Reliability Indexes for Use in Educational Experiments: Cronbach's Alpha versus a *G* Study

Marion Kimball Slack
Darrell Sabers
Lon N. Larson
William F. McGhan
J. Lyle Bootman

INTRODUCTION

Research in education has developed along two different lines, experimental and correlational (1). The experimental method is concerned primarily with changing conditions and observing the effects on the student. For example, a researcher compares two lesson presentation formats, lecture and computer-assisted instruction. In experiments, the variance due to the treatment (in the example, presentation format) is the variance of interest; other sources of variance, such as individual variance, are considered error variance. In contrast, the correlational method is concerned primarily with individual variation, the desire to sort and rank individuals according to their individual differences, and then, if relevant, to predict the criterion performance of individuals. For example, a researcher ranks

Marion Kimball Slack, Ph.D., Lon N. Larson, Ph.D., and J. Lyle Bootman, Ph.D., are in the Department of Pharmacy Practice at the University of Arizona College of Pharmacy, Tucson, AZ 85721. Darrell Sabers, Ph.D., is in the Department of Educational Psychology, College of Education, University of Arizona. William F. McGhan, Pharm.D., Ph.D., is in the Department of Pharmacy Practice and Pharmacy Administration at the Philadelphia College of Pharmacy and Science, Philadelphia, PA 19104.

Journal of Pharmacy Teaching, Vol. 3(1) 1992

a group of individuals according to their intelligence. In correlational research, the variance due to individuals is of interest, and any variance due to treatment or other factors may be considered error variance.

Because the goals are different for each type of research, measurement methods and their accompanying reliability indexes are also different. Experimenters want measures that closely estimate the treatment effect. Precise measures will measure only the treatment effect and will be influenced little by individual differences or other factors; that is, there will be little variation around the mean. An appropriate reliability index will indicate the extent to which sources of variance other than the treatment contribute to variation around the mean. The sources of variance may be either systematic bias or random error. For example, individual differences may contribute to variation around the mean and may be either random variation in the population or systematic variance due to differences in intelligence.

Correlational researchers want measures that precisely sort and rank individuals; precise measures will measure only individual variability and will be influenced little by treatments or other environmental effects. An appropriate reliability index will indicate the extent to which sources of variance other than individual differences—for example, environmental effects—affect the relative rank of the individuals.

Although researchers in educational measurement carefully differentiate between the two types of measurement and their respective reliability indexes, researchers in pharmacy education tend to use reliability indexes derived from correlational research in experimental research settings (1-3). Researchers will use a correlational reliability index, such as Cronbach's alpha, to measure the reliability of an instrument used to differentiate the effects of two or more treatments. Differences in treatments are more difficult to detect when an instrument exhibits high correlational reliability (4). Therefore, when correlational reliability indexes have been used, researchers in pharmacy education may misinterpret the results from comparisons of educational interventions.

Because reliability indexes derived from correlational research may provide misleading information when used in a research set-

ting, the purpose of this paper is to compare the information from a correlational reliability index, Cronbach's alpha, to the information obtained from a *G* study. A *G* study is a type of analysis that quantifies the variance due to extraneous factors. By comparing the information from each type of study, we will demonstrate that correlational reliability indexes are inappropriate in experimental settings and that better information is available from a *G* study.

Since Cronbach's alpha and a *G* study represent different conceptual approaches to assessing reliability, we will first review classical test reliability, including Cronbach's alpha, and *G* theory, from which a *G* study is derived. Then we will compare them by using a Cronbach's alpha and a *G* study to assess the measurement reliability of a dependent variable in an experimental setting.

A REVIEW OF THE ASSESSMENT OF MEASUREMENT RELIABILITY

Classical test theory is a test theory of individual differences (5). The object of measurement is the individual (person), and the purpose of measurement is to differentiate between individuals or to rank an individual in relation to other persons in the group. Classical test theory rests on the assumption that the observed score is the sum of the true score and error (6). The true score is generally conceptualized as the mean of an infinite number of measurements on the same person. Therefore, variation in scores is assumed to be caused by variation between persons. Variation between persons is considered true score variation, while variation due to other factors is considered error variance.

