DOCUMENT RESUME

ED 314 434 TM 014 215

AUTHOR Eason, Sandra
TITLE Why Generalizability Theory Yields Better Results than Classical Test Theory.
PUB DATE Nov 89
NOTE 34p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Generalizability Theory; Test Reliability; *Test Theory
IDENTIFIERS *Classical Test Theory; *Error Analysis (Statistics); Research Results

ABSTRACT

Generalizability theory provides a technique for accurately estimating the reliability of measurements. The power of this theory is based on the simultaneous analysis of multiple sources of error variances. Equally important, generalizability theory considers relationships among the sources of measurement error. Just as multivariate inferential statistics consider relationships among variables that univariate statistics cannot detect, generalizability theory considers relationships of error measurement that classical theory cannot. An extensive discussion of the concept of reliability and its use in classical test theory and generalizability theory is presented. A comparison of classical test theory and generalizability theory illustrates how generalizability theory subsumes all other reliability estimates as special cases. A hypothetical data set provides examples of when the failure to use generalizability theory can lead to seriously erroneous estimates of test reliability. The framework of generalizability theory incorporates two stages of analysis: (1) a generalizability study; and (2) a decision study. The former analyzes the extent to which results are generalizable to a population, while the latter uses information from the generalizability study to determine other generalizability coefficients for variations of the measurement protocol. Six data tables are provided, and an appendix presents the GENOVA program code used. (TJH)

# WHY GENERALIZABILITY THEORY YIELDS BETTER RESULTS THAN CLASSICAL TEST THEORY

Sandra Eason

University of New Orleans 70148

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 8, 1989.

ABSTRACT

Generalizability theory provides a technique for most accurately estimating the reliability of measurements. The power of generalizability theory is based on the simultaneous analysis of multiple sources of error variances. A comparison of classical test theory and generalizability theory illustrates how generalizability theory subsumes all other reliability estimates as special cases. Further, a hypothetical data set provides examples of when the failure to use generalizability theory can lead to seriously erroneous estimates of test reliability.

## WHY GENERALIZABILITY THEORY YIELDS BETTER RESULTS THAN CLASSICAL TEST THEORY

Behavioral measurements that yield reliable results are of paramount importance for social scientists. Ghiselli (1964) suggests that quantitative descriptions which compare traits among and within individuals must give a precise characterization of an individual in order to be very useful. Nunnally (1982, p. 1589) notes that

> Science is concerned with repeatable experiments. If data obtained from experiments are influenced by random errors of measurement, the results are not exactly repeatable. Thus, science is limited by the reliability of measuring instruments and by the reliability with which scientists use them.

Historically, reliability of measurements has been determined by theory first articulated decades ago. This body of thought has come to be called classical test theory. Reliable information about individual differences is obtained by measurements that have minimum amounts of error variance and maximum amounts of systematic variance. Within classical test theory various coefficients are available for investigating single sources of error variance.

However, in classical theory consideration of multiple sources of error variance within one analysis is unavailable. The inability to analyze more than one source of error variance at a time severely limits classical test theory as a psychometric technique. With the conceptualization and

development of generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) the limitation of classical test theory, i.e., the inability to examine multiple sources of error variance simultaneously, was resolved. Further, generalizability theory provides a technique for more accurately estimating the reliability in measurements.

Reliability within classical test theory refers to how consistently test scores are measured under various circumstances (Gronlund, 1985; Nunnally, 1972). The importance of reliable measurement lies in how much confidence can be placed in the results. A measurement that consistently yields similar results over several administrations is dependable. Decisions based on the results can be made with confidence. However, unreliable results indicate the presence of error in a measurement. Data obtained from such a measurement are not dependable and consequently are of little value. Yet, the quest for the perfect test is a futile one. Measurement theorists suggest that no perfectly reliable measurement exists (Ebel & Frisbie, 1986; Nunnally, 1972). Inconsistency or measurement error is present in all instruments.

There is some confusion concerning the referent for reliability coefficients. Frequently, an instrument or test is referred to as having reliability. Such references are, strictly speaking, incorrect. Data, not a test, have the characteristics of reliability. To illustrate, if the Scholastic Achievement Test (SAT) was administered in a city that had suffered a devastating tornado the day before, test

scores would likely not reflect true abilities of the examinees. Concentration on the SAT would have been hindered by the emotional upheavel and fatigue caused by the disaster. Thus, the data from the SAT would be unreliable, not the SAT. Similarly, a high school history exam given immediately before an important pep rally for a football district championship might appear unreliable. In an otherwise dependable test, low reliability must be attributable to the test results rather than the actual test itself. In addition to these factors, other common factors that contribute to inconsistency of test scores are anxiety and guessing.

Understanding the functional characteristics of reliability leads researchers to recognize that measurement error can also invalidate significance testing. Nunnally (1972) states that attentuation occurs because measurement error tends to reduce correlations, i.e., makes them closer to zero. Thus, measurement error obscures true effect sizes. For example, if a student with a high ability for geography was measured by an unreliable geography test, the observed score would not reflect the student's actual ability. And systematic effects in experimental investigations would have been blurred by the prescence of error variance. Researchers strive to eliminate error so that observed scores reflect the actual capabilities of students, not extraneous factors. Thus, as information related to measurement error sources is gained, the greater are the chances of reducing measurement error and of detecting systematic influences via significance testing.

