DOCUMENT RESUME

ED 421 529 TM 028 855

AUTHOR Kieffer, Kevin M.
TITLE Why Generalizability Theory Is Essential and Classical Test Theory Is Often Inadequate.
PUB DATE 1998-04-11
NOTE 48p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (New Orleans, LA, April 9-11, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Error of Measurement; *Generalizability Theory; Heuristics; Interaction; *Test Theory

ABSTRACT

This paper discusses the benefits of using generalizabilty theory in lieu of classical test theory. Generalizability theory subsumes and extends the precepts of classical test theory by estimating the magnitude of multiple sources of measurement error and their interactions simultaneously in a single analysis. Since classical test theory examines only one source of measurement error at a time (e.g. occasions, forms, or internal consistency), it is not possible to estimate the magnitudes of all sources of measurement error and the magnitude of measurement error interaction effects concurrently. As this paper explores the shortcomings of classical test theory and the strengths afforded by using generalizability theory, a small heuristic data set is used to make the discussion concrete. (Contains 9 tables and 28 references.) (SLD)

Running head: GENERALIZABILITY THEORY

Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate

Kevin M. Kieffer

Texas A&M University 77843-4225

Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, April 11, 1998.

Abstract

The present paper discusses the benefits of utilizing generalizability theory in lieu of classical test theory. Generalizability theory subsumes and extends the precepts of classical test theory by estimating the magnitude of *multiple* sources of measurement error and their interactions *simultaneously* in a single analysis. Since classical test theory examines only one source of measurement error at a time (e.g., occasions, forms, or internal consistency), it is not possible to concurrently estimate the magnitudes of all sources of measurement error and the magnitude of measurement error interaction effects. Thus, the present paper explores the shortcomings of classical test theory and the strengths afforded by employing generalizability theory. A small heuristic data set is utilized to make the discussion concrete.

Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate

Inquiry in the social sciences has always been concerned with obtaining accurate and reliable measurements so that substantive studies are based on the analysis of meaningful data (Eason, 1991). Since science is concerned primarily with repeatable and replicable experiments, the influence of random effects that might contaminate results (e.g., measurement error, sampling error) needs to be examined and reduced. Science, therefore, is constrained and influenced by the soundness of measuring instruments and the dependability of the data generated by them. As Nunnally (1982) stated,

> 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. (p. 1589)

The dependability of data generated from measuring instruments, however, is an area that is frequently overlooked in the social sciences, both in practice and in educating student-researchers. Pedhazur and Schemlkin (1991, pp. 2-3) noted that,

> Measurement is the Achilles' heel of sociobehavioral research. Although most programs in sociobehavioral sciences... require a medium of exposure to statistics and research design, few seem to require the same where

> measurement is concerned.... It is, therefore, not surprising that little or no attention is given to the properties of measures used in many research studies.

The generation of accurate and reliable data, therefore, is critical to strong scientific inquiry and essential to furthering scientific knowledge, but is an area that is glossed over by many researchers. As expressed by Willson (1980, pp. 8-9),

> Only 37% of the AERJ studies explicitly reported reliability coefficients for the data analyzed...[and a]nother 18% reported only indirectly through reference to earlier research.... [U]nreported [reliability coefficients] in almost half the published research is... inexcusable at this late date.

What Does Reliability Mean?

Reliability can be defined as, "the degree to which test scores are free from errors of measurement" (American Psychological Association, 1985, p. 19). When measuring phenomena of interest, two types of variance can be generated: systematic or unsystematic. The former is associated with real differences and is likely to be replicated in future measurements. Systematic variance, therefore, is often referred to as "good variance." Unsystematic variance, however, represents variability that is likely to be indigenous only to the measurement or sample being investigated and will probably not replicate in future measurements or will vary in unpredictable ways. Thus, unsystematic variance, or error variance, is considered "bad

variance" and is one of many random effects that psychometricians attempt to reduce or eliminate when collecting observations or measurements.

