DOCUMENT RESUME

ED 395 950 TM 025 017

AUTHOR Arnold, Margery E.
TITLE Influences on and Limitations of Classical Test Theory Reliability Estimates.
PUB DATE 25 Jan 96
NOTE 30p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, January 25, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Generalizability Theory; Heuristics; *Interrater Reliability; *Research Methodology; Test Interpretation; *Test Reliability; Test Theory

ABSTRACT

It is incorrect to say "the test is reliable" because reliability is a function not only of the test itself, but of many factors. The present paper explains how different factors affect classical reliability estimates such as test-retest, interrater, internal consistency, and equivalent forms coefficients. Furthermore, the limits of classical test theory are demonstrated, and it is recommended that researchers, teachers, and psychologists instead utilize generalizability-theory estimates of reliability. Heuristic examples and detailed explanations make this discussion accessible even to those who are uninitiated in either classical test theory or generalizability theory. (Contains 2 figures, 10 tables, and 18 references.) (Author)

Influences on and Limitations of Classical Test Theory Reliability Estimates

by

Margery E. Arnold
Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, LA, January 25, 1996.

Abstract

It is incorrect to say "the test is reliable" because reliability is a function not only of the test itself, but of many factors. The present paper explains how different factors affect classical reliability estimates such as test-retest, inter-rater, internal consistency and equivalent forms coefficients. Furthermore, The limits of classical test theory are demonstrated, and it is recommended that researchers, teachers and psychologists instead utilize generalizability-theory estimates of reliability. Heuristic examples and detailed explanations make this discussion accessible even to those who are uninitiated in either classical test theory or generalizability theory.

The reliability of test scores concern teachers, psychologists and researchers who want to know that the scores on the tests which they administer are consistent and generalizable. Unfortunately, many training programs in the disciplines of education and psychology still emphasize classical methods of deriving reliability coefficients (such as test-retest reliability, internal consistency reliability). In addition, typical ways of teaching about selecting and using standardized tests may unwittingly teach students that instruments or tests can possess a quality called "reliability."

The present paper demonstrates the extent to which classical reliability estimates derived using the true score model can yield accurate estimates of measurement error, as well as explains the limits of these classical reliability estimates. Furthermore, the paper illustrates how reliability is a quality of *scores* on tests, and *not* a quality of tests or instruments. Many factors affect the magnitude of reliability estimates. These contributing factors include qualities of test items to be sure, but they also include characteristics of examinees--which certainly are not the same from testing to testing. For these reasons researchers, psychologists and teachers are advised to neither write nor say that "the test is reliable (or not reliable)." Rather, what should be said or written is, "the scores from this testing are reliable (or not reliable)."

Definitions of Reliability According to Classical Test Theory

To explain how tests cannot be reliable, only scores can be reliable, it is helpful to give a brief review of the concept of reliability. Reliability expresses the relationship between observed scores and true scores. A concrete example using the spelling test scores of a second grade class clarifies these constructs. "True score" refers to each classmember's true ability in the domain of 2nd grade spelling. "Observed score" refers to each classmember's actual score on the spelling test. So,

reliability concerns the relationship between what the children actually know about spelling (true score) and what they made on their spelling test (observed score). This relationship between true score and observed score can be conceptually explained as a mathematical model, a statistic, and an illustration, all of which are demonstrated below.

The relationship between true score and observed score as expressed in a mathematical model was explained by Charles Spearman (1907, 1913, cited in Crocker & Algina, 1986) in what has become known as the true score model or classical test theory. According to Crocker and Algina (1986), "the essence of Spearman's model was that any observed test score could be envisioned as the composite of two hypothetical components--a true score and a random error component" (p. 107). Thus, the equation is Observed score=True score + Random error. In other words, observed scores are composed of true scores (which, by definition are reliable) and an error component that is not reliable. From the true score model one can construct another equation for reliability as expressed in a ratio of true score variance to observed score variance. So, reliability is "the proportion of observed score variance that may be attributed to variation in the examinees' true scores" (Crocker & Algina, 1986, p. 116).

It is important to note that the error term in this model can have a positive effect or a negative effect on observed scores. For example, if a student does not know the answer to a question, but guesses correctly, his or her observed score will be higher than his or her true score. In this case the measurement error that was introduced by the guess has a *positive* effect on the observed score. Alternatively, another student might know the correct answer and mismark the answer, so that she or he answers incorrectly. In this case, the true score exceeds the observed score and the error that was introduced by mismarking had a *negative* effect on the student's observed score.

As noted above, reliability can also be expressed as a statistic such as coefficient alpha or as the pearson product-moment correlation, r. Alpha will be explained later, so this demonstration focuses on r. Expressed as a correlation, reliability is the correlation between the true scores (actual ability in the domain of 2nd grade spelling) and observed test scores (the class's scores on the spelling test). [The correlation between true scores and observed scores is called the *reliability index*. The *reliability coefficient* is the reliability index squared (Crocker & Algina, 1986, p. 115-116).] This relationship between true scores and observed scores can also be illustrated as the graph found in Figure 1, or the diagram found in Figure 2.