Generalizability theory considers the multiple sources of error that may influence scores. Although introduced as generalizability theory by Cronbach, Gleser, Nanda, and Rajaratnam (1972), related developments upon which generalizability theory was based were reported earlier (Hoyt, 1941; Lindquist, 1953; Medley & Mitzel, 1963). Generalizability theory provides the framework to simultaneously examine multiple sources of error variance. By so doing, measurement reliability can be more optimally maximized through better informed test revision.

The purpose of the present paper is to provide an introduction to the powerful measurement theory called generalizability theory. A comparison of classical test theory and generalizability theory will illustrate how generalizability theory subsumes all other reliability estimates as special cases. In addition, the paper will demonstrate that failure to use generalizability theory can lead to seriously erroneous estimates of test reliability.

## Classical Test Theory

In classical test theory observed score variance is partitioned into true score variance and error variance. If tests were perfectly reliable, true score variance would equal observed score variance. However, since several error factors exist, classical theory provides estimates for at least three types of reliability: internal consistency, stability, and equivalence. Each reliability estimate considers one source of error either error in items, test occasions, or test forms. The

estimated reliability is represented by a coefficient which indicates the ratio of true score variance to total observed score variance.

Clarification of the types of reliability coefficients will be facilitated by the inclusion of a hypothetical measurement situation. A subsequent generalizability analysis will be performed on the same data to provide a comparison of the two theories.

The hypothetical example depicts a researcher attempting to establish a measurement protocol that reliably assesses college student attitudes towards the teaching profession. The instrument is administered to college seniors majoring in education. The researcher, being rather ambitious, hopes for a lucrative future as a psychometrician and therefore exerts sufficient energy to design two parallel forms of the same test. In a pilot study, six students are each administered the two parallel forms in which each form contains a different set of five items, i.e., items are nested in each test. Each of the forms is administered on two occasions. A hypothetical data set, describing the example, is presented in Table 1. The small sample size of the data set was intentional so readers who wish to pursue the paper's purpose may replicate these analyses.

---

INSERT TABLE 1 ABOUT HERE.

---

Classical reliability coefficients for the research situation are presented in Table 2. Each of the three types of reliability examines a separate source of variance. Internal

consistency examines the homogeneity of performance of items within a test. In internal consistency analysis the items can be evaluated in a variety of ways depending on the coefficient selected, e.g., split-half, coefficient alpha, and Kuder-Richardson 20 (Ebel & Frisbie, 1986). A high reliability suggests that the items are homogeneous with respect to statistical characteristics of interest. Table 2 presents four internal consistency reliabilities analyzed on each of the parallel forms over the two occasions. The hypothetical young researcher in the senario has found the varying reliabilities disturbing. Form A's reliability estimate is 0.48 on occasion one but 0.91 on occasion two. Cognizant that expertly designed tests often yield reliability coefficients of 0.90 or higher (Ebel & Frisbie, 1986; Nunnally, 1972), the researcher is perplexed about the true reliability of the data from Form A. In addition, the reliability coefficients of Form B are more stable but offer no consolation because of the estimates of 0.72 and 0.74 indicate the presence of substantial measurement error.

---

INSERT TABLE 2 ABOUT HERE.

---

Undaunted by the confusing internal consistency reliabilities, this ambitious hypothetical researcher continues with other reliability analyses. A second reliability estimate was obtained by evaluating stability. Stability of a test scores is estimated to determine how stable an instrument is over time. A high degree of reliability indicates that measurements given on two occasions are relatively the same.

However, a low reliability coefficient, indicates instability is present. The stability coefficients of Ø.93 and Ø.88 are presented in Table 2. Encouraged by the high reliability, the researcher infers that student attitudes have remained consistent over the two occasions.

The researcher is now confidently ready to test for a third type of reliability, equivalence. Equivalence reliability indicates the degree to which parallel forms of a test measure the same domain of interest. By correlating the scores of the six students, each taking Form A and Form B, an equivalence reliability coefficient is obtained. High reliability would indicate that the rank ordering on the two forms remained relatively unchanged and that the parallel forms could be used interchangeably with confidence. Unfortunately for the aspiring psychometrician, Form A and Form B were estimated to have a relatively unstable equivalence reliability. The coefficients of Ø.88 and Ø.72, presented in Table 2, indicate that the two forms were measuring somewhat different attitudes in regard to the teaching profession.

The researcher has now obtained three types of reliability coefficients: internal consistency, stability, and equivalence. However, the coefficients have confused rather than clarified the reliability of the attitude measure because different estimates yield contradictory results. The coefficients of internal consistency and equivalence perplex the researcher as to what to do to increase the reliability of the attitude measurement. Zealous to become famous, the researcher

determines by reading the scientific journals that generalizability theory is more appropriate than classical theory and can better address the inconsistencies presented by the data.