Properties of Reliable Data

The reliability of scores directly corresponds to the degree to which error is present or absent in a behavioral measurement. As such, reliability is a concept that inures to the scores generated by a particular measuring instrument and not to the instrument itself. Thus, Rowley (1976) noted 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). Similarly, Eason (1991, p. 84) argued:

> Though some practitioners of the classical measurement paradigm speak of reliability as a characteristic of tests, in fact reliability is a characteristic of data, albeit data generated on a given measure administered with a given protocol to given subjects on given occasions.

Thus, the scores of a particular group of individuals are reliable or unreliable, not the measuring instrument used to generate the scores. Unfortunately, based on the common and thoughtless usage of the phrase "the test is reliable", many researchers have erroneously become to believe that tests are reliable. As indicated above, nothing could be more untrue.

Since it is the scores of a group of measured individuals that ultimately impacts reliability, the consistency of the sample used to estimate reliability is critical. Like all statistics,

reliability coefficients are affected by the variability of the scores in the sample of chosen individuals. Reliable data are generated by maximizing the systematic variance (e.g., real differences between people or groups) and minimizing the variance attributed to random sampling or measurement error (e.g., variance that will not replicated in future samples and is specific to only the sample under investigation).

As a group of individuals becomes more heterogeneous, the real or systematic differences between the individuals becomes more pronounced. This divergence in the systematic variance of the group typically leads to more variability in the set of scores. Consequently, the increased score variance for the set of individuals leads to higher reliability for the group. Since more heterogenous groups of individuals tend to generate more variable scores than homogenized groups, the same measure administered to more heterogenous or homogeneous groups will likely yield scores with differing reliabilities. By conceptualizing reliability in this manner, it becomes obvious why it is incorrect and inappropriate to use the statement "the test is reliable," as the same test can produce radically different results when administered to different groups of people.

Indeed, Vacha-Haase (1998) has proposed an important new technique for investigating the variability in score reliability: "reliability generalization." Reliability generalization can be used (a) to characterize the variability in score reliability across studies and (b) to identify those design features that best

explain or predict the variations in score reliabilities across studies.

Classical Test Theory

History and Theoretical Underpinnings

The concept of reliability has been explored since around the turn of the century. The area of inquiry related to estimating the reliability of data, however, has been slow to develop in comparison to techniques utilized to analyze substantitive data. The groundwork for what has recently been called classical test theory was originally articulated by Thorndike (1904). Thorndike argued that reliable information about individuals can be obtained by collecting measurements that have a maximum amount of systematic variance and a minimum amount of measurement error variance. This theoretical conception of reliability was operationalized in a mathematical formula offered by Spearman (1907) called the "true score" model of measurement.

Spearman (1907) extended the notion proffered by Thorndike (1904) to include the influences of both random and measurement in explaining a score that a given respondent will exhibit on a given occasion. As noted by Crocker and Algina (1986, p. 107), "The essence of Spearman's model was that any observed score could be envisioned as the composite of two hypothetical components--a true score and a random error component." Thus, the true score formula can be written in the equation

Observed Score = True Score + Random Error.

In the Spearman (1907) model, a true score refers to the

average score that the same respondent will obtain over infinite administrations of the same version of a given test (Crocker & Algina, 1986). Thus, a true score indicates the actual performance of a particular respondent and is, by definition, a perfectly reliable score. An observed score, however, is the score that a respondent actually generates on a given administration of a test. This observed score may or may not be sufficiently reliable to be useful. Random error is composed of random measurement error that can accentuate or attenuate a given respondent's true score. Based on the positive or negative influences of the random error component, it is possible for a respondent's observed score to be higher or lower than the same respondent's true score.

