. The difference between the observed score and the person's true score is the error. It is uncorrelated from one measurement to another—another statement of the independence principle. The concept of error is that random errors are unrelated to the true score and have a mean of zero over persons, or over repeated measurements.

The conception of true score is indefinite until equivalent is endowed. This did not occur until Lord (1955) cataloged various degrees in which parallel tests might resemble one another. At one extreme, there could be parallel tests where the content of Item 5 appeared in a second form of the instrument in other wording as, let us say, Item 11. That is to say, the specific content of the two tests, as well as the general dimensions running through many items, were duplicated. At the other extreme were random-parallel tests, where each test was (or could reasonably be regarded as) a random sample from a specified domain of admissible test items. It was the latter level of parallelism that seemed best to explain the function of coefficient alpha; it measured the consistency of one random sample of items with other such samples from the same domain.

A rather obvious description of the accuracy with which an instrument measures individual differences in the corresponding true score is the correlation of the observed score with the true score. Coefficient alpha is essentially equal to the square of that correlation. (The word *essentially* is intended to glide past a full consideration of the fact that each randomly formed instrument will have a somewhat different correlation with the true score.) Reliability formulas developed with assumptions rather different from those entering alpha are also to be interpreted as squared correlations of observed score with the corresponding true score, so alpha is on a scale consistent with tradition. It might seem logical to use the square root of alpha in reports of reliability findings, but that has never become the practice.

Table 1b
Person × Item Score (X_{pi}) Infinite (Population-Universe) Matrix

	Item					
Person	1	2	. . .	i	. . .	$k \to \infty$
1	X_{11}	X_{12}	. . .	X_{1i}	. . .	X_{1k}
2	X_{21}	X_{22}	. . .	X_{2i}	. . .	X_{2k}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
p	X_{p1}	X_{p2}	. . .	X_{pi}	. . .	X_{pk}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
$n \to \infty$	X_{n1}	X_{n2}	. . .	X_{ni}	. . .	X_{nk}

Note. Table added by the editor.

The observed score is regarded as the sum of the true score and a random error. That statement and the independence assumption, which has its counterpart in the development of other reliability formulas, lead to the simple conclusion that the variance of observed scores is the sum of the error variance and the true score variance. It will be recalled that variance is really the square of the standard deviation. Each individual taking a test has a particular true score, which I may label *T*, and the true scores have a variance. The observed score has been broken into fractions, its presenting error and true score. We may, therefore, interpret alpha as reporting the percentage of the observed individual differences (as described in their variance) that is attributable to true variance in the quality measured by this family of randomly parallel tests.[11]

In thinking about reliability, one can distinguish between the coefficient generated from a single set of *n* persons and *k* items, or about the value that would be obtained using an exceedingly large sample and averaging coefficients over many random drawings of items. The coefficient calculated from a finite sample is to be considered an estimate of the population value of the coefficient. Little interest attaches to the consistency among scores on a limited set of items and a particular group of people. This is the usual consideration in research where data from the sample are used to infer relations in the population.

In the history of psychometric theory, there was virtually no attention to this distinction prior to 1951, save in the writings of British-trained theorists. My 1951 article made no clear distinction between results for the sample and results for the population. It was not until Lord's (1955) explicit formulation of the idea of random parallel tests that we began to write generally about the sampling, not only of persons, but of items. This two-way sampling had no counterpart in the usual thinking of psychologists. No change in procedures

was required, but writing had to become more careful to recognize the sample-population distinction.

The alpha formula is constructed to apply to data where the total score in a row of Table 1a will be taken as the person's observed score. An equivalent form of the calculation applicable when the average is to be taken as the raw score yields the same coefficient. The alpha coefficient also applies to composites of k conditions. When an investigator wants to know what would happen if there were k' conditions, the solution known as the Spearman-Brown Formula applies.

My 1951 article embodied the randomly parallel-test concept of the meaning of true score and the associated meaning of reliability, but only in indefinite language. Once Lord's (1955) statement was available, one could argue that alpha was almost an unbiased estimate of the desired reliability for this family of instruments. The *almost* in the preceding sentence refers to a small mathematical detail that causes the alpha coefficient to run a trifle lower than the desired value.

This detail is of no consequence and does not support the statement made frequently in textbooks or in articles that alpha is a lower value to the reliability coefficient. That statement is justified by reasoning that starts with the definition of the desired coefficient as the expected consistency among measurements that had a higher degree of parallelism than the random parallel concept implied. We might say that my choice of the true score as the expected value over random parallel tests and the coefficient as the consistency expected among such tests is an assumption of my argument.