The Impact of Examinees on Test Score Reliability

Now that the concept of reliability has been reviewed and illustrated as the correlation statistic, r, one can easily demonstrate why it is incorrect to say that "the *test* is reliable" and more accurate to say that "the *scores* are reliable." Hinkle, Wiersma and Jurs (1994) describe how r is affected by homogeneity of the group. When a group is very homogeneous all of the members tend to score similarly to one another. When this happens the range of scores is very small and thus the standard deviation of the scores is very small. After reflecting on the formula for the correlation coefficient, r=Covariance of X & Y /[(standard deviation of X)(Standard deviation of Y)], one notices that,

> If a group is sufficiently homogeneous on either or both variables, the variance (and hence the standard deviation) tends toward zero. . . . When this happens, we are dividing by zero, and the formula becomes meaningless. In essence, the variable has been reduced to a constant. As a group under study becomes increasingly homogeneous, the correlation coefficient decreases. (Hinkle, Wiersma & Jurs, 1994, pp. 115-116)

Since reliability is a correlation coefficient, it is affected by the homogeneity of the group to whom the test is given. As the group being studied becomes increasingly

homogeneous, the reliability coefficient decreases. Thus reliability is affected not only by the properties of the items on the test, but also by the persons taking the test.

An illustration using the spelling test example makes this more clear. Figure 1 displays the relationship between the entire 2nd grades *true* spelling scores and their *observed* spelling test scores. The reliability of the spelling scores is calculated in Table 1 to be .9977, which is considered to be rather reliable. A second reliability coefficient has also been calculated. This second coefficient is the reliability coefficient for the scores of the top five students in the 2nd grade. The top five students are likely to do homogeneously well on the spelling test. The range of their scores is smaller than the range of scores of the entire second grade. Notice that the reliability coefficient for this group is lower (reliability = .7456) than the coefficient for the entire grade. This example illustrates one reason why it is incorrect to say "the test is reliable" or to say "the test is not reliable". As Gronlund and Linn (1990) noted,

> Reliability refers to the *results* obtained with an evaluation instrument and not to the instrument itself.... Thus it is more appropriate to speak of the reliability of the "test scores" or of the " measurement " than of the "test " or the " instrument". (p. 78, emphasis in original)

A test, in and of itself, cannot be reliable because reliability is a function not only of the items on a test, but also a function of *who* takes the test. As Rowley (1976) states, "It needs to be established that an instrument itself is neither reliable nor unreliable. . . .A single instrument can produce scores which are reliable and other scores which are unreliable". (p. 53)

Implications for Psychometrists, Therapists and Researchers

Education and psychology training programs teach students to read test manuals to examine reliability and validity coefficients. When evaluating norm statistics (such as reliability) reported in test manuals, psychometrists must be very

careful that the intended examinee is similar to the normed group. As the example in Figure 1 illustrates, however, even when one cautiously selects a test that has yielded reliable scores for similar examinees in the past, it is incorrect to assume the test will yield reliable scores for all future uses of the test. A psychometrist is also cautioned to look at the homogeneity or heterogeneity of the norm group. For example, as the spelling test example has shown, if the norm group is incredibly heterogeneous compared to the group for whom the test is designed, one might expect that the reliabilities calculated for the intended groups will be lower than the ones reported in the manual. For this reason, and other reasons, it is important to calculate reliability *for every group to whom the test is given.*

Because reliability is a function of at least both the test and the test-takers, researchers as well as psychometrists should calculate reliability statistics on the scores of every group of persons that they measure. Researchers calculate reliability statistics on their own data for two reasons. The first reason to calculate reliability statistics on one's own data has been discussed above: to discover the extent to which measurement error has affected one's data. The second reason for calculating reliability statistics on one's own data is to determine the extent to which measurement error is limiting the effect sizes in the study of interest. Reinhardt (1991) cautioned researchers on this point, noting that "Prospectively, researchers must select measures that will allow detection of effects at the level desired; retrospectively, researchers must take reliability into account when interpreting findings" (p. 1).

An example using the effect size for a correlation coefficient explains this principle clearly. As Thompson (1991) explained, the correlation coefficient is the basis for all parametric stat.stics; "*all* classical analytic methods are correlational" (emphasis in original, p. 87). Therefore, the principle involving reliability coefficients and effect sizes has implications for *all* effect sizes in *all* common

statistical procedures such as ANOVA, multiple regression, MANOVA, factor analysis, discriminant function analysis, and canonical correlation analysis. As Locke, Spirduso and Silverman (1987) noted, "the correlation between scores from two tests cannot exceed the square root of the product for reliability in each test" (p. 28). Written in equation form, the relationship between reliability and correlation looks like this:

$$r_{xy} \leq [(\text{reliability of X})(\text{reliability of Y})]^{.5}$$

This formula for the correlation between scores on X and scores on Y can be algebraically changed by squaring both sides of the equation to create an r^2 type of effect size. The new equation explains how reliability is related to effect sizes.

$$r^2_{xy} \leq [(\text{reliability of X})(\text{reliability of Y})]$$

The effect size can be no greater than the square root of the product of the reliability coefficients for the two measures that are being correlated.