Consider the following example. Suppose the respondent of interest is a sixth grade math student who is capable of correctly answering 16 out of 20 questions on a math test (the student's true score). Now suppose the student is ill the day of the exam, and fluctuates throughout the exam in cognitive functioning ability, and therefore randomly misses four questions that could have been answered correctly by this student. This situation illustrates the negative influence of random error as the student's true score is higher than the observed score. Suppose that on another occasion with the same exam, the respondent guessed correctly on two items that should have been missed. This condition represents the positive effect of random error, as the student's true score is lower than the observed score. Both positive and negative effects of random error lead to unreliable

scores and potentially misleading estimates of the respondents' abilities.

Reliability Estimates in Classical Test Theory

The following cursory review of classical test theory reliability estimates presumes some knowledge of the subject matter and it not intended to be a thorough coverage of the material. Interested readers are referred to Arnold (1996) and Eason (1991) for more detailed explanations. As noted previously, classical test theory partitions observed score variance into only two components, true score variance and random error variance. Consequently, it is possible to examine only a single source of measurement error at any given time. This poses a serious dilemma for the researcher, since in reality several types of measurement error can exist concurrently. In classical test theory, it possible to consider estimates of measurement error due to inconsistency in forms (equivalence), or observers (inter-rater agreement), or sampling the item domain (internal consistency or split-half) or time (test-retest or stability) (Arnold, 1996). Only one measurement error influence can be considered in a given analysis.

Each type of reliability estimate (e.g., time, observers) can be used to determine the degree to which true scores deviate from observed scores. The problem, however, is that classical test theory is unable to examine inconsistencies in test forms, raters, items, or occasions simultaneously. That is, classical test theory for example makes the potentially erroneous assumption that the

error observed in calculating a stability coefficient is the same error indicated by calculating an internal consistency alpha. As stated by Thompson and Crowley (1994, p. 3),

> An embedded assumption of many researchers using the classical score approach is that sources of error substantially overlap each other (e.g., the 15% of error from a test-retest analysis is the [sic] essentially the same [emphasis in original] error as the 10% or 15% measurement error detected in an internal consistency analysis) and that the sources of error do not interact to create additional error variance.

Much to the dismay of the researcher utilizing classical test theory reliability estimates, error components corresponding to items, occasions and forms are actually distinct sources of variability and often are not equivalent. Further, in addition to only examining one source of error variance at a time, classical test theory also fails to examine the unique contribution of the interaction sources of measurement error variance (e.g., occasions with items, items with forms). As noted by Thompson (1991, p. 1072),

> ...a practitioner may do classical internal consistency, test-retest, and equivalent forms reliability analyses, and may find in all three that measurement error comprises 10% of score variance. Too many classicists would tend to assume that these 10 percents are the same and also tend to not realize that in addition to being

> unique and cumulative, the sources may also interact to define disastrously large interaction sources of measurement error not considered in classical theory. The effects of these assumptions are all the more pernicious because of their unconscious character.

Thus, classical test theory does not consider sources of measurement error simultaneously nor does it examine the effects of their unique interaction effects.

Estimates of reliability in classical test theory are represented in the form of reliability indexes or reliability coefficients (Arnold, 1996). A reliability index is simply the Pearson product moment correlation coefficient between observed scores and true scores (Crocker & Algina, 1986). In order to examine the amount of shared variance between observed scores and true scores, it is necessary to square the Pearson r between them. This statistic, termed a reliability coefficient, indicates the percentage of true score variance accounted for by the observed scores. Coefficient alpha, another type of reliability coefficient, also indicates the amount of shared variance between observed and true scores.

Since several estimates of reliability involve computing correlation coefficients between two sets of scores (i.e., split-half and test re-test), it is necessary to first square the Pearson r prior to reporting it as a reliability coefficient. Regrettably, too few researchers recognize that reliability coefficients are squared statistics and that the Pearson rs

typically used to estimate test retest and split-half reliability are instead reliability indices (Arnold, 1996). Errors such as these only exacerbate the problems incurred in employing classical test theory in analyzing score reliability and, when coupled with the problems previously described, predisposes the wary researcher to search for alternative methods of examining the dependability of behavioral observations.