## Generalizability Theory

Generalizability theory (G theory) subsumes classical theory as a special case. G theory encompasses the concepts of classical theory as well as accomodating complex measurement designs. The power of G theory lies in the consideration of multiple sources of error variance simultaneously. Classical test theory is limited to analyses of single sources of error variance (Thompson, 1989a; Webb, Rowley, & Shavelson, 1988).

The two theories estimate measurement characteristics using different frameworks. A reliability coefficient in classical theory concerns the dependability of an instrument or procedure that is to be used on different occasions or with different forms. If over several test administrations the results remain relatively the same, the instrument is said to yield reliable information. Contrastingly, G theory looks not at how reliable an instrument is over varying situations but rather how generalizable the results are to a universe. A generalizability coefficient represents the ratio of universe score variance (systematic variance) to observed score variance. The fundamental differences between classical test theory and generalizability theory have been stated by Shavelson, Webb, and Rowley (1989, p. 922):

The concept of reliability, so fundamental to classical

theory, is replaced by the broader and more flexible notion of generalizability. Instead of asking how accurately observed scores reflect their corresponding true scores, generalizability theory asks how accurately observed scores permit us to generalize about persons' behavior in a defined universe of situations.

The framework of generalizability theory incorporates two stages of analyses. The first stage analyzes the degree that results are generalizable to a population and is termed a generalizability study (G study). The second stage, decision study (D study), uses information from the G study to determine other generalizability coefficients for variations of the measurement protocol. In other words, a G study estimates magnitudes of error variance and a D study uses the information to determine the best measurement design to get the most reliable scores in the most efficient manner.

The conceptual foundation of a G study is based on a universe of admissible observations. This universe is an infinite set of conditions from which the sampling is representative. Within the universe of admissible observations are variables or areas of measurement called facets. Facets provide information about the multiple sources and amounts of error in a measurement. Facets can be of many types. Items, tests, occasions, raters, or observers are facets typically of interest to researchers. For example, in a G study designed to measure the oral English proficiency of foreign teaching

assistants, the facets of raters and occasions formed the universe of admissible observations (Bolus, Hinofotis, & Bailey, 1982). Facets are samples from a universe of all possible items, tests, occasions, raters, or observers, i.e., from the universe of admissible observations. Further, each facet is composed of conditions which vary. Thus, a G study takes into consideration a representative sample from a population of factors or variables, i.e., facets, with each having a range of conditions. Shavelson, Webb, and Burstein (1986) present generalizability studies that illustrate these issues.

## G Study Analyses

Bringing the previous example of the tenure seeking researcher into a generalizability context, the measurement design provides concrete examples of the terms germaine to generalizability theory. The researcher, somewhat mystified and weary from estimating individual reliabilities in the classical approach, hopes to salvage the remains of previous efforts. To further pursue the noble goal of a highly generalizable attitude measure, the researcher determines that the universe of admissible observations of the G study will contain the facets of items, forms, and occasions. After careful thought, the researcher defined the facets. Items would reflect attitudes toward the teaching profession with two forms of the test given on two occasions three weeks apart.

The development of a comprehensive design was due to the researcher's extensive knowledge newly gained from the library. Coefficients of generalizability can only be estimated to the

degree that the universe of admissible observations has been defined (Brennan, 1983; Shavelson, Webb, & Rowley, 1989). Desiring the highest generalizability, the researcher optimizes the research design by including all facets that could affect generalizability. For example, without testing for error from forms or from more than one occasion, information is unattainable as to the error that may originate from this source (Thompson, 1989a). In summary, for a G study to provide the most accurate estimate of generalizability, all facets, representing error variance within the measurement design, must be included in the analysis.

One additional generalizability term not previously introduced is object of measurement. Object of measurement usually refers to persons and in the above scenario specifically refers to senior education students. However, in a study on school-level variables, schools were the object of measurement (O'Brien & Jones, 1986). An object of measurement is the variance which the researcher considers legitimate, e.g., student ability variations on a posttest in an experiment, and about which the researcher wishes to generalize. Facets contain error variance. Objects of measurement contain systematic variance and are analogous to the classical true score variance. In generalizability the estimated variance component for persons is the universe score variance. The remaining variance components represent error variance.

A G study employs the statistical procedure of analysis of variance (ANOVA) to estimate variance components. Variance

components are central to the framework of generalizability theory. Brennan (1983) suggests the importance of variance components: "generalizability theory emphasizes the estimation, use, and interpretation of variance components associated with universes" (p. xiii). For several years variance components were employed in statistical analyses (Guilford, 1950). However, the use of mean squares from which variance components are determined changed as F statistics and F tests became more popular (Brennan, 1983). The overriding concern of researchers became statistical significance testing. The importance of such tests appear to be prevalent today. In a recent article, Thompson (1989b) suggests that too many researchers attend only to statistical significance disregarding other important issues such as effect size and replicability. Consequently, researchers may be unfamiliar with the use of mean squares for estimating variance components. Nevertheless, the concept of estimated variance components, not statistical significance, is important in generalizability theory.