There is a fundamental assumption behind the use of alpha, an assumption that has its counterpart in many other methods of estimating reliability. The parts of the test that identify columns in the score table (see Table 1a) must be independent in a particular sense of the word. The parts are not expected to have zero correlations. But it is expected that the experience of responding to one part (e.g., one item) will not affect performance on any subsequent item. The assumption, like all psychometric assumptions, is unlikely to be strictly true. A person can become confused on an item that deals with, say, the concept of entropy, and have less confidence when he encounters a later item again introducing the word. There can be fatigue effects. And, insofar as performance on any one trial is influenced by a person's particular state at the time, the items within that trial are, to some degree, influenced by that state.

One can rarely assert, then, that violations of independence are absent, and it is burdensome (if not impossible) to assess the degree and effect of nonindependence.[12] One therefore turns to a different method or makes a careful judgment as to whether the violation of the assumption is major or minor in its consequence. If the problem is minor, one can report the coefficient with a word of caution as to the reasons for accepting it and warning that the nonindependence will operate to increase such coefficients by at

least a small amount. When the problem is major, alpha simply should not be used. An example is a test given with a time limit so that an appreciable number of students stop before reaching the last items. Their score on these items not reached is inevitably zero, which raises the within-trial correlation in a way that is not to be expected of the correlations across separately timed administrations.

The alpha formula is not strictly appropriate for many tests constructed according to a plan that allocates some fraction of the items to particular topics or processes. Thus, in a test of mathematical reasoning, it may be decided to make 20% of the items around geometric shapes. The several forms of the test that could be constructed by randomly sampling geometric items will be higher than the correlation among items in general. The tests are not random parallel.

When the distribution of content is specified formally, it is possible to develop a formula to fit those specifications, but this is difficult and not appropriate when the allocation of items is more impressionistic than strict. In such an instance, one is likely to fall back on alpha and to recognize in the discussion that the coefficient underestimates the expected relationship between observed scores and true scores formed from tests, all of which satisfy the constraint. That is to say, alpha tends to give too low a coefficient for such tests. An extension of alpha to fit specifically the stratified parallel test (sometimes called stratified alpha; Cronbach, Schonemann, & McKie, 1965) can be based on the battery reliability formula that Jackson and Ferguson published in an obscure monograph.[13]

Variance Components and Their Interpretation

I no longer regard the alpha formula as the most appropriate way to examine most data. Over the years, my associates and I developed the complex generalizability (*G*) theory (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach et al., 1972; see also Brennan, 2001; Shavelson & Webb, 1991), which can be simplified to deal specifically with a simple two-way matrix and produce coefficient alpha. From 1955 to 1972, we exploited a major development in mathematical statistics of which psychologists were unaware in the early 1950s. Subsequently, I had occasion to participate in the analysis of newer types of assessments, including the use of performance samples where the examinee worked on a complex realistic problem for 30 minutes or more, and as few as four such tasks might constitute the test (Cronbach, Linn, Brennan, & Haertel, 1997). The performance was judged by trained scorers so that the data generated could be laid out in a two-way matrix.[14]

Here I sketch out the components of variance approach to reliability focusing on the simplest case where coefficient alpha applies, the Person × Condition data matrix (see Table 1a). Random sampling of persons and conditions (e.g., items, tasks) is a central assumption to this approach.

Giving Sampling a Place in Reliability Theory

Measurement specialists have often spoken of a test as a sample of behavior, but the formal mathematical distinction between sample of persons and populations of persons, or between a sample of tasks and a population [a universe] of tasks, was rarely made in writings on test theory in 1951 and earlier [see discussion of Fisher above]. Nevertheless, the postwar mathematical statistics literature suggested that one or both factors in a two-way design might be considered random. This opened the way for a method, the components of variance method, that reached beyond what Fisher's interpretation offered.[15]

Random sampling, now, is almost invariably an assumption in the interpretation of psychological and educational data where conclusions are drawn, but the reference is to sampling of persons from the population. We are thinking now of a person universe matrix from which one can sample not only rows (persons) but also columns (conditions). Thus, the alpha article flirted with the thought that conditions are randomly sampled from the universe, but this idea did not become explicit until much later. Now, it is most helpful to regard the random sampling of persons as a virtually universal assumption and the random sampling of conditions that provide the data as an assumption of the alpha formula when the result is interpreted as applying to a family of instruments that are no more similar to each other than random samples of conditions would be. Investigators who want to postulate a higher degree of similarity among the composites would find alpha and related calculations underestimating the accuracy of the instrument.

The [random sampling] assumptions just stated are not true in any strict sense, and a naive response would be to say that if the assumptions are violated, the alpha calculations cannot be used. No statistical work would be possible, however, without making assumptions and, as long as the assumptions are not obviously grossly inappropriate to the data, the statistics calculated are used, if only because they can provide a definite result that replaces a hand-waving interpretation. It is possible at times to develop a mathematical analysis based on a more complex set of assumptions, for example, recognizing that instruments are generally constructed according to a plan that samples from domains of content rather than being constructed at random. This is more troublesome in many ways than the analysis based on simple assumptions, but where feasible it is to be preferred.