A Heuristic Example Using Classical Test Theory

In an effort to facilitate greater understanding of classical test theory and thus its limited usefulness in the social sciences, a small heuristic data set will be explored. Suppose the data contained in Table 1 was generated by a hypothetical psychological instrument that measures the degree of happiness exhibited by a respondent. In an attempt to develop a marketable instrument, the psychometrician develops an instrument that is short (five items), easy to administer, and which adequately represents the item domain of interest. The psychometrician uses a seven-point Likert-type scale in constructing the five item test, thus generating item responses ranging from 0 to 6. In an effort to examine some of the psychometric properties of interest, the same version of the instrument is administered to one group of 10 people on four separate occasions. Thus, based on the reliability estimates available in classical test theory and the given information, the psychometrician can examine two types of reliability: stability over time (test-retest) and accurately sampling the item domain (internal consistency). Since the

researcher has elected to utilize classical test theory in ascertaining these reliability estimates, it is not possible to examine these measurement error influences simultaneously, nor it is possible to determine the additional separate and unique measurement error variance introduced by the interaction effect of time with internal consistency. Consequently, the classical theory researcher must calculate each reliability estimate separately.

Insert Table 1 About Here

A quick perusal of the results presented in Table 2 is disconcerting to the hopeful researcher. As a result of examining the reliability of the scores based on their stability and internal consistency, the psychometrician has generated ten different reliability coefficients. The internal consistency alphas on occasion three and four appear promising (.7974 and .8393, respectively) but the alphas on occasion one and two are disastrously low (.4799 and -.2128, respectively). The alpha on occasion two is even negative (see Reinhardt, 1996), indicating that the scores are less reliable than if they were simply generated at random (negative alphas, even less than -1, are possible with coefficient alpha and indicates a serious psychometric problem). Thus, based on examining reliability in terms of inconsistencies in items, the researcher cannot conclude with any certainty that the item domain was adequately represented by the five happiness items included on the instrument and is left

to employ other methods of assessing the quality of the derived scores.

Insert Table 2 About Here

Similarly to the internal consistency alphas, the stability coefficients are quite variable and range from a low of .0705 (occasion two with four) to a high of .3547 (occasion one with four). Obviously, the budding psychometrician has a major dilemma as both types of reliability estimates provide divergent and contradictory results. If the researcher had simply examined the coefficient alpha on occasion four (.8393), a different conclusion may have been reached. After examining all reliability coefficients in respect to each other, the researcher is left befuddled and perplexed as to the true reliability of the scores generated by the instrument. The despondent psychometrician avidly researches developments in reliability theory and discovers that another theory of score reliability, generalizability theory, is superior to classical test theory and might help uncover the true characteristics of the scores generated by the newly devised happiness instrument.

Generalizability Theory

Generalizability theory (G theory), much like classical test theory, is a theory of the dependability of a set of behavioral observations or measurements. G theory, however, considers simultaneously both multiple sources of measurement error variance

and also their unique interaction effects (Eason, 1991). As stated by Thompson and Crowley (1994, p. 2),

> Generalizability theory subsumes and extends classical score theory. "G" theory is able to estimate the magnitude of multiple sources of error simultaneously. Therefore, sources of error variance and interactions among these sources can be considered simultaneously in a single generalizability analysis. This is unlike classical test score analyses which allow for only a single source of error to be considered at one time....and does not consider the possible, completely independent or separate interaction effects of the sources of measurement error variance.

Thus, the power of G theory lies in its ability to examine multiple sources of measurement error variance and their unique interaction effects simultaneously while allowing researchers to accurately assess the dependability of scores in complex measurement designs.

History and Background

Although most major developments in G theory were made several decades ago in the early 1970's, the original groundwork for the theory can be traced even further back to the work of Hoyt (1941), Lindquist (1953) and Medley and Mitzel (1963). Each of these theorists understood the limitations of classical test theory and explored new ways to examine reliability that would allow the simultaneous examination of several sources of measurement error.