An example illustrates how reliability influences effect size. One researcher is interested in the effects of self-esteem on achievement test scores. She uses the Self-Esteem Scale of the Behavioral Assessment System for Children (BASC) (Reynolds & Kamphaus, 1992) to measure the self-esteem of third graders at a local elementary school. To measure achievement, she uses the achievement scores from the standardizes testing of the school district. She obtains two reliability coefficients, one for the self-esteem scores (r_{xx}= .60) and one for the achievement scores (r_{yy}= .90). Using the formula above, the researcher learns that the maximum effect size that she can obtain when she correlates self-esteem scores with achievement test scores will be .54 (effect size $\leq$ [(.6)(.9)]=.54). The researcher in this hypothetical example obtained an effect size of .52. Her uninformed colleague told her that the effect size was only "moderate". She replied to the colleague, "Moderate? How can you say that it is 'moderate' when the maximum effect size I could have found was .54? This is not a 'moderate' effect size. In the context of

what could be (maximum =.54), .52 is a rather strong effect size." Thompson (1994) warned the would-be researcher to weigh the effects of reliability on effect sizes when planning and evaluating research.

> The failure to consider score reliability in substantive research may exact a toll on the interpretations within research studies. For example, we may conduct studies that could not possibly yield noteworthy effect sizes given that score reliability inherently attenuates effect sizes. Or we may not accurately interpret the effect sizes in our studies if we do not consider the reliability of the scores we are actually analyzing. (p. 840)

Methods of Estimating Measurement Error Using the True Score Model

Returning to the Classical True score model (Observed Score= True Score + Random Measurement Error), what is of interest to the researcher, the psychologist or the teacher is the examinee's true score. Crocker and Algina (1986) defined true score "as the average of the observed scores obtained over an infinite number of repeated testings with the same test" (p. 109). Unfortunately it is impossible (and impractical) to calculate true scores in this manner. True scores cannot be exactly calculated. They can only be estimated. True scores are estimated using what the researcher *can* obtain--observed scores, measurement error estimates (i.e.., reliability coefficients) and the true score model.

The true score is predicted by estimating the amount of measurement error that occurred in the administration of a test and then adjusting the observed scores using that estimation of error. If one knows the measurement error, then one can estimate the extent to which the measurement error has caused the observed scores to deviate from the true scores. It follows, then, that the estimation of measurement error is the key to finding true scores. Crocker and Algina (1986) defined error in the classical true score model as "an error of measurement. . .the discrepancy between an examinee's observed test score and his or her true score" (p. 110). [Note that this

type of error is *measurement error* as opposed to *sampling error* or *model error*. Sampling error is the difference between the statistic one obtained by measuring a sample and the statistic that one would have obtained had one sampled the entire population. Model error is the variance in the observed dependent variable score that is not explained by the independent variables in the model. Using the earlier example, model error would be the variance in the achievement test scores that cannot be explained by the self-esteem scores.]

In classical test theory, there are four sources of measurement error that are often estimated: (a) inconsistencies in occasions, (b) inconsistencies in forms, (c) inconsistencies between raters and (d) inconsistencies in sampling the content domain. What follows is a discussion for each source of error explaining how each of these inconsistencies is a source of measurement error and how to compute the reliability coefficient that corresponds to that source of error. Before moving directly to those explanations, however, it is important to note Gronlund's warning regarding reliability and particular sources of error.

> An estimate of reliability always refers to a particular type of consistency [e.g., consistency across occasions **or** across forms **or** across raters **or** across items sampled]. . . .It is possible for test scores to be consistent in one of these respects and not in another. The appropriate type of consistency in a particular case is dictated by the use to be made of the results. . . . The reliability coefficient resulting from each method [of calculating reliability] must be interpreted in terms of the type of consistency being investigated. (pp. 106-108)

With this warning having been heeded, an explanation of different sources of error and how to calculate reliability coefficients for each of those sources is considered next. A researcher may be concerned with how stable his or her observed test scores will be over time. In other words, if the test were administered to the

same group of people on a future occasion, how different will the test scores obtained on the second occasion be from the scores observed on the first occasion? The difference between the scores on these two occasions is a source of measurement error--measurement error due to occasions. Once this source of error has been measured (by testing the same group of persons with the same test on two different occasions) one can compute a reliability coefficient called the "stability coefficient". The stability coefficient is calculated by computing the pearson product moment correlation between the scores on the two different occasions (Crocker & Algina, 1986).

A second source of measurement error comes from inconsistencies in test forms. A teacher who would like to deter cheating on an exam may give two different forms of the same test. To understand how giving two different forms of the test might have introduced measurement error into the observed scores, the teacher may compute a reliability coefficient called the "equivalence coefficient". The equivalence coefficient is computed by calculating the pearson-product moment correlation between the scores on the two forms. Note that to compute this coefficient it is necessary to give both forms to at least a portion of the persons taking the test. One form is given and then the second form is administered within a short period of time. Usually the order of the forms is counterbalanced, so that order of test form will not affect scores or reliability estimates.

A third source of measurement error arises from inconsistencies between raters. (This type of error, of course, only occurs when one uses raters and will not occur during the administration of an objective exam with accurate machine scoring.) For example, one rater may have a slightly different method of assigning scores than another rater. To estimate the extent to which this type of measurement error has differentiated the observed scores from the true scores, one may calculate the inter-rater per cent agreement, or some similar coefficient.