Components of Variance

In the random model with persons crossed with conditions, it is necessary to recognize that the observed score for person p in condition i (X_{pi}) can be divided into four components, one each for the (*1*) grand mean, (*2*) person

(p), condition (i), and residual consisting of the interaction of person and condition (pi) and random error (e, actually pi,e):

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu).$$

The first of these, the grand mean, μ, is constant for all persons. The next term, $\mu_p - \mu$, is the person's true score (μ_p) expressed as a deviation from the grand mean (μ)—the person effect. The true score, it will be recalled, is the mean that would be expected if the person were tested by an indefinitely large number of randomly parallel instruments drawn from the same universe. (In *G* Theory, it is referred to as the universe score because it is the person's average score over the entire universe of conditions.) The μ_i term represents the average of the scores on item i in the population and is expressed as a deviation from μ—the item effect. The fourth term is the residual consisting of the interaction of person p with item i, which, in a $p \times i$ matrix, cannot be disentangled from random error, e. The residual simply recognizes the departure of the observed score from what would be expected in view of the μ_i level of the item and the person's general performance level, μ_p. (In most writings, the residual term is divided into interaction and error, although in practice it cannot be subdivided because with the usual matrix of scores X_{pi} from a single test administration, there is no way to take such subdivision into account.)

Except for μ, each of the components that enter into an observed score vary from one person to another, one item to another, and/or in unpredictable ways. Recognizing that score components vary, we now come to the critically important equation that decomposes the observed-score variance into its component parts:

$$V(X_{pi}) = V_p + V_i + V_{Res}.^{16}$$

Here, V is a symbol form of the population variance. (In the technical literature, the symbol σ^2 is used.) The term on the left refers to the variation in scores in the extended matrix that includes all persons in the population and all items in the universe [see Table 1b]. It characterizes the extent of variation in performance. The equation states that this variance can be decomposed into three components, hence the name Components of Variance approach.

The first term on the right is the variance among persons, the true-score variance. This is systematic, error-free variance among persons, the stuff that is the purpose and focus of the measurement. This variance component gives rise to consistency of performance across the universe of conditions. The i component of variance describes the extent to which conditions (items, tasks) vary. And the residual represents what is commonly thought of as error of measurement, combining the variability of performance to be expected when an individual can sometimes exceed his norm by gaining insight into a

question and sometimes fall short because of confusion, a lapse of attention, and so forth.

The last equation is only slightly different from the statement made in connection with alpha and more traditional coefficients: The observed variance is the sum of true-score variance and error variance. The *novelty lies in the introduction of the* μ_i. In the long history of psychological measurement that considered only individual differences, the difference in item means is disregarded, having no effect on individual standings when everyone responds to the same items.

Spearman started the tradition of ignoring item characteristics because he felt that the person's position on the absolute score scale was of no interest. He reasoned that the person's score depended on a number of fairly arbitrary conditions, for example, the size and duration of a stimulus such as a light bulb, and on the background, as well as on the physical brightness itself. His main question was whether the persons who were superior at one kind of discrimination were superior at the next kind, and for this he was concerned only with ranks. Psychologists shifted attention from ranks to deviation scores, partly because these are sensitive to the size of differences between individuals in a way that ranks are not, are easier to handle mathematically, and fit into a normal distribution. (For a time, it was believed that nearly all characteristics are normally distributed, as a matter of natural law.) When psychologists and educators began to make standardized tests, some of them tried to use natural units, but this quickly faded out because of the sense that the individual's score depended on the difficulty of the items chosen for the test. The rankings on arithmetic tests could be considered stable from one set of items to another, where the score itself was seen as arbitrary. Consequently, it was the statistics of individual differences observed in tests that received the greatest emphasis.

Nonetheless, the absolute level of the person's performance is of significance in many circumstances. This is especially true in the many educational tests used to certify that the person has performed adequately. The critical score indicating minimal adequate performance is established by careful review of the tasks weighed by experts in the domain of the test. This score is established for the family of tests in general, not separately for each form in turn. When a candidate takes a form for which μ_i is unusually low, the number of examinees passing are reduced for no good reason. Therefore, persons using tests for absolute decisions must be assured that the choice of form does not have a large effect on a person's chances of passing, which means that a low $V\mu_i$ is wanted.