The hypothetical researcher's measurement study incorporates 6 x 2 x 2 x 5 design with items nested within the tests. Nested items (I:T) refer to each person responding to a different set of items for each test (the score of person P on item I nested in both test T) in contrast to a crossed design where each person would respond to the same items on each test (the score of person P on item I in both test T1 and T2). Partitioning through a factorial ANOVA provides estimated variance components for the sources of variation in this

example: Persons (P), Occasions (O), Tests (T), and Items nested in the Test (I:T), the two-way interactions PO, PT, PI:T, OT, and OI:T, and the three-way interactions POT and POI:T and error. Table 3 presents the ANOVA for Table 1 data. The GENOVA computer program was used to calculate the analysis (Brennan, 1983).

INSERT TABLE 3 ABOUT HERE.

Using an ANOVA, specifically using the mean squares, a G study determines the estimated variance components. Of concern within these various methods of estimating variance components is a means to treat negative estimates. Estimates with negative variance sometime occur but are conceptually not possible. Variance can never be negative. Several methods are available to calculate variance components and to resolve negative estimates (Shavelson, Webb, & Rowley, 1989). In some methods the components are converted to zero. GENOVA uses the two methods of (a) algorithms and (b) expected mean square equations (EMS) to estimate variance components. These estimates are presented in Table 4. Thompson (1989a) provides a non-technical discussion with mathematical examples of variance components.

INSERT TABLE 4 ABOUT HERE.

The generalizability calculations for the data in Table 1 are presented in Table 5. Since the objective of the researcher's measurement was to obtain scores reflecting

individual differences of attitudes toward the teaching profession, a relatively large variance component (Ø.59) for the object of measurement, persons, was reassuring. The astute researcher knows that in an accurately measuring instrument most of the observed variance is systematic variance. Drawing more careful consideration from the researcher were the error components. Although seven of the variance components reflected little or no error, three components were troublesome. A two-way interaction between persons and items involved an error component of Ø.65. The relatively large component suggested that persons were inconsistent in their attitudes across items. Another variance component represented the three-way interaction of persons by occasions by items (Ø.33) was troublesome. Interactions, especially three-way interactions, are difficult to explain. For this data set, the explanation may have been that individual attitude items by individual persons across occasions lacked consistency. The final variance component which reflected error in measurement was the main effect variance component for test (Ø.17). The estimate indicated that Form A and Form B were correlated, but not so highly as might be hoped. With all of the information from the variance components, the researcher is anxious to obtain the long awaited coefficient of generalizability. Was fame and fortune just one coefficient away?

---

INSERT TABLE 5 ABOUT HERE.

---

However, before the researcher's curiosity could be satisfied, another theoretical source of great importance became apparent. The researcher became aware of two types of G coefficients. One important feature of generalizability theory that classical test theory is unable to address is the distinction between relative and absolute decisions (Shavelson, Webb, & Rowley, 1989). Relative decisions are based solely on a person's rank order within a group, such as a score in the 90th percentile on a norm-referenced test. For instance, the California Achievement Test provides percentiles for the purpose of comparing the ability of one student to the ability of other students in several academic areas. A specific score is not used as a reference. The researcher cares only whether the relative position of the object of measurement is consistent across measurements, and does not care about the scores per se.

Absolute decisions, on the other hand, involve concerns both about consistency of relative placement and about consistency of placement in relation to some absolute criterion such as a cutoff score or reference point. Several professions come to mind where competence must be demonstrated in relation to an absolute standard. Medical personnel, certified public accountants, and lawyers must achieve a passing score before being granted a license to practice. Similarly, an applicant for a driver's license must demonstrate a set level of competency on a driving test before legally getting behind the wheel of a car. For example, in a generalizability study on constructing diagnostic test profiles, the major purpose of

testing was to assess individual status with respect to the knowledge domain of pronoun usage (Webb, Herman, & Cabello, 1987). Mastery of the domain provided the basis for decisions from the diagnostic test. In short, the purpose of the measurement determines which type of coefficient is appropriate.

Error variance for relative and absolute decisions is estimated using different combinations of variance components. A relative decision is determined only by the variance components that affect the relative standing of an individual in a group. For instance, in the hypothetical aspiring researcher's nested design, a relative decision was determined by the variance components which interact with the object of measurement, i.e., PO, PT, PI:T, and error. Main effect variance components are not reflective of the relative standing of an individual and are not included in the analysis. At last, with unbounded enthusiasm the researcher obtains from Table 5 the generalizability coefficient, Ø.86. Although not as high as might be hoped, the researcher accepts the coefficient with a degree of relief. The powerful measurement technique of G theory has resolved the confusing conflicting reliability coefficients of classical test theory by yielding one coefficient representing the generalizability of the attitude instrument. The generalizability coefficient of Ø.86 represents the degree that scores are generalizable for a relative decision.