A fourth major source of measurement error in classical test theory arises from inconsistencies in sampling the content domain or from inconsistencies in items. When a teacher (or psychologist) gives a test, he or she cannot possibly ask all of the questions in the domain of the content area being tested. Therefore, he or she must select possible items from the larger content domain. The teacher or psychologist hopes that he or she selected the correct items such that scores on the test he or she created can generalize to the domain of questions that might have been asked (Crocker & Algina, 1986). One may calculate the "internal consistency reliability coefficient" to estimate the extent to which this type of measurement error has caused the observed scores to deviate from the true scores. There are several ways to compute the internal consistency coefficient; all are based on the correlation between separately scored parts of the test (Crocker & Algina, 1986). "If examinees' performance is consistent across subsets of items within a test, the examiner can have some confidence that this performance would generalize to other possible items in the content domain" (Crocker & Algina, 1986, p. 135). Three common ways to calculate internal consistency reliability are the split-half coefficient, KR-20 and Cronbach's (1958) alpha. Split-half coefficients are the Pearson product-moment correlations between scores on two halves of the same test. KR-20 and Cronbach's alpha are computed with similar formulas, which are outlined below. Notice that the formulas are identical with one exception--how they compute the sum of the item variances. Because KR-20 is used only with dichotomously scored data, a simpler formula for item variance can be used.

$$\text{KR-20} = [k/(k-1)][1-(\varepsilon(pq)/\sigma_X^2)]$$

$$\text{Cronbach's Alpha} = [k/(k-1)][1-(\varepsilon\sigma_i^2/\sigma_X^2)]$$

k = number of items

p = percent of persons answering the item correctly

q = percent of persons answering the item incorrectly

$\varepsilon\sigma_i^2$ = sum of the item variances

σ_X^2 = test score variance

Problems with True Score Model Estimates of Measurement Error

The reliability coefficient has been previously defined. One definition that lends itself particularly well to graphic illustration is the one offered by Crocker and Algina (1986). The reliability coefficient is "the proportion of observed score variance that may be attributed to variation in the examinees' true scores". (p. 116) An example has been drawn in Figure 3. The outer rectangle represents the observed score variance. The shaded area of the observed score variance is the true score variance (or 90% of the observed score variance). The unshaded portion of the observed score variance is the measurement error variance (10% of the observed score variance). Notice how the reliability coefficient (r_{xx} =.9) defines how much of the observed score variance is measurement error variance. Recall that in classical test theory, reliability is defined by the source of error being estimated (occasions, raters, forms or items).

From this diagram and the calculations that are used to estimate reliability in classical test theory, it becomes apparent that classical test theory only allows for the estimation of one type of error at a time--e.g., only inconsistencies across forms, but not across raters, items or occasions (Webb, Rowley & Shavelson, 1988). It does not allow for estimations of simultaneously occurring measurement error. This major flaw in the true score model causes problems in the everyday use of reliability coefficients that have been derived using classical methods.

For example, a school psychologist is testing a boy to see if the boy will be given a mental retardation diagnosis. If the boy is given the diagnosis, then he will carry that diagnosis for many years to come. Therefore, it is very important that the

test that determines whether or not the boy receives a diagnosis yields scores that tend to have very high stability coefficients. That is to say, if the test diagnoses him as mentally retarded today, the test should diagnose him as mentally retarded at many points in the future, because he will be carrying this label for many years. Would it not also be important to simultaneously evaluate whether or not the items on the test enable the test administrator to generalize results on this test to the greater domain of mental retardation (i.e., that the test yields scores that tend to have high internal consistency coefficients)? Certainly, it would be important to ensure stability and internal consistency for a test with such far reaching implications. Classical test theory does not permit the researcher, psychometrist or teacher to simultaneously evaluate the effects of both of these possible sources of error on examinees' observed scores. Therefore, in Figure 3 the portion of the diagram that represents error variance can represent only one source of error at a time and does not meet the needs of most researchers, teachers or psychologists (Webb, Rowley & Shavelson, 1988).

There is a second problem with this limited model of measurement error (i.e., the True score model). The True score model does not account for error variance that may be caused by *interactions* between the different components of measurement error. Consider for example a testing scenario in which two judges assign ratings to candidates for entry into a graduate program based on 10 criteria. Table 2 outlines the 10 criteria and the ratings of the two judges on one of the applicants. Notice that the judges both gave the candidate a total score of 5. According to classical test theory, the candidate's score is consistent across raters; thus, the inter-rater reliability coefficient seems to be high. Also notice that the candidate received a "1" on all of the criteria. Therefore, according to the true score model, the candidate's scores seem to be consistent across items and thus, internal consistency reliability is high.

However, the true score model does not detect the obvious item-by-rater interaction effect. Such interactions can occur in common measurement situations, and can create sizeable and additional unique measurement error components. Thuse, it is a serious flaw that true score theory does not calculate or evaluate interaction sources of measurement error. This problem would be detected had the researcher used generalizability theory to derive his or her reliability estimates. Generalizability theory subsumes classical test theory and the True score model (Thompson, 1994). Generalizability theory is a topic too large for discussion in this paper. The interested reader is directed to Shavelson and Webb (1991) for more information. A note from Jaeger (1991), however, gives a flavor of the thoughts on generalizability theory in comparison to classical test theory: "Thousands of social science researchers will no longer be forced to rely on outmoded [classical theory] reliability estimation procedures when investigating the consistency of their measurements" (Jaeger, 1991, p. x).