The analysis that generates estimates of the three components is simple. One first performs an analysis of variance, ordinarily using one of the readily available computer programs designed for that purpose. Instead of calculat-

ing F ratios, one converts the mean squares (MS) for rows, columns, and a residual to components of variance. These equations apply:

$$\hat{V}_{Residual} = MS_{Residual}$$
$$\hat{V}_i = (MS_i - MS_{Residual}) / n_p$$
$$\hat{V}_p = (MS_p - MS_{Residual}) / n_i$$

It is to be understood that these components describe the contributions of the three sources to variation in scores at the item level. We are looking not at the decomposition of a particular item but at a typical result, in a sense averaged over many persons and items. These estimates are readily converted to estimates that would apply to test scores and to averages over specified numbers of persons. The components of variance are determined with the assumption that the average of scores in the row (see Table 1a) would lead to the composite score. Specifically, if randomly sampled tests of 20 items are applied, and the average score on the 20 items is reported, then $\hat{V}_{Residual}$ for this average score is 1/20 of $V_{Residual}$ for a single item score. Results reached with that understanding are readily converted to the total score scale. If your interpretation is based on the total scores over 20 items, $\hat{V}_{Residual}$ for this total score is 20 times greater than $\hat{V}_{Residual}$, but I shall stay with averages for observed scores because this keeps formulas a bit simpler.

Interpreting the Variance Components

The output from the analysis of variance is a set of estimates of characteristics of the population-universe matrix [see Table 1b]. The estimates are assumed to apply to any sample matrix. Obviously, they apply to the sample from which they were taken, and, for want of an alternative, the other possible sample matrices are assumed to be similar statistically.

Variance components are generally interpreted by converting them to estimates of the corresponding standard deviations. Thus, the square root of the $\hat{V}_p$ is a standard deviation of the distribution of individuals' true scores, that is to say, the average score they would obtain if they could be tested on all conditions in the universe. One might consider forming a composite instrument by combining many conditions, the usual test score being a prominent example. If the test score is expressed as a per-condition average, then the standard deviation just calculated applies to the true score on such composites. If, however, as is often the case, the total score over conditions is to be used, then the value of the standard deviation must be multiplied by the number of items to put it on the scale of the composite.

The usual rule of thumb for interpreting standard deviations is that two thirds of the scores of persons will fall within one standard deviation of the

mean, and 95% of the persons will fall within two standard deviations of the mean. The standard deviation of true scores gives a clearer picture of the spread of the variable being measured than the standard deviation that is calculated routinely from observed scores, because the effect of random errors of measurement is to enlarge the range of observed scores. Working from the $\hat{V}_p$ indicates whether the variable of interest is spread over much of the possible score scale or is confined to a narrow range.

μ_p is the row mean in the population-universe matrix [see Table 1b], and μ_i is the column mean, that is to say, the population mean for all p under condition i. The variance of column means V_i is therefore the information about the extent to which condition means differ. A standard deviation may be formed and interpreted just as before, this time with the understanding that the information refers to the spread of the items (or, more generally, the spread of the conditions) and not the spread of persons. The standard deviation for condition means gives a direct answer to questions such as the following: Do the items in this ability test present similar difficulty? Do the statements being endorsed or rejected in a personality inventory have similar popularity? Do some of the persons scoring this performance exercise tend to give higher scores than others? It is important to reiterate that we are concerned with characteristics of the population and universe. We are arriving at a statement about the probable spread in other samples of conditions that might be drawn from the universe. Where we have a composite of k' single conditions, the estimated variance for μ_i must be divided by k' (i.e., $\hat{V}_i$ / k'). The standard deviation is reduced correspondingly, and if the composite is being scored by adding the scores on the elements, the estimated value of $\hat{V}_i$ is k' times as large as that for single conditions.

A comparatively large value of this standard deviation raises serious questions about the suitability of an instrument for typical applications. If students are being judged by whether they can reach a level expressed in terms of score units (e.g., 90% of simple calculations), then the student who happens to be given one of the easier tests has a considerable advantage and the test interpreter may get too optimistic an impression of the student's ability. Similarly, when one of a group of scorers is comparatively lenient, the students who are lucky enough to draw that scorer will have an advantage over students who draw one of the others.

To introduce the residual or the RES, it may help to think of a residual score matrix that would be formed by adjusting each X_{pi} by subtracting out μ_p for person p and μ_i for condition i, then adding in the constant (μ) equal to the overall mean of scores in the population. These are scores showing the inconsistency in the individual's performance after you make allowance for his level on the variable being measured, and the typical scores on the conditions in the universe. The residual scores spread around the value of zero. They represent fluctuations in performance, some of which can be explained by systematic causes, and some of which are due to nonrecurrent variation such

as those due to momentary inattention or confusion. A few of the possible systematic causes can be listed:

- In an ability test, the student finds certain subtopics especially difficult and will consistently have a negative residual on such items; for example, the student taking a math test may be confused about tangents, even when he or she is at home with sines and cosines. Deviations can also arise from picking the high-scoring alternative when choosing between attractive options, and also from sheer good or bad luck in guessing.
- In an anxiety inventory, a student who can generally say that he or she has no emotional problems in situation after situation may recognize a timidity about making speeches or otherwise exposing himself or herself to the scrutiny of a group, and thus respond to the related items in a way that deviates from his or her typical response.