Intelligence tests represent a typical test in classical test theory. The purpose of testing is to rank one student as more (or less) intelligent than a second student. Because people are ranked according to their individual differences in intelligence, a reliable test will rank people consistently high (or low) in the group if they are retested on a later occasion. The relative consistency of ranks between test occasions is measured by test-retest reliability.

Test-retest reliability, also known as the coefficient of stability, may be calculated in several ways. When only two test occasions are considered, a Pearson correlation coefficient obtained by corre-

lating the scores on the first test with the scores on the second test measures test-retest reliability (6). The Pearson correlation indicates the extent to which a person's rank within the group is consistent from the first test occasion to the second test occasion.

If more than two test occasions are included in the assessment, an ANOVA approach is needed. A randomized blocks ANOVA (repeated measures ANOVA) with persons as one factor and test as the second factor provides the mean squares needed to estimate reliability. Each score in the matrix represents a single subject's total test score for the specified test occasion. The test occasions factor can include as many test occasions as desired. Test-retest reliability is the ratio of between-persons variance to between-persons variance plus mean error variance (7).

The coefficient of equivalence represents another form of test reliability in classical test theory. The coefficient of equivalence indicates the consistency of subjects' ranks from one test to a parallel form or forms of the test. It is calculated as a Pearson correlation coefficient for two parallel tests. When more than two tests are involved, a randomized blocks ANOVA in which the factors are persons and test forms is used. Like test-retest reliability, the coefficient of equivalence is the ratio of between-person variance to the sum of between-person variance and mean error variance.

The coefficients of stability and equivalence describe the reliability of total test scores. Similar approaches can be used to obtain coefficients to describe the homogeneity of items within a test (7, 8). Homogeneity indexes include Cronbach's alpha and the intraclass correlation coefficient. Again, the randomized blocks ANOVA is used to calculate homogeneity coefficients, except that item scores replace total scores in the body of the matrix and the test occasions factor is replaced with a test items factor. The test items factor is analogous to the test forms factor of test-equivalence reliability. In assessing homogeneity, the subject is measured across items rather than across test forms (8). Cronbach's alpha is the ratio of between-person variance to the sum of between-person variance and the mean error variance, and the intraclass correlation coefficient is the ratio of between-person variance to the sum of between-person variance plus the total error variance (9).

Cronbach's alpha and the intraclass correlation coefficient (ICC)

were described above as ratios of between-person variance and error variance. Other formulations for alpha are available; the three principal formulations are shown with an ICC in Figure 1. The classical test and the Kuder-Richardson 20 formulations were derived from correlational testing (see Nunnally for derivations) (3). The ANOVA formulation was derived from the variance definition, which states that observed score variance is composed of true score variance and error variance (6-8). The correlational and ANOVA formulations are equivalent (8).

FIGURE 1

Classical Test Form of Alpha[a]

$$\text{alpha} = \frac{k}{k-1}\left(1 - \frac{\sigma_i^2}{\sigma_T^2}\right)$$

ANOVA Form of Alpha[b]

$$\text{alpha} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2/k}$$

Kuder-Richardson 20 Form of Alpha[c]

$$r_{kk} = \frac{k}{k-1}\left(1 - \frac{\Sigma_{pq}}{\sigma_T^2}\right)$$

Intraclass Correlation Coefficient (ICC)[d]

$$\text{ICC} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}$$

a See Cronbach et al. (8): k = number of items; σ_i^2 = sum of item variances and σ_t^2 = total variance.

b See Brennan (9): aσ_p^2 = person variance and σ_e^2 = error variance.

c See Nunnally (3) for derivation of (alpha) and K-R 20. K-R 20 is a special case of alpha when items are scored 1 (correct) or zero (incorrect); p = proportion correct and q = proportion incorrect.

d See Cronbach et al. (8): the Spearman-Brown formula converts the ICC to alpha; that is, it estimates the reliability of a test with k items.