Other Factors Affecting the Magnitude of Reliability Coefficients

Several factors influence the magnitude of reliability coefficients. Among those elements are homogeneity of the examinees, time limits placed on the test, the spread or variability of the scores, the length of the test and the difficulty of the items. The earlier spelling test example illustrated how homogeneity of examinees can attenuate reliability coefficients. A spreadsheet program created from the formula for KR-20 demonstrates how time limits, test score variability, test length and item difficulty affect reliability coefficients.

To understand how time limits affect coefficient alpha, it is important to recall that the True score model assumes that measurement error is random, not systematic (Observed Score= True Score [systematic] + Error [random]). The speed at which an examinee can complete a test, however, is systematic, not random. Therefore, speed is an ability that would fall under the systematic part of the True

score model. Tables 3 and 4 demonstrate the effects of speed on a seven-item, reading comprehension test taken by 10 persons. (Note. The grid represents persons' scores (0=incorrect and 1=correct) in the seven items. The final column lists each individual's total score. Across the bottom of the grid one can find "p" or the difficulty for each item and "v" the variance of each item. Below the grid one can find the elements of the KR-20 formula for alpha [k/k-1][1- (the sum of the item variances/ total test score variance)].) In Table 3, the examinees were given as much time as they wanted to complete the test. In Table 4, the test was timed and four members of the class--Todd, Nancy, Lu and Tammi--did not do as well as they did when they had all the time that they needed. (The items that they missed are in bold.)

One could reasonably argue that the first, untimed test scores illustrate the abilities of the class on reading comprehension better than do the timed test scores. Nonetheless, the reliability of the timed test (α= .74) is greater than the reliability of the untimed test (α= .20). This occurs because the timed test measures two abilities--reading comprehension *and* speed--as opposed to the untimed test which only measures reading comprehension. This measuring of two abilities provides for a greater spread in the total test scores. (Notice that the range on the timed test is 7-1=6, while the range on the untimed test is 7-4=3.)

The spread, or variance, in total test scores is the element of the KR-20 equation that *most greatly affects* the magnitude of coefficient alpha (Reinhardt, 1991). The spreadsheet program used in the timed-test example can also be used to illustrate how total test score variance affects coefficient alpha. Table 5 illustrates that when there is very little test score variance, coefficient alpha is at an absolute low (α= -21). **Note that coefficient alpha can be negative.** (This occurs when the sum of the item variances is greater than the total test score variance.) Also note that if there is no variability in total test scores, then it is impossible to compute coefficient

alpha. (It is impossible to divide by zero.) Therefore, in Table 5, the test score variance has been reduced to the lowest possible level without becoming zero.

Table 5 was transformed (changed values are in bold in Table 6) to create maximum test score variance. The item responses were changed so that half of the students answered all of the items correctly, while the other half of the students answered all of the items incorrectly. Arranging the test scores this way creates maximum deviation from the mean test score. (Recall that the formula for variance has as its numerator the sum of the squared deviations from the mean.) While these test scores are probably not test scores desired by any teacher, they are the test scores that will produce maximum total test score variance, and thus, maximum coefficient alpha.

The above describes mathematically how test score variance can increase reliability. Gronlund (1976) offers a conceptual explanation.

> Since the larger reliability coefficients result when individuals tend to stay in the same relative position in a group, from one testing to another, it naturally follows that anything which reduces the possibility of shifting positions in the group also contributes to larger reliability coefficients. In this case greater differences between the scores of individuals reduce the possibility of shifting positions. (p. 118)

A related concept concerns the effects of length of test on reliability coefficients. Longer tests, generally speaking, are likely to create more test score variance and thus increase reliability coefficients. It is possible, however, for one test to be longer than a second test and still yield scores with the exact same or lower reliability estimates than the shorter test. "There is one important reservation in evaluating the influence of test length on the reliability of the scores, . . . [this rule] . . . assume[s] that the test will be lengthened by adding test items of the same quality as those already in the test" (Gronlund, 1976, p. 118).

In other words, if the items added to the test are worse than the items already on the test, the longer test may actually yield lower reliability coefficients than the shorter test. Tables 7 through 10 demonstrate how adding items to a test may make reliability better, worse or the same depending upon the quality of the added items. Table 7 lists the scores of 10 persons on a seven item test (α= .84). Two items were added to this test and scores on the new items can be seen in Table 8. Notice that the two items that are added do not change the rankings of the examinees. Everyone answered the two added items incorrectly. Notice also that the coefficient alpha exactly equals the alpha of the test without the added items (α= .84).

A third example is given in Table 9. Notice that the added items increased the spreadoutness of the test scores; the range is now 9 instead of 7. This increase in total test score variance increased coefficient alpha (α= .89). However, in Table 10 one can see how adding items of lesser quality than the original items can actually decrease coefficient alpha (α= .69). In this example, the added items were of lesser quality than the original items because persons who had scored low (Skip and Jan) on the original test answered the items correctly while persons who scored high (Alex and Gina) on the original test answered the added items incorrectly. Furthermore, the added items decreased the variance of the total test scores. (Notice that the range on the original test was 7, while the range on the new test is 6.)