Additive Combinations of Variance Components

The interpretation of components gives information about the population-universe matrix, but it is combinations of components that more directly yield answers to the questions of a prospective user of an instrument, including the following: How much do the statistics for the instrument change as k' is increased or decreased? How much greater precision is achieved by using a crossed rather than a nested design for the instrument? How much is the score from a sample of conditions expected to differ from the universe score? How much is the uncertainty about the universe score arising from such errors of measurement?

Adding two or three variance components in an appropriate way estimates the expected observed-score variance for measures constructed by sampling conditions. The word *expected* signifies that we can estimate only for a particular new set of randomly sampled conditions.

I take up first the estimate for nested conditions where different individuals are assessed under different sets of conditions (see Table 2). The most common example is where scores on observations of performance tasks for each individual are assigned by different scorers selected haphazardly from a pool of qualified scorers. The expected observed-score variance here is a weighted sum of all three components. Assume that there are k' conditions and that the average score over conditions will be used: $\hat{V}_{X_{pi}} = \hat{V}_p + \hat{V}_{Res} \ / \ k'$ where the residual consists of three variance components confounded with one another $\hat{V}_i, \hat{V}_{pi}, \hat{e}$. The weight of $\hat{V}_p$ is 1. The other two components (conditions confounded with the *pi* interaction and error) are weighted by $1/k'$. This allows for the fact that as more conditions are combined, random variability of the average decreases. If future observations will be made by means of a crossed design, everyone being observed under the same set of conditions, then the expected observed variance is V_P plus $V_{Res/k'}$. The variation in conditions (*i*) makes no contribution, because everyone is exposed to the

same conditions and all scores are raised or lowered on easy and difficult items (respectively) by a constant amount.

In the crossed $p \times i$ design (see Table 1a), each person is observed under each condition. The most common example is where scores are available for each individual on each item on a test. The expected observed-score variance here (see Table 2) is a weighted sum of V_p and V_{Res}, where $V_{Residual}$ consists of V_{pi}, e. Again, the weight of V_p is 1. The residual is weighted by $1/k'$. A comparison of the residual terms for the nested and crossed design shows that in the nested design, the variance due to conditions cannot be disentangled from the variances due to the person by condition interaction and random error. With a crossed design, condition variance can be disentangled from variance due to the person by condition interaction and error. Consequently, the nested-design residual will be larger than or equal to the crossed-design residual.

The Standard Error

A much more significant report on the measuring instrument is given by the residual (error) variance and its square root, the standard error of measurement (SEM). This describes the extent to which an individual's scores are likely to vary from one testing to another when each measurement uses a different set of conditions. In the nested design, the error variance equals the expected observed score variance as calculated above minus V_p. This leaves us with the weighted sum of the i and residual components of variance, both of which represent sources of error.

The rule of thumb I suggest for interpreting the standard error assumes that errors of measurement for any person are normally distributed, and the standard error tends to be the same in all parts of the range. Both of these assumptions can be questioned. Indeed, when complex analyses are used to estimate a standard error in each part of the range, it is usual for the standard error to show a trend, higher in some ranges of universe [true] scores than others. Here again, we rely on the rule of thumb, because it is impractical to interpret the standard error without them.

Observed scores depart in either direction from the person's universe score. Two thirds of the measurements, according to the usual rule of thumb, fall within one SEM of the universe score, and 95% fall within two SEM. Here we have a direct report on the degree of uncertainty about the person's true level of performance. The figure is often surprisingly large and serves as an important warning against placing heavy weight on the exact score level reached.

For many purposes, a useful scheme is to report scores as a band rather than a single number. Thus, in a profile of interest scores, one would have an array of bands, some spanning a low range and some spanning a high range,

Table 2
Statistics Applying to Two Types of Designs and Two Types of Decisions

	Measurement	
Design	Absolute	Differential
Nested: Conditions (i) within Persons (p) – i:p		
Universe-score variance	V_p	V_p
Expected observed-score variance	$V_p + (V_i + V_{Res})/k'$	$V_p + (V_i + V_{Res})/k'$
Error variance	$(V_i + V_{Res})/k'$	$(V_i + V_{Res})/k'$
Crossed: Conditions (i) crossed with Persons (p) – $p \times i$		
Universe-score variance	V_p	V_p
Expected observed-score variance	$V_p + (V_i + V_{Res})/k'$	$V_p + (V_{Res})/k'$
Error variance	$(V_i + V_{Res})/k'$	$(V_{Res})/k'$

Note. It is assumed that each person responds to a sample of k' conditions and that the score for the person is the average of these scores under separate conditions. If the totals were used instead, the entries in the table would be increased but the patterning would remain the same. The standard error of measurement is the square root of the error variance. The reliability coefficient pertains only to differential measurement and is obtained by dividing the universe-score [true-score] variance by the expected observed-score variance.

but usually with a good many that overlap to a large degree. This discourages emphases on which interest is strongest and encourages attention to the variety of categories in which the person expresses interest.