However, if an absolute decision had been the researcher's focus, all facet variance components including main effects

would have been used in the generalizability calculations: O, T, I:T, PO, PT, PI:T, OT, OI:T, POT, and POI:T, and error. Needless to say, the researcher was relieved that absolute decisions would not be necessary since the generalizability coefficient of Ø.77 declined, as represented by phi in Table 5. Further discussion and formulas relating to relative and absolute decision are available by Brennan (1983), Gillmore (1983), or Webb, Rowley, and Shavelson (1988). Classical test theory, unlike generalizability theory, cannot distinguish the differential reliability of scores employed for relative as against absolute decisions (Brennan, 1983, p. 18), again reflecting the limits of classical theory.

## D Study Analyses

The newly energized researcher forges ahead to the second stage of generalizability theory, the decision study (D study). D studies use variance components information from the G study to design a measurement protocol that both minimizes error variance and is most efficient, i.e., yields the most reliable scores with the least effort. Shavelson, Webb, and Rowley (1989, p. 925) state, "In distinguishing a G study from a D study, G theory recognizes that the former is associated with the development of a measurement procedure whereas the latter then applies the procedure."

A concept central to a D study is the universe of generalization. The concept refers to the universe the researcher wishes to generalize. A D study can include all of the facets in the universe of admissible observations, or a

reduction in one level or condition of a facet, or a facet can even be eliminated. However, a D study cannot include facets that were not present in the universe of admissible observations during the G study. Conditions to be sampled can vary in a D study but must be present in the G study so that the necessary variance components are available to estimate the effects of various changes in the measurement protocol.

Within the D study analysis, the researcher alters the measurement design of the G study by varying the conditions of the facets. The analysis is performed by dividing the variance components estimated in the G study by the number of levels in their facet design. For example, one D study analyzed a design with one occasion, one form, and five items and yields an estimated generalizability coefficient of 0.70. In another analysis with a similar design containing 10 items, the coefficient increased to 0.79. By increasing the items to 25 in one test, one occasion design, the coefficient increased to 0.85. The improvement in the coefficient by increasing the items is reasonable since in the present example a large error component was present for person by items nested in a test interaction. The use of more items divides the variance from this measurement error source by a larger number, resulting in a larger estimated generalizability. Therefore, the D study provided two measurement designs by which the researcher can achieve a similar degree of generalizability; either two tests with five items each given on two occasions (0.86) or one test with 25 items given once (0.85). If the researcher is satisfied

with this outcome, the protocol which is most efficient or practical can now be selected in an informed manner.

Failure to use G theory, however, could have led the researcher to very seriously erroneous estimates of test reliability (Thompson, 1989a). If the researcher had administered only Form A of the test on the first occasion, the classical reliability of Ø.48 would have suggested to the researcher that the project be abandoned. In addition, if the researcher had measured for stability of Form A, a Ø.93 reliability estimate would have stimulated unwarranted confidence in the measure. Deflated estimates of reliability would have been obtained if only Form B's internal consistency on occasion one (Ø.73) and occasion two (Ø.74) had been computed. Importantly, a total of four of the eight coefficients in Table 2 would have been lower than the generalizability coefficient of Ø.86.

One final interesting data set will further clarify the discussion concerning potentially erroneous classical reliabilities. Table 6 presents a similar data set representing the same measurement design. However, subjects score identical results for Form A and Form B on each occasion. A classical reliability of equivalence yields an incredible 1.Ø--perfectly reliable forms! Conversely, a G study indicates that the measurement's generalizability is Ø.82. Error, present in the measurement design, went undetected by the single-source reliability estimate in the classical approach. Although both the single analysis and the identical scores in the data set are

unrealistic, they do demonstrate a point. Multiple sources of error variance are important. Generalizability theory provides the framework needed to determine the influence of measurement error. Only generalizability theory can simultaneously consider all the multiple sources of measurement error.

INSERT TABLE 6 ABOUT HERE.

Put differently, a researcher may calculate internal consistency, stability, and equivalence reliability coefficients to all be 0.90 for a data set, and yet the generalizability coefficient for the same data might be 0.60 because only generalizability theory considers the interaction of measurement error sources. Only generalizability theory honors complex reality in which measurement error sources may interact to compound each other!

## Conclusion

Measurement theory has advanced beyond classical test theory. A more powerful analysis, generalizability theory, considers all sources of error variance simultaneously. Equally important, generalizability theory considers relationships among the sources of measurement error. Just as multivariate inferential statistics considers relationships among variables that univariate statistics cannot detect, generalizability theory considers relationships of error measurement that classical theory cannot. Nunnally (1982) suggests that generalizability theory goes even beyond the evaluation of

measurement error:

> There really is no sharp borderline dividing studies of reliability from studies of validity. Consequently, the concepts and mathematical models relating to generalizability theory can be extended to wider, more important issues in the behavioral sciences than just the investigation of measurement error. (p. 1600)

Thus, there is every possibility that reflective researchers will increasingly turn to generalizability theory as the measurement model of choice.

## References


Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. Language Learning, 32, 245-258.

Brennan, R. L. (1933). Elements of generalizability theory. Iowa City, IA: ACT Publications.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of measurements. New York: Wiley & Sons.