The item difficulty affects reliability in much the same way as test length does--by increasing or decreasing total test score variance. If all of the items on a test are rather difficult for all of the examinees, then the test score variance will be small and the reliability coefficient will be low. (The range of scores will be restricted, with everyone scoring near 0% correct.) The same phenomenon occurs if all of the examinees answer almost all of the items correctly. Reliability will be low because test score variance is low. (The range of scores will be restricted, with everyone scoring near 100% correct.) If, however, the test is of a medium difficulty for the

examinees, the scores will have a greater range, and reliability will be increased. Gronlund (1976) explains that to maximize reliability one should design a test so that

> the average score is 50 per cent correct and that the scores range from near zero to near perfect. . . . We can estimate the ideal average difficulty for a selection-type test by taking the point midway between the expected chance score and the maximum possible score. Thus for a 100 item true-false test the ideal average difficulty would be 75 (midway between 50 and 100), and for a 100 item five-choice multiple choice test the ideal average difficulty would be 60 (midway between 20 and 100). (p. 121)

Conclusion

The present paper has demonstrated that several factors influence reliability coefficients as derived using classical test theory. While the qualities of a test do contribute to the magnitude of the reliabilities of the scores that the test yields, the qualities of that test certainly do not control whether or not all scores on test can be called "reliable." Other factors including homogeneity of examinees, ability level of examinees vis-a-vis the test items, score variance, and test time-limits all have the potential to greatly influence reliability of scores. These potentialities have been demonstrated in the present paper. Furthermore, the limits of classical test theory reliability estimates have been detailed.

For reasons outlined in the present paper, the author makes two recommendations. First, teachers, psychologists and researchers should teach and/or understand the limitations of classical test theory reliability estimates. Second, researchers, psychologists and teachers should never write or say that "the test is reliable (or not reliable)." Rather, the author recommends that such persons

be accurate, writing and sayi. g, "the scores from this testing are reliable (or not reliable)." As Thompson has noted,

> This is not just an issue of sloppy speaking [or writing]--the problem is that sometimes we unc nsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. (Thompson, 1992, p. 436)

References

Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Fort Worth: Harcourt Brace Jovanovich College Publishers.

Gronlund, N. E. (1976). Measurement and evaluation in teaching (3rd ed.). New York: Macmillan.

Gronlund, N. E. & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Hinkle, D. E., Wiersma, W. & Jurs, S. G. (1994). Applied statistics for the behavioral sciences. Boston: Houghton Mifflin.

Jaeger, R. (1991). Foreword. In R. J. Shavelson & N. M. Webb, Generalizability theory: A primer (pp. ix-x). Newbury Park, CA: SAGE Publications.

Joint Committee on Standards for Educational Evaluation. (1994). The program evaluation standards: How to assess evaluations of educational programs. Thousand Oaks, CA: Sage.

Locke, L. F., Spirduso, W. W., & Silverman, S. J. (1987). Proposals that work: A guide for planning dissertations and grant proposals (2nd ed.). Newbury Park, CA: Sage.

Reinhardt, B. M. (1991, January). Factors affecting coefficient alpha: A mini Monte Carlo study. Paper presented at the annual meeting of the South west Educational Research Association, San Antonio. (ERIC Document Reproduction Service No. ED 327 574)

Reynolds, C. R. & Kamphaus, R. W. (1992). Behavioral Assessment System for Children. Circle Pines, MN: American Guidance Service, Inc.

Rowley, G. L. (1976). The reliability of observational measures. American Education Research Journal, 13, 51-59.

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

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. American Journal of Psychology, 18, 161-169.

Spearman, C. (1913). Correlations of sums and differences. British Journal of Psychology, 5, 417-426.

Thompson, B. (1991). A primer on the logic and use of canonical correlation analysis. Measurement and Evaluation in Counseling and Development, 24, 80-95.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-438.

Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

Thompson, B. & Crowley, S. (1994, April). When classical measurement theory is insufficient and generalizability theory is essential. Paper presented at the annual meeting of the Western Psychological Association, Kailua-Kona, Hawaii. (ERIC Document Reproduction Service No. ED 377 218)

Webb, N. M, Rowley, G. L. & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81-90.

Table 1
Reliability Calculated as the Correlation between True Scores and Observed Scores

2nd Graders	True Score	Observed Score
Jane	57	58
Julio	96	95
Max	63	62
Maria	100	99
Jason	63	63
Leticia	98	100
Margaret	65	64
Anna	98	98
Michael	65	65
David	96	97
Emily	65	66
Sara	96	95
Cathy	68	67
Arnoldo	93	94
Ramiro	67	68
Wayne	93	93
Mitchell	68	69
Matthew	90	91
April	70	71
Amy	88	89
Joan	72	73
Craig	88	87
Linda	75	75
Ruth	85	85
Todd	76	77
Andrew	82	83
Fred	79	79
Juan	82	82
Virginia	81	80
Sum	2319	2325
Mean	77.3	77.5

Reliability of the Entire Class

r_{xx}= .9977

Reliability of the top 5 persons in the Class (bolded names)

r_{xx}= .7456

Table 2
Example of an Interaction Effect that is not Detected Using Classical Test Theory Reliability Estimates