The Kuder-Richardson 20 formulation is a special form of alpha used when items are scored correct or incorrect (one or zero). The ANOVA formulation is the most general formulation and can be used for items scored one and zero as well, although Winer argues that a matrix of one and zero scores should be interpreted as a profile of scores (7, 10). The same values will be obtained for alpha regardless of which formulation is used. [Note: Dichotomous scores, such as one and zero, violate the assumption of a normal distribution needed for hypothesis testing in ANOVA; however, a normal distribution is not required for obtaining variance estimates (5).]

The intraclass correlation coefficient (ICC) was derived outside test measurement by Fisher and other researchers using analysis of variance research designs but was found to be closely related to classical reliability measures (8). Conceptually, alpha and the ICC apply to slightly different situations. If one is assessing the reliability of several items, alpha will indicate the reliability of the score represented by the mean of the items. The ICC represents the average reliability of the score provided by a single item. For a discussion of the relationship between the different reliability indexes, see Cronbach and colleagues; see Shavelson for a discussion of reliability within an ANOVA framework (7, 8, 11).

The variance formula and the mean squares needed to estimate Cronbach's alpha from a randomized blocks ANOVA are shown in Table 1. See Shavelson for a description of the relationship between the mean squares and variance components (7). Note that the error variance is divided by k, the number of items; therefore, longer tests are likely to be more reliable than shorter tests because the error variance will be reduced. Only the information needed for Cronbach's alpha is shown in the table because alpha is considered representative of the reliability indexes from classical test theory.

Classical reliability indexes are a family of related indexes, with each index representing reliability under specific conditions, conditions of retesting, equivalence, etc. Further, in classical reliability, several assumptions are made a priori. The object of measurement is assumed to be the individual, and the error term is undifferentiated; that is, all sources of error are included in a single term, so the specific source of error cannot be identified. Thus, test site could be

TABLE 1

Component	Variance	Mean Square (MS) Estimate
Persons (p)	$(\sigma_p^2)^a$	(MSp - MSe) / k[b]
Items (i)	$(\sigma_i^2)^c$	(MSi - MSe) / n[d]
Error (e)	(σ_e^2)	MSe[e]

a Cronbach's alpha = $\sigma_p^2/(\sigma_p^2 + \sigma_e^2/k)$, therefore, only the person and error variance components are needed to determine alpha (Brennan, 1983, p. 18).
b k = Number of items.
c A G-study is concerned with the quantity of item variance as well as with person variance.
d n = Number of persons.
e alpha = (MSp-MSe)/MSp when the mean square estimates are substituted for the variance components and simplified (Cronbach et al., 1963, p 142). See Shavelson (7) for formulas to calculate the mean square estimates.

introducing large amounts of variation, but a classical reliability index would indicate only that a large amount of error was present, not the source of the error.

G theory is more general in its approach to reliability than classical test theory. In fact, Cronbach and colleagues developed *G* theory because classical reliability theory was restrictive in its assumptions and limited in its application (8). To expand reliability assessment in *G* theory, an observed score was conceptualized as a single score from a class of observations. A reliability index, then, represents the degree to which the observed measurement can be generalized to a class of similar observations. The class of similar observations may include observations concerning the relative rank of individuals, as in classical test theory, or observations of groups, items, raters, etc. The class of observations—that is, the object of measurement—is not specified a priori by *G* theory. Rather, the researcher must specify the object of measurement. Because one is concerned with generalizing to a class of similar observations, *G*

theory is known as generalizability theory and represents a more comprehensive way of thinking about reliability.

The variation of observed scores is defined as the sum of the variance due to the object of measurement and the variance due to error. Multiple sources of error variance can be considered simultaneously, including test site, test occasion, test form, and test items. Therefore, a multifactorial analysis of variance, such as a randomized blocks ANOVA, is needed to assess error variance. Unlike classical test theory, which has an undifferentiated error term, *G* theory requires specification of the sources of error variance that researchers consider relevant based on past research or theory.