Ebel, R. L., & Frisbie, D. A. (1986). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Ghiselli, E. E. (1964). Theory of psychological measurement. New York: McGraw-Hill.

Gilmore, G. M. (1983). Generalizability theory: Applications to program evaluation. In L. J. Fyans, Jr. (Ed.), New Directions for Testing and Measurement (pp. 3-16). San Francisco: Jossey-Bass.

Gronlund, N. E. (1985). Measurement and evaluation in teaching (5th. ed.). New York: MacMillan.

Guilford, J. P. (1950). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Lindquist, E. F. (1953). Design and analysis of experiments in education and psychology. Boston: Houghton Mifflin.

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching (pp. 247-328). Chicago: Rand McNally.

Nunnally, J. C. (1972). Educational measurement and evaluation (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C. (1982). Reliability of measurement. In H. E. Mitzel (Ed.), Encyclopedia of Educational Research (pp. 1589-1601). New York: Free Press.

O'Brien, R. M., & Jones, B. (1986). The reliability of school-level aggregate variables: An application of generalizability theory. Journal of Research and Development in Education, 20, 21-27.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), Handbook of Research on Teaching (3rd. ed.) (pp. 50-91). New York: Macmillan.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Thompson, B. (1989a, January). Why generalizability coefficients are an essential aspect of reliability assessment. Paper presented at the meeting of the Southwest Educational Research Association, Houston, TX.

Thompson, B. (1989b). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. Measurement and Evaluation in Counseling and Development, 22, 2-5.

Webb, N. M., Herman, J. L., & Cabello, B. (1987). A domain-referenced approach to diagnostic testing using generalizability theory. Journal of Educational Measurement, 24, 119-130.

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
Data for Study Example

| | Occasion | Form | Items | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| PERSON 1 | 1 | A | 4 | 4 | 3 | 5 | 2 | 18 |
| | | B | 2 | 3 | 2 | 2 | 3 | 12 |
| | 2 | A | 4 | 4 | 4 | 5 | 3 | 20 |
| | | B | 2 | 2 | 1 | 2 | 3 | 10 |
| PERSON 2 | 1 | A | 5 | 2 | 5 | 4 | 5 | 21 |
| | | B | 4 | 3 | 4 | 4 | 2 | 17 |
| | 2 | A | 5 | 2 | 5 | 4 | 4 | 20 |
| | | B | 5 | 3 | 4 | 5 | 2 | 19 |
| PERSON 3 | 1 | A | 4 | 3 | 4 | 4 | 4 | 19 |
| | | B | 5 | 3 | 3 | 2 | 3 | 16 |
| | 2 | A | 3 | 4 | 4 | 4 | 4 | 19 |
| | | B | 5 | 4 | 3 | 2 | 4 | 18 |
| PERSON 4 | 1 | A | 4 | 1 | 3 | 2 | 2 | 12 |
| | | B | 3 | 1 | 2 | 1 | 1 | 8 |
| | 2 | A | 2 | 2 | 2 | 2 | 1 | 9 |
| | | B | 2 | 2 | 1 | 3 | 2 | 10 |
| PERSON 5 | 1 | A | 2 | 3 | 4 | 2 | 4 | 15 |
| | | B | 3 | 1 | 3 | 3 | 2 | 12 |
| | 2 | A | 2 | 3 | 3 | 2 | 2 | 12 |
| | | B | 1 | 1 | 3 | 2 | 1 | 8 |
| PERSON 6 | 1 | A | 5 | 4 | 5 | 3 | 2 | 19 |
| | | B | 3 | 2 | 4 | 5 | 5 | 19 |
| | 2 | A | 5 | 4 | 5 | 4 | 4 | 22 |
| | | B | 2 | 2 | 5 | 5 | 5 | 19 |

Table 2
Classical Test Theory Reliabilities

Internal consistency reliabilities:

| | |
|---|---|
| Form A, Occasion 1 | 0.48 |
| Form B, Occasion 1 | 0.73 |
| Form A, Occasion 2 | 0.91 |
| Form B, Occasion 2 | 0.74 |

Stability reliability:

| | |
|---|---|
| Form A, Occasions 1 & 2 | 0.93 |
| Form B, Occasions 1 & 2 | 0.88 |

Equivalence reliability:

| | |
|---|---|
| Forms A & B, Occasion 1 | 0.88 |
| Forms A & B, Occasion 2 | 0.72 |