Criteria	Judge #1's Ratings	Judge #2's Ratings	Total Score
Criterion 1	0	1	1
Criterion 2	0	1	1
Criterion 3	0	1	1
Criterion 4	0	1	1
Criterion 5	0	1	1
Criterion 6	1	0	1
Criterion 7	1	0	1
Criterion 8	1	0	1
Criterion 9	1	0	1
Criterion 10	1	0	1
	Total Rating = 5	Total Rating = 5	Total Score = 10

Table 3
Ten Students' scores on a 7-item, nonspeeded test with coefficient alpha calculations

	Items								
	1	2	3	4	5	6	7	Score	
Todd	1	1	0	1	1	1	0	5	
Lu	0	1	1	1	1	1	1	6	
Nancy	1	1	1	1	1	1	1	7	
Tammi	1	1	0	1	0	1	0	4	
Karen	1	1	1	1	1	0	1	6	
Avery	0	1	1	1	1	1	1	6	
Art	1	0	1	0	1	0	1	4	
Cris	1	1	0	1	1	1	1	6	
Kris	1	1	1	1	1	1	1	7	
Brad	1	1	1	1	1	1	1	7	
P	0.8	0.9	0.7	0.9	0.9	0.8	0.8	5.8	M
V	0.16	0.09	0.21	0.09	0.09	0.16	0.16	1.16	Total test variance
								1.17	K/K-1
0.20 =α								0.96	Sum of item variances
								0.83	Sum of item variances/total test varian
								0.17	1- (sum of item variances/total test variance)

Table 4
Ten Students' scores on a 7-item, speeded test with coefficient alpha calculations

	Items								
	1	2	3	4	5	6	7	Score	
Todd	1	1	0	**0**	**0**	**0**	0	2	
Lu	0	1	1	1	1	1	0	5	
Nancy	1	**0**	**0**	**0**	**0**	**0**	**0**	1	
Tammi	1	1	0	1	0	**0**	0	3	
Karen	1	1	1	1	1	0	1	6	
Avery	0	1	1	1	1	1	1	6	
Art	1	0	1	0	1	0	1	4	
Cris	1	1	0	1	1	1	1	6	
Kris	1	1	1	1	1	1	1	7	
Brad	1	1	1	1	1	1	1	7	
P	0.8	0.8	0.6	0.7	0.7	0.5	0.6	4.7	M
ιτεμ σ2	0.16	0.16	0.24	0.21	0.21	0.25	0.24	4.01	Totai test variance
								1.17	K/K-1
0.74 =α								1.47	Sum of item variances
								0.37	Sum of item variances/total test varian
								0.63	1- (sum of item variances/total test variance)

Note. Answers that Todd, Lu, Nancy and Tammi had answered correctly in the nonspeeded test, but answered incorrectly on the speeded test are in bold.

Table 5
Minimal test score variance leads to minimal coefficient alpha

	Items							Score	
	1	2	3	4	5	6	7		
Buzz	0	0	0	0	0	0	0	0	
Meg	0	0	0	0	0	0	0	0	
Skip	1	1	1	1	0	0	0	4	
Jan	1	1	1	0	0	0	1	4	
Mark	1	1	0	0	0	1	1	4	
Joy	1	0	0	0	1	1	1	4	
Max	0	0	0	1	1	1	1	4	
Lucy	0	0	1	1	1	1	0	4	
Alex	0	1	1	1	1	0	0	4	
Gina	1	1	0	0	1	1	0	4	M
P	0.5	0.6	0.6	0.6	0.6	0.6	0.4	3.9	Total test variance
V	0.25	0.24	0.24	0.24	0.24	0.24	0.24	0.09	
									K/K-1
α=-21								1.17	Sum of item variances
								1.69	Sum of item variances/total test varianc
								18.8	1- (sum of item variances/total test va
								-18	

Table 6
Maximal test score variance leads to maximal coefficient alpha

	Items							Score	
	1	2	3	4	5	6	7		
Buzz	0	0	0	0	0	0	0	0	
Meg	0	0	0	0	0	0	0	0	
Skip	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	0	0	
Mark	0	0	0	0	0	0	0	0	
Joy	1	1	1	1	1	1	1	7	
Max	1	1	1	1	1	1	1	7	
Lucy	1	1	1	1	1	1	1	7	
Alex	1	1	1	1	1	1	1	7	
Gina	1	1	1	1	1	1	1	7	
P	0.5	0.5	0.5	0.5	0.5	0.5	0.5	3.5	M
V	0.25	0.25	0.25	0.25	0.25	0.25	0.25	12.3	Total test variance
α=1								1.17	K/K-1
								1.75	Sum of item variances
								0.14	Sum of item variances/total test varianc
								0.86	1- (sum of item variances/total test va

Table 7
Ten persons' scores on a seven item dichotomously scored test

	Items							Score	
	1	2	3	4	5	6	7	Score	
Buzz	0	0	0	0	0	1	1	2	
Meg	0	0	0	1	0	0	1	2	
Skip	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	1	1	
Mark	0	0	0	0	0	1	1	2	
Joy	0	0	0	0	1	1	1	3	
Max	0	0	0	1	1	1	1	4	
Lucy	0	0	1	1	1	1	1	5	
Alex	0	1	1	1	1	1	1	6	
Gina	1	1	1	1	1	1	1	7	
P	0.1	0.2	0.3	0.5	0.5	0.7	0.9	3.2	M
V	0.09	0.16	0.21	0.25	0.25	0.21	0.09	4.56	Total test variance
0.844 α								1.17	K/K-1
								1.26	Sum of item variances
								0.28	Sum of item variances/total test variance
								0.72	1- (sum of item variances/total test var