The information concerning extraneous error variance can be used to modify the test procedure to decrease the variance due to error. To assess the variance due to different sources of error, variance estimates are obtained from a randomized blocks ANOVA for each source of error. Then the quantity of variance due to one source of error is compared directly with other sources of error. The test procedure is modified to eliminate or reduce large sources of error variance. If one source of error (for example, test site) is much larger than the other sources of error (such as test items or test form), the test procedure is modified to reduce the error due to test site.

A *G* study also provides the information for estimating several different reliability coefficients, including Cronbach's alpha and the ICC, if alpha or the ICC are appropriate methods for assessing reliability (9). A generalizability coefficient can also be calculated. Generalizability coefficients are like an ICC except that they are not restricted to considerations of between-person variance or to ranking individuals. The generalizability coefficient can be calculated for any object of measurement and for either relative ranks or absolute measurements. In fact, the relationship between the reliability coefficients and the relative advantages and disadvantages of each can be considered within *G* theory (8, 9). Thus, *G* theory includes classical test reliability as a special case.

Since *G* theory makes no a priori assumptions concerning the object of measurement or possible sources of error variance, the researcher must specify both when undertaking a *G* study. For experimental research, Kerlinger's MAXMINCON principle provides

a useful guide (12). According to this principle, the researcher wants maximum differences on levels of the independent variable but minimum differences or variance on extraneous variables. If the purpose of an educational experiment is to compare different presentation formats, variance between the groups is desirable, while variance due to the items of the test or to subjects is undesirable and should be minimized in the experiment. A *G* study conducted for this example would estimate the quantity of variance due to each factor to determine if the variance contributed by items or persons is indeed minimal.

METHODS

The reliability study reported here was conducted separately from the experimental study. The experiment compared the group means of pharmacists and pharmacy technicians on recognition, judgment, and decision tests. For the reliability study, the subjects were pooled into one group ($n = 20$), and no consideration was given to the different groups.

The items for the tests were brief descriptions of patients or drug names. Two additional pharmacists rated each item on a seven-point scale with anchors of typical and atypical. The correlation between the ratings assigned by the 2 pharmacists was 0.94 for the patient items and 0.95 for the drug items. The items were divided into a recognition test consisting of drug names plus a judgment test and a decision test, each consisting of a drug name accompanied by a patient description. Each test contained two sections, one section containing eight typical items and the other, eight atypical items. The items were scored zero or one so that the scores for each section could vary from zero to eight. On the recognition test, a choice was scored one if the drug name chosen was the correct drug name. On the judgment test, a choice indicating that the drug was appropriate was scored one, and a choice indicating that the drug was inappropriate was scored zero. On the decision test, a choice indicating that the drug should be dispensed to the patient was scored one, and a choice indicating that the physician should be consulted before dispensing was scored zero.

The test items were presented on the screen of a lap-top com-

puter, and subjects responded by pressing one of two keys on the keyboard. A computer random number program was used to determine the order of the items and the order of the responses on each item. Each subject was tested individually in an office at the clinic site.

The data for the reliability study were analyzed using a random model, randomized blocks design in which test items were the repeated factor and the person times items interaction was the error term, as described by Shavelson (7). The mean squares obtained from this analysis were used to estimate the variance components needed for Cronbach's alpha and the *G* study.

Cronbach's alpha was calculated as shown in Table 1. Note that variance due to items is not explicitly included in alpha. As described above, alpha is the ratio of between-persons variance to the sum of between-persons variance and mean error variance. Alpha is expected to be high when the quantity of between-persons variance is large.

For the *G* study, groups were designated as the object of measurement, and the sources of error variance examined were between-persons variance and item variance. Thus, both between-persons variance and item variance are considered error variance in the *G* study. The reliability of the group means is expected to be lower when the quantity of either or both is large.

The variance estimates needed for the *G* study are estimated in a manner identical to those used for Cronbach's alpha. However, for the *G* study, the variance due to items is of interest as well as the variance due to persons. Therefore, both variances were estimated according to the formulas in Table 1.