Table 3
Random Effects ANOVA from GENOVA

| EFFECT | DEGREES OF FREEDOM | SUMS OF SQUARES FOR MEAN SCORES | SUMS OF SQUARES FOR SCORE EFFECTS | MEAN SQUARES |
|---|---|---|---|---|
| Persons | 5 | 1234.20000 | 68.56667 | 13.71333 |
| Occasions | 1 | 1165.66667 | 0.03333 | 0.03333 |
| Tests | 1 | 1177.66667 | 12.03333 | 12.03333 |
| Items:T | 8 | 1190.66667 | 13.00000 | 1.62500 |
| PO | 5 | 1237.40000 | 3.16667 | 0.63333 |
| PT | 5 | 1252.40000 | 6.16667 | 1.23333 |
| PI:T | 40 | 1330.00000 | 64.60000 | 1.61500 |
| OT | 1 | 1177.73333 | 0.03333 | 0.03333 |
| OI:T | 8 | 1193.33333 | 2.60000 | 0.32500 |
| POT | 5 | 1258.80000 | 3.16667 | 0.63333 |
| POI:T | 40 | 1352.00000 | 13.00000 | 0.32500 |
| MEAN | | 1165.63333 | | |
| TOTAL | 119 | | 186.36667 | |

NOTE: FOR GENERALIZABILITY ANALYSES, F-STATISTICS SHOULD BE IGNORED

Table 4
Variance Components Estimated from Random Effects ANOVA

| EFFECT | DEGREES OF FREEDOM | MODEL VARIANCE COMPONENTS USING ALGORITHM | MODEL VARIANCE COMPONENTS USING EMS EQUATIONS | STANDARD ERROR |
|---|---|---|---|---|
| Persons | 5 | 0.6240000 | 0.5895000 | 0.3687613 |
| Occasions | 1 | 0.0000000 | -0.0100000 | 0.0080050 |
| Tests | 1 | 0.1898333 | 0.1683333 | 0.1647915 |
| Items:T | 8 | 0.0008333 | 0.0008333 | 0.0686416 |
| PO | 5 | 0.0000000 | 0.0000000 | 0.0478755 |
| PT | 5 | -0.0690000 | -0.0690000 | 0.0823673 |
| PI:T | 40 | 0.6450000 | 0.6450000 | 0.1797435 |
| OT | 1 | -0.0200000 | -0.0200000 | 0.0125388 |
| OI:T | 8 | 0.0000000 | 0.0000000 | 0.0269541 |
| POT | 5 | 0.0616667 | 0.0616667 | 0.0691760 |
| POI:T | 40 | 0.3250000 | 0.3250000 | 0.0709208 |

NOTE: THE "ALGORITHM" AND "EMS" ESTIMATED VARIANCE COMPONENTS WILL BE IDENTICAL IF THERE ARE NO NEGATIVE ESTIMATES

Table 5
Generalizability Calculations from GENOVA

VARIANCE COMPONENTS IN TERMS OF D STUDY UNIVERSE (OF GENERALIZATION) SIZES

| EFFECT | VARIANCE COMPONENTS FOR SINGLE OBSERVATIONS | FINITE UNIVERSE CORRECTIONS | D STUDY SAMPLING FREQUENCIES | VARIANCE COMPONENTS FOR MEAN SCORES | |
|---|---|---|---|---|---|
| | | | | ESTIMATES | STANDARD ERRORS |
| Persons | 0.58950 | 1.0000 | 1 | 0.58950 | 0.36876 |
| Occasions | 0.00000 | 1.0000 | 2 | 0.00000 | 0.00400 |
| Tests | 0.16833 | 1.0000 | 2 | 0.08417 | 0.08240 |
| Items:T | 0.00083 | 1.0000 | 10 | 0.00008 | 0.00686 |
| PO | 0.00000 | 1.0000 | 2 | 0.00000 | 0.02394 |
| PT | 0.00000 | 1.0000 | 2 | 0.00000 | 0.04118 |
| PI:T | 0.64500 | 1.0000 | 10 | 0.06450 | 0.01797 |
| OT | 0.00000 | 1.0000 | 4 | 0.00000 | 0.00313 |
| OI:T | 0.00000 | 1.0000 | 20 | 0.00008 | 0.00135 |
| POT | 0.06167 | 1.0000 | 4 | 0.01542 | 0.01729 |
| POI:T | 0.32500 | 1.0000 | 20 | 0.01625 | 0.00355 |

| | VARIANCE | STANDARD DEVIATION | STANDARD ERROR OF VARIANCE |
|---|---|---|---|
| UNIVERSE SCORE | 0.58950 | 0.76779 | 0.36876 |
| EXPECTED OBSERVED SCORE | 0.68567 | 0.82805 | 0.36650 |
| LOWER CASE DELTA | 0.09617 | 0.31011 | 0.04074 |
| UPPER CASE DELTA | 0.18042 | 0.42475 | 0.08864 |
| MEAN | 0.19853 | 0.44556 | |

GENERALIZABILITY COEFFICIENT = 0.85975 ( 6.12998)
PHI = 0.76567 ( 3.26744)