For a design with conditions (e.g., scorers) nested within persons, the residual or measurement error includes differences in condition means as well as unsystematic (random) variation (due to the $p \times i$ interaction confounded with random error; see Table 2). In this case, we speak about what may be called absolute measurement, where the level of a person's score, and not just his or her standing among peers, is of concern. Many educational applications of tests require a judgment as to whether the examinee has reached a predetermined score level. Examinees are not in competition; all may meet the standard, or none.

For a design with conditions (e.g., items) crossed with persons, the residual or measurement error does not include differences in condition means. So the residual is an index of relative or differential error disentangled from differences in conditions means. In contrast to absolute measurement, this differential measurement is concerned with the relative standing of persons. In selection, when there are a limited number of positions to be allotted, the highest scoring individuals are given preference. Few practical decisions are based directly on such simple rankings, but this is the formulation that permits statistical analysis. It should be noted also that where the correlation between one instrument and another is to be the basis for interpreting data, the interpretation is differential. It was his interest in correlations that led

Spearman originally to define the reliability coefficient so that it applied to differential measurement (which ignores the contribution of variation in μ_i to error). This tradition dominated the literature on reliability down through the alpha article.

Many tests convert the raw score to a different form for use by interpreters. Thus, the raw score on an interest inventory is often expressed as a percentile rank within some reference distribution. There is no way to apply internal consistency analysis directly to such converted scores. One can, however, express the bounds on the probable true score on the raw score scale, as has been illustrated. Then each limit can be rescaled to apply to the new scale. As an illustration, suppose that raw scores 40, 50, and 60 convert to percentile scores 33, 42, and 60, respectively. Then an observed score of 50 converts to a percentile score of 42. If we have established that two thirds of the raw scores fall between 43 and 57, these can be converted to the new scale supplying an asymmetric confidence range running from approximately 37 to 56. Note that the interval is no longer symmetric around the observed score.

Reliability Coefficients

We come now to reliability coefficients estimated with variance components. These coefficients describe the accuracy of the instrument on a 0-to-1 scale; the alpha coefficient fits this description. The assumptions underlying the formulas for estimating variance components are quite similar to the assumptions made in connection with alpha. We discuss here only the analysis of the crossed design, which matches the basis for alpha. The principal change is that because variance components are used to make inferences to the population-universe matrix [see Table 1b] rather than describing the sample, the random sampling of persons and of conditions becomes a formal assumption.

In general, the coefficient would be defined as V_p divided by the expected observed variance. We have seen above that the expected observed variance takes on different values, depending on the design used in data collection. Coefficients differ correspondingly. The alpha coefficient applies to a crossed design implying k conditions. It refers to the accuracy of differential measurement with such data. Computing components of variance has the advantage that an observed-score variance is estimated in terms of k', which may take on any value. Thus, direct calculation of the expected observed variance (with the implied and important standard error) reaches the result for which Spearman-Brown Formula has traditionally been utilized.[17]

As the expected observed variance is larger for a nested design than a crossed design [See Table 2], the coefficient is smaller than that from the crossed design. This is important because an instrument developer often sets up the crossed design in checking the accuracy of the instrument when practi-

cal conditions make it likely that the actual data obtained will have a nested design.

Differential and absolute measurements and reliability. It will be noted that the alpha coefficient is included as one of the statistics reported with differential decisions and not with absolute decisions. A coefficient could be calculated by formal analogy to the entry in the differential column, but it would be meaningless. A coefficient is concerned with individual differences, and those are irrelevant to absolute decisions.

Homogeneity/heterogeneity of samples of conditions. Whereas the topic of homogeneity was the subject of heated discussion in the late 1940s, it has faded from prominence. There are, however, investigators who believe that good psychological measurement will rely on homogeneous instruments, where homogeneity can be thought of as consistency from one condition to another in the ranking of individuals. A contrary position emphasizes that one needs to represent all aspects of the variable that is the focus of measurement, not narrowing it to a single focal topic. An appropriate statistic for evaluating the homogeneity of conditions is the value of the reliability coefficient when k' is set at 1. The value of this coefficient is held down not only by diversity among conditions, but also by the sheer unreliability of an individual's performance in responding many times to the same condition. More advanced techniques, such as factor analysis, can remove much of the ambiguity.

Recommendations

General Observations and Recommendations

I am convinced that the standard error of measurement, defined in accordance with the relevant cell of Table 2, is the most important single piece of information to report regarding an instrument, and not a coefficient. The standard error, which is a report on the uncertainty associated with each score, is easily understood not only by professional test interpreters but also by educators and other persons unschooled in statistical theory, and also to lay persons to whom scores are reported.