To demonstrate that relatively large between-persons variance and a high alpha coefficient are undesirable when comparing groups, a *t*-test was conducted for comparisons between the pharmacist and technician groups. The *t*-test and corresponding *p* value were calculated for tests having low, medium, and high quantities of between-persons variance. The difference between the groups was set at one unit for all between-groups comparisons so that any changes in the *p* value could be attributed to the quantity of between-persons variance present.

RESULTS AND DISCUSSION

The variance components for the reliability analyses appear in Table 2. The variance due to error appears quite constant for all of the tests. However, the variance for items and persons differs across tests and across item type. On each test, the between-persons variance is greater for atypical items than for typical items. Item variance is also greater for atypical than for typical items on the judgment and decision tests. The variance for both typical and atypical

TABLE 2

	Source		
Test/Item Type	Items[a]	Persons[b]	Error[c]
Recognition			
Typical	0.0052	0.0115	0.117
Atypical	0.0049	0.0245	0.144
Judgment			
Typical	0.0177	0.0029	0.167
Atypical	0.0192	0.0280	0.203
Decision			
Typical	0.0284	0.0250	0.157
Atypical	0.0325	0.0701	0.156

a Item variance = (MSi - MSe)/n where MSi = square items, MSe = mean square error, and n = total number of persons.
b Person variance = (MSp-MSe)/k where MSp = mean square persons, MSe = mean square error, and k = number of items.
c Error variance = mean square error (MSe).

is at least 1.6 times as great on the decision test as on the corresponding items for the other tests.

Within a *G* study, the variance components can be examined directly for their contribution to undesirable variance in the measurement of the group means. Undesirable variance in an experiment was identified above as variance attributable to persons or items. Some variance is unavoidable, but the person variances on both the typical and atypical items on the decision test are substantially higher than on the recognition or judgment tests. There was also increased item variance on the decision test when compared to the recognition test or the judgment test. The increased variance could indicate a problem with the decision test. In fact, several subjects commented that they were not sure how to respond on the decision test, indicating that, indeed, there was a problem with that test.

The increased variance for the atypical items probably does not indicate any problem with these items. Responses to atypical items are characterized by greater variance both between persons and within persons (13). Therefore, increased variance for atypical items in general is to be expected. Given the increased between-persons variance on atypical items, one can expect that statistically significant differences between groups are more difficult to find when the items are atypical.

The values for alpha appear in Table 3 for the three different tests and for typical and atypical items. According to the alpha values, the atypical items on the decision test produced the most reliable scores. However, this conclusion is counter to the conclusion reached with the *G* study. Direct examination of the variance components in the *G* study indicated undesirable variance in the scores on the decision test. The apparent contradiction in the results can be resolved by considering the purpose of and underlying assumptions made by Cronbach's alpha.

As described above, alpha was developed in the context of testing for individual differences; therefore, the purpose of alpha is to indicate the extent to which individuals are likely to be ranked consistently when tested over multiple items. This purpose is considerably different from identifying the factors that contribute variance in the measurement of group means. Further, alpha is based on the

TABLE 3

Test	Item Type	
	Typical	Atypical
Recognition	0.441[a]	0.576
Judgment	0.122	0.524
Decision	0.561	0.782

a Alpha = $\sigma_p^2/(\sigma_p^2 + \sigma_e^2/k)$

assumption that between-persons variance is desirable, while for the measurement of group means, between-persons variance is undesirable. As indicated by Nunnally, alpha is appropriate only when large individual differences are desirable (3).

The contradictory results obtained by alpha and the *G* study are demonstrated by the comparison of *p* values obtained from conducting an independent groups *t*-test using scores from tests with low, medium, and high values of alpha. As shown in Table 4, as the value of alpha increases (or as the quantity of between-persons variance increases), the *p* value increases. Therefore, a difference between groups is less likely to be demonstrated when alpha is high than when alpha, and the corresponding between-persons variance, is low.

In addition to obtaining contradictory results, using alpha to assess reliability in an experiment may mislead the researcher. A high value for alpha may be interpreted by the researcher as indicating that the measurement of the effects due to the independent variable is reliable; that is, as indicating that there is little extraneous variation present. Instead, a large amount of extraneous variation is likely to be present due to the large amount of between-persons variability needed to produce a high value for alpha.