NOTE: SIGNAL/NOISE RATIOS ARE IN PARENTHESES

Table 6
Data for Example

| | Occasion | Form | Items | | | | | Total |
|---|---|---|---|---|---|---|---|---|
| PERSON 1 | 1 | A | 2 | 1 | 2 | 1 | 1 | 7 |
| | | B | 2 | 1 | 2 | 1 | 1 | 7 |
| | 2 | A | 1 | 2 | 3 | 3 | 3 | 12 |
| | | B | 1 | 2 | 3 | 3 | 3 | 12 |
| PERSON 2 | 1 | A | 2 | 1 | 2 | 5 | 5 | 15 |
| | | B | 2 | 1 | 2 | 5 | 5 | 15 |
| | 2 | A | 2 | 1 | 2 | 3 | 3 | 11 |
| | | B | 2 | 1 | 2 | 3 | 3 | 11 |
| PERSON 3 | 1 | A | 3 | 3 | 4 | 2 | 3 | 15 |
| | | B | 3 | 3 | 4 | 2 | 3 | 15 |
| | 2 | A | 3 | 2 | 4 | 1 | 2 | 12 |
| | | B | 3 | 2 | 4 | 1 | 2 | 12 |
| PERSON 4 | 1 | A | 4 | 4 | 5 | 4 | 5 | 22 |
| | | B | 4 | 4 | 5 | 4 | 5 | 22 |
| | 2 | A | 4 | 3 | 4 | 5 | 5 | 21 |
| | | B | 4 | 3 | 4 | 5 | 5 | 21 |
| PERSON 5 | 1 | A | 3 | 4 | 4 | 3 | 4 | 18 |
| | | B | 3 | 4 | 4 | 3 | 4 | 18 |
| | 2 | A | 4 | 4 | 3 | 4 | 4 | 19 |
| | | B | 4 | 4 | 3 | 4 | 4 | 19 |
| PERSON 6 | 1 | A | 3 | 2 | 1 | 4 | 2 | 12 |
| | | B | 3 | 2 | 1 | 4 | 2 | 12 |
| | 2 | A | 2 | 2 | 2 | 3 | 5 | 14 |
| | | B | 2 | 2 | 2 | 3 | 5 | 14 |

Appendix A
GENOVA Program Code

```
GSTUDY      @@@@@@@@  - GENERALIZABILITY THEORY -
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + O 2 0
EFFECT      + T 2 0
EFFECT      + I:T 5 0
NAME        P "Person"
NAME        O "Occasion"
NAME        T "Test Form"
NAME        I "Items within Test"
FORMAT      (29X,5F5.0)
REWIND      9
PROCESS     9 DEFAULT
DSTUDY      #1   P x O x T x I:T MODELS use ems    --GENER9
DEFFECT     $ P    6   6   6   6   6   6   6   6
DEFFECT       O    2   1   1   2   1   1   1   1
DEFFECT       T    2   1   2   1   1   1   1   1
DEFFECT       I:T  5   5   5   5  10  15  20  25
ENDDSTUDY
DSTUDY       #2   P x O x T x I:T MODELS use algorithm   --GENER9

OPTIONS     ALGORITHM
DEFFECT     $ P    6   6   6   6   6   6   6   6
DEFFECT       O    2   1   1   2   1   1   1   1
DEFFECT       T    2   1   2   1   1   1   1   1
DEFFECT       I:T  5   5   5   5  10  15  20  25
ENDDSTUDY
GSTUDY          ###1  Form  A  Occasion  #1  internal consistency
reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + I 5 0
NAME        P "Person"
NAME        I "Items within Test"
FORMAT      (29X,5F5.0///)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY            ###2  Form  B  Occasion #1 internal consistency
reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + I 5 0
NAME        P "Person"
NAME        I "Items within Test"
FORMAT      (/29X,5F5.0//)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY          ###3  Form  A  Occasion  #2  internal consistency
reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
```

```
EFFECT      + I 5 0
NAME        P "Person"
NAME        I "Items within Test"
FORMAT      (//29X,5F5.0/)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY            ###4  Form  B  Occasion #2 internal consistency
reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + I 5 0
NAME        P "Person"
NAME        I "Items within Test"
FORMAT      (///29X,5F5.0)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY      ###5 Form A  test-retest  stability reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + O 2 0
EFFECT      + I 5 0
NAME        P "Person"
NAME        O "Occasion"
NAME        I "Items within Test"
FORMAT      (29X,5F5.0//29X,5F5.0/)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY      ###6 Form B  test-retest  stability reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + O 2 0
EFFECT      + I 5 0
NAME        P "Person"
NAME        O "Occasion"
NAME        I "Items within Test"
FORMAT      (/29X,5F5.0//29X,5F5.0)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY      ###7 Occasion #1  equivalence  reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + T 2 0
EFFECT      + I:T 5 0
NAME        P "Person"
NAME        T "Test"
NAME        I "Items within Test"
FORMAT      (29X,5F5.0/29X,5F5.0//)
REWIND      9
PROCESS     9 DEFAULT
GSTUDY      ###8 Occasion #2  equivalence  reliability
OPTIONS     RECORDS ALL CORRELATION NEGATIVE
EFFECT      * P 6 0
EFFECT      + T 2 0
EFFECT      + I:T 5 0
```

```
NAME          P "Person"
NAME          T "Test"
NAME          I "Items within Test"
FORMAT        (//29X,5F5.0/29X,5F5.0)
REWIND        9
PROCESS       9 DEFAULT
FINISH
```