There has been a shift in the character of the way measurement is used. The change is obvious in much of educational assessment, where the purpose is to judge individuals or student bodies relative to specified performance standards. Rankings are irrelevant. A similar change is to be seen in screening applicants for employment, where the employer now bears a burden of proof that the choice of a higher scoring individual is warranted, a policy that

seems to work against minority candidates. In making comparisons between candidates, the employer wants to know whether a difference in favor of one of the two would probably be confirmed in another testing. (Questions about the predicted job performance of the candidates are more significant than questions about accuracy of measurement, but inaccurate measurement sets a limit on the accuracy that predictions can obtain.)

The investigator charged with evaluating reliability ought to obtain information on the most prominent sources of potential error. For instruments that make use of judgment of scorers or raters, a simple $p \times i$ design is inadequate. The alpha coefficient, which relies on that design, is appropriate enough for objectively scored tests where items can be considered a sample from the domain. But even in the limited situation contemplated in a $p \times i$ design, the application of the alpha formula does not yield estimates of the three components of variance or the sums listed in Table 2. I cannot consider here data structures in which conditions are classified in more than one way.

In general, a person responsible for evaluating and reporting the accuracy of a measurement procedure ought to be aware of the variety of analyses suggested by Table 2 and include in the report on the instrument information for all of the potential applications of the instrument. Sometimes the investigator will know that the instrument is to be used in correlational research only, in which case a reliability coefficient may be the only report needed. But most instruments lend themselves to more diversified applications. I suggest that the person making judgments about the suitability of an instrument or its purposes, or about the trust that can be placed in observed scores, consider these questions: In my use of the instrument, will I be concerned with the absolute standing of persons, or groups, or the comparative standing?

The choice of a single statistic to summarize the accuracy of an instrument is not the best report that can be made. I recommend that the three separate components of variance be reported. Given this information, the investigator can combine the components or not, according to the competence of his or her likely readership.

Considerations in Conducting a Reliability Study

Aspects of the test plan. The investigator conducting a reliability study should consider a number of points in taking advantage of the information laid out. I write here as if the investigator believes that his or her instrument is likely to be useful in future studies by him or her or by others, and that the investigator is therefore providing guidance for instrumentation in those studies. Of course, the case may be that the investigator is interested in the current set of data and only that set, and has no intention of making further use of the instrument. If so, the investigator will run through these considerations, giving much weight to some and little weight to others in deciding of

the adequacy of the scores for the purpose of that one study. I assume that the investigator is starting with a matrix of scores for persons crossed with conditions, such as are used with the alpha formula.

Independence in sampling. The first step is to judge whether assumptions behind the calculations are seriously violated by the data being used. Violations of the independence assumption can often be regarded as having little consequence, but some violations are serious. The most prominent and frequent misuse of the computations discussed in this article is to apply them to a test where the examinees are unable to complete many items on which they have a reasonable probability of earning a nonzero score. The data may then be used only if it is considered reasonable to truncate the data set, eliminating persons who have too many items not completed, or omitting items toward the end of the set from the calculation. This is a makeshift solution, but it may be necessary.

Heterogeneity of content. Another common difficulty is that conditions fall into psychologically distinct classes, which calls into question the assumption that conditions are randomly sampled. There is no reason to worry about scattered diversity of items, but if, for example, a test in mathematics is planned with some number of geometric-reasoning items and a certain number of numeric reasoning items, the sampling is not random. This type of heterogeneity is not a bar to use of the formulas. It needs only to be recognized that an analysis that does not differentiate between the two classes of items will report a larger standard error than a more subtle analysis.

How the measurement will be used. Decide whether future uses of the instrument are likely to be exclusively for absolute decisions, for differential decisions, or may include both uses (not necessarily in the same study). If either type of decision is unlikely to be made with this instrument in future applications, no further information need be stated for it. Once this decision is made, I recommend that the investigator calculate estimates for the components of variance and combine these to fill in numerical values for the rows of each relevant column of Table 2.

With respect to differential decisions, the standard error from a nested design will be at least a bit larger than the standard error from a crossed design. This larger error, plus the appearance of greater fairness, favors use of crossed designs wherever feasible. However, in large-scale programs such as tests for college admissions, it may seem easy to provide crossed data, when in fact the data are from a nested design. Examinees tested on different dates, or perhaps in different locales, will take different forms of the test and yet be compared with each other. Where it is practical to obtain crossed data for a reliability study, the program itself will always have a nested design. Like-

wise, a crossed design with a small group of scorers is feasible for the reliability study, but the crossing is impractical in operational scoring of the instrument.