That Cronbach's alpha and *G* theory lead to different conclusions

TABLE 4

Between-Person Variance	Alpha	t[b]	p
0.0029[c]	0.122	3.62	0.002
0.0250[d]	0.561	2.65	0.017
0.0701[e]	0.782	1.87	0.081

a To facilitate the between-groups comparisons, the difference between groups was equal for all three comparisons.
b "t" is for an independent group comparison on total mean scores for each test; "p" is the significance of the t value.
c For typical items on the judgment test; SD = 1.234, N = 20.
d For typical items on the decision test; SD = 1.689, N = 20.
e For atypical items on the decision test; SD = 2.395, N = 20.

about the reliability of the test used to measure the effects of the treatment indicates that any reliability index should be selected based on the purpose of the measurement and the characteristics of the particular index. Otherwise, the researcher may select a reliability index that is inappropriate for the setting in which the measurement is made.

While a *G* study provides information relevant to assessing the measurement reliability in an experiment, there are several disadvantages to conducting a *G* study. Except for very small data sets, a *G* study requires a computer program for statistics that includes a random effects, randomized blocks model (e.g., BMDP or a program developed specifically for *G* theory). A *G* study can also become very complicated and difficult to interpret if a number of sources of extraneous variance are considered. However, the real impediment to conducting a *G* study is the necessity for collecting data to estimate the variance due to situational factors.

LIMITATIONS

The results of the study comparing the performance of pharmacists and pharmacy technicians on the recognition, judgment, and decision tests may have limited generality due to the small number of subjects and the computer presentation format. However, the reliability study was concerned with the selection of an appropriate reliability index and with demonstrating the conflicting results that are possible if an index is selected inappropriately. Thus, the generalizability of the reliability study is concerned with the degree to which the issues raised in this example are applicable to other situations in which the researcher must select a reliability index. Because the selection of an appropriate reliability index depends on purpose and not on sample size or presentation format, the issues illustrated by this example should be applicable to other research situations.

CONCLUSIONS

As demonstrated by the reliability study, a *G* study provides specific information about the effects of extraneous variation on the measurement of the dependent variable. The *G* study identified excessive between-persons variation on the decision test, which could be due to subjects' inability to understand how they were to perform. The study also identified differences in between-persons variance on the typical and atypical items. Such specific information was not available from the reliability assessment using Cronbach's alpha. In fact, alpha was shown to be an inappropriate reliability measure for an experiment in which the means of independent groups are compared.

REFERENCES

1. Cronbach LJ. The two disciplines of scientific psychology. Am Psychol 1957;12:671-84.

2. Thorndike RL. Reliability. In: Linquist EF, ed. Educational measurement. Washington, DC: American Council on Education, 1951:561-73.

3. Nunnally JD. Psychometric theory. New York: McGraw-Hill, 1978.

4. Nicewander WA, Price JM. Reliability of measurement and the power of statistical tests: some new results. Psychol Bull 1983;94:524-33.

5. Kane MT, Brennan RL. The generalizability of class means. Rev Educ Res 1977;47:267-92.

6. Brown FG. Principles of educational and psychological testing. New York: Holt, Rinehart and Winston, 1976.

7. Shavelson RJ. Statistical reasoning for the behavioral sciences. Needham Heights, MA: Allyn and Bacon, 1988.

8. Cronbach LJ, Gleser GC, Rajaratnam N. Theory of generalizability: a liberalization of reliability theory. Br J Math Stat Psychol 1963;16:137-73.

9. Brennan RL. Elements of generalizability theory. Iowa City, IA: ACT Publications, 1983.

10. Winer BJ. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

11. Cronbach LJ, Gleser GC, Nanda H et al. The dependability of behavioral measurements. New York: John Wiley, 1972.

12. Kerlinger FN. Foundations of behavioral research. New York: Holt, Rinehart and Winston, 1973.

13. McCloskey MC, Glucksberg S. Natural categories: well defined or fuzzy sets? Memory Cognition 1978;6:462-72.