Table 8
Scores from test in Table 7 with two added items that everyone answers incorrectly

	Items									Score	
	1	2	3	4	5	6	7	8	9		
Buzz	0	0	0	0	0	1	1	0	0	2	
Meg	0	0	0	0	0	0	1	0	0	1	
Skip	0	0	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	1	0	0	1	
Mark	0	0	0	0	0	1	1	0	0	2	
Joy	0	0	0	0	1	1	1	0	0	3	
Max	0	0	0	1	1	1	1	0	0	4	
Lucy	0	0	1	1	1	1	1	0	0	5	
Alex	0	1	1	1	1	1	1	0	0	6	
Gina	1	1	1	1	1	1	1	0	0	7	
P	0.1	0.2	0.3	0.4	0.5	0.7	0.9	0	0	3.1	M
V	0.09	0.16	0.21	0.24	0.25	0.21	0.09	0	0	4.89	Total test variance
0.84 =α										1.13	K/K-1
										1.25	Sum of item variances
										0.26	Sum of item variances/total t
										0.74	1- (sum of item variances/to

Table 9
Scores from test in Table 7 with two added items that add score variability

	Items									Score	
	1	2	3	4	5	6	7	8	9		
Buzz	0	0	0	0	0	1	1	0	0	2	
Meg	0	0	0	0	0	0	1	0	0	1	
Skip	0	0	0	0	0	0	0	0	0	0	
Jan	0	0	0	0	0	0	1	0	0	1	
Mark	0	0	0	0	0	1	1	0	0	2	
Joy	0	0	0	0	1	1	1	0	0	3	
Max	0	0	0	1	1	1	1	0	0	4	
Lucy	0	0	1	1	1	1	1	0	0	5	
Alex	0	1	1	1	1	1	1	0	0	6	
Gina	1	1	1	1	1	1	1	1	1	9	
P	0.1	0.2	0.3	0.4	0.5	0.7	0.9	0.1	0.1	3.3	M
V	0.09	0.16	0.21	0.24	0.25	0.21	0.09	0.09	0.09	6.81	Total test variance
0.89	=α									1.13	K/K-1
										1.43	Sum of item variances
										0.21	Sum of item variances/total t
										0.79	1- (sum of item variances/to

Table 10
Scores from test in Table 7 with two added items that decrease score variability

	Items									Score	
	1	2	3	4	5	6	7	8	9		
Buzz	0	0	0	0	0	1	1	0	0	2	
Meg	0	0	0	0	0	0	1	0	0	1	
Skip	0	0	0	0	0	0	0	1	1	2	
Jan	0	0	0	0	0	0	1	1	1	3	
Mark	0	0	0	0	0	1	1	0	0	2	
Joy	0	0	0	0	1	1	1	0	0	3	
Max	0	0	0	1	1	1	1	0	0	4	
Lucy	0	0	1	1	1	1	1	0	0	5	
Alex	0	1	1	1	1	1	1	0	0	6	
Gina	1	1	1	1	1	1	1	0	0	7	
P	0.1	0.2	0.3	0.4	0.5	0.7	0.9	0.2	0.2	3.5	M
V	0.09	0.16	0.21	0.24	0.25	0.21	0.09	0.16	0.16	3.45	Total test variance
										1.13	K/K-1
0.61	= α									1.57	Sum of item variances
										0.46	Sum of item variances/total t
										0.54	1- (sum of item variances/to

Figure 1

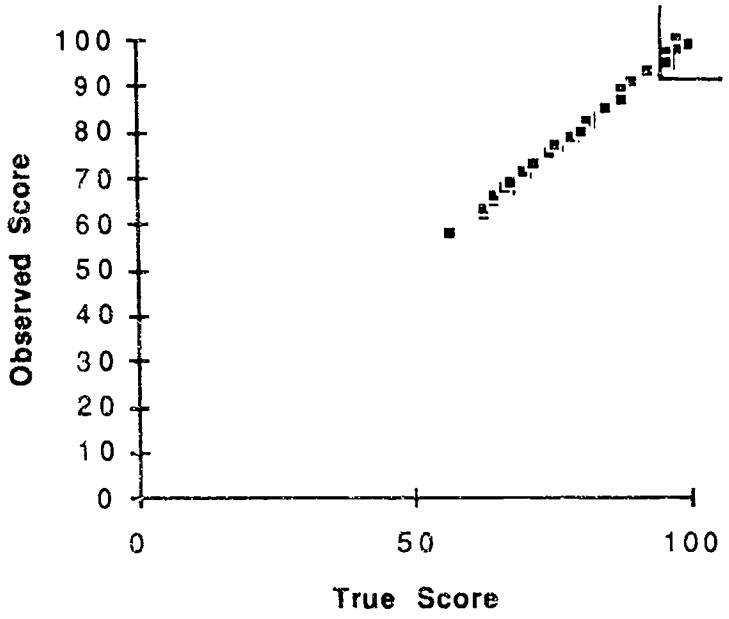

Figure 2