Number of conditions for the test. Next, specify the standard error considered acceptable for the purpose of the measurement. Calculate the value of k', which changes the previously calculated standard error. The original value assumed the decisions would be based on responses to k conditions, the new calculation may produce a higher or lower value of k'. Increasing k' to the value just calculated may prove too costly, and a compromise must be made between cost and precision. When a test will be used in a variety of contexts, different users may specify different standard errors as acceptable. Anticipating that problem, the original investigator could well set up a table with several values of the standard error and the corresponding k' required to achieve each one. If the instrument is to be used in correlational research only, it may be easier to specify an acceptable reliability coefficient than a standard error. The equations in the differential column make it simple to convert the acceptable coefficient detailed and acceptable probable error.

Main Message of These Notes

The alpha coefficient was developed out of the history that emphasized a crossed design used for measuring differences among persons. This is now seen to cover only a small perspective of the range of measurement uses for which reliability information is needed. The alpha coefficient is now seen to fit within a much larger system of reliability analysis.

Notes

1. [All Editor's Notes in text, as well as in subsequent endnotes, are in brackets.]

2. [To give some notion of how extraordinary this annual citation frequency is for a psychometric piece, Noreen Webb and I published *Generalizability Theory: A Primer* in 1991. The average number of social science citations over the past 5 years was 11 per year!]

3. [Cronbach, Rajaratnam, & Gleser (1963).]

4. [In "Coefficient Alpha," Cronbach (1951, p. 300) cites both Spearman (1910) and Brown (1910) as providing the first definition of a split-half coefficient.]

5. [As applied to reliability, intraclass correlation is a ratio of true-score (typically person) variance to observed-score variance for a single condition which is composed of true-score variance plus error variance.]

6. The articles by others working with Fisher's ideas employed a number of statistical labels that gave a result identical to my formula but that were unfamiliar to most persons applying measurements. This explains why so little use was made of these formulas. Priority in applying the appropriate intraclass correlation to measurements probably goes to R. W. B. Jackson (Jackson & Ferguson, 1941). So far as I recall, no one had presented the version that I offered in 1951, except for the Kuder-Richardson report, which did not give a general formula.

7. Violation of independence usually makes the coefficient somewhat too large, as in the case where the content of each test form is constrained, for example, by the requirement that 10%

of items in a mathematical reasoning test should be concerned with geometric reasoning. Then, the items can be described as chosen at random within the category specified in the [test] plan, but this is stratified random sampling rather than random sampling. The alpha formula will underestimate the reliability of such instruments (Cronbach, Schonemann, & McKie, 1965).

8. [Cronbach is likely referring to Burt (1936).]

9. Realistically, of course, conditions themselves may be classified in two or more ways, for example, test questions being one basis for classification and scorer being another. The matrices that result when persons are combined with such complex systems of conditions are the subject of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam 1972), and did not enter into the 1951 article.

10. To avoid confusion, my colleagues and I adopted the convention of referring to the domain of items from which tests were presumably sampled as the *universe* of items, reserving the term *population* for the persons represented in a study.

11. The statements in the preceding two paragraphs are in no way peculiar to alpha. They appear in the theory for any other type of reliability coefficient, with the sole reservation that some coefficients rest on the assumption that every test in a family has the same correlation with the corresponding true score.

12. This assumption of independence enters the derivation of any internal-consistency formula.

13. [Cronbach is likely referring to Jackson and Ferguson (1941).]

14. Most of the analyses involved more complex structures, for instance, a three-way matrix in which persons, tasks, and scorers were treated as separate bases for sorting scores.

15. It may be said at the outset that these methods retained Fisher's calculations but then went beyond them to an interpretation that would have been meaningless with fixed factors such as species.

16. [$V_p = E(\mu_p - \mu)^2$; $V_i = E(\mu_i - \mu)^2$; $V_{Residual} = E(X_{pi} - \mu_p - \mu_i + \mu)^2$; $V_{X_{pi}} = V_p + V_i + V_{Res}$, where E is the expectation operator.]

17. [Alpha, expressed in variance-component terms, is

$$\alpha = \frac{V_p}{V_p + \frac{V_{Res}}{k'}}$$

where k' provides the Spearman-Brown adjustment for length of test (or, alternatively, number of tests).]

References

Brennan, R. L. (2001). *Generalizability theory.* New York: Springer-Verlag.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296-322.

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners* (pp. 245-314). London: Macmillan.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J. (Ed.). (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow.* Mahwah, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity among profiles. *Psychological Bulletin*, *50*(6), 456-473.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373-399.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizbility: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163.

Cronbach, L. J., Schonemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, *25*, 291-312.

Jackson, R. W. B., & Ferguson, G. A. (1941). *Studies on the reliability of tests* (Bulletin No. 12, Department of Educational Research, Ontario College of Education). Toronto, Canada: University of Toronto Press.

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271-295.