The work of Hoyt (1941) is most readily comparable to modern G theory and indicates the first effort to partition score variance into more than one component at a time. Hoyt was influenced by the work of Rulon (1939) who developed a short method of assessing internal consistency reliability through split-half analysis. Hoyt noted that if the odd-even split was "an unlucky one" (Hoyt, 1941, p. 153), the derived split-half coefficient may be an under- or over-estimate of the true dependability of the scores. Thus, Hoyt developed a method which allowed score variance to be partitioned into three components: variance among individuals, variance among items, and random error variance. This was a radical departure from classical test theory which partitions score variance into only true score and random error variance.

The method developed by Hoyt (1941) directly utilized analysis of variance (ANOVA) to estimate the reliability of scores. The main shortcoming of the method, however, is that even though variance is further partitioned into components corresponding to items, individuals and error, the estimate generated by the Hoyt method is exactly equal to the estimate generated by calculation of coefficient alpha, developed later by Cronbach (1951). Further, the Hoyt method can be employed with only dichotomously scored answers (i.e., right versus wrong) which limits its pragmatic value. Thus, after the development of coefficient alpha (which can handle dichotomously or polychotomously scored items), the Hoyt method had limited utility

in the social sciences since it was more difficult to calculate and was appropriate only for certain types of items. The important contribution of Hoyt's method to the development of G theory, however, is that score variance was decomposed into more than two constituent components.

The following example of the Hoyt (1941) method will serve as a basis for understanding the more complex G theory and will further help illustrate the shortcomings of previous theories of score reliability. Consider the following example. An instructor gives a five item spelling examination to a group of 6 students in which each item is classified as right (1) or wrong (0). The data could be placed in a table similar to the one presented in Table 3.

Insert Table 3 About Here

The calculation of reliability through ANOVA involves several steps. The first step is to calculate the sums of squares (SOS) for individuals, items, error and total (Hoyt, 1941, p. 154). The SOS for individuals is

$$\frac{1}{n} \sum (tot)^2 - \frac{(\sum tot)^2}{nk}$$

where n = number of items, k = number of individuals and tot = total score on the test. The SOS for items is

$$\frac{1}{k} \sum (item)^2 - \frac{(\sum tot)^2}{nk}$$

where n = number of items, k = number of individuals, $item$ = the sum of each item for all individuals and tot = total score on the

test. The SOS total for all individuals on all items is

$$\frac{(\Sigma \text{ tot})\ (nk - \Sigma \text{ tot})}{nk}$$

where n = number of items, k = number of individuals and tot = total score on the test. The final component, SOS residual, is computed by subtracting the SOS for items and individuals from the total SOS. Thus, the results could be presented in an ANOVA summary table as illustrated in Table 4.

Insert Table 4 About Here

The next step is to calculate the degrees of freedom (df) for each component. For individuals and items, the df is the number of items or individuals minus one. The df total is calculated by multiplying the number of items and individuals and subtracting one. The df residual is calculated by subtracting the df for items and individuals from the total df (note that SOS and df are always cumulative in ANOVA). The formula for the respective df calculations are presented in Table 4.

The variance for each source of variation in the Hoyt method must then be computed. In fixed effects classical ANOVA, the variance for sources of variation are typically referred to as mean squares (MS). For each of the sources of variation, the MS is simply calculated by dividing the respective SOS component by the corresponding df. The formulas for each source of variation are also presented in Table 4.

The final step in the Hoyt (1941) method is to calculate the

reliability coefficient. Remember that a reliability coefficient is the ratio of true score variance to observed score variance. Thus, Hoyt (1941, p. 155) delineated the reliability coefficient as

$$r_{xx} = \frac{MS_{ind} - MS_{res}}{MS_{ind}}$$

where MS_{ind} = mean square individual, MS_{res} = mean square residual. Remember in ANOVA that the MS is a measure of variance; it is the variance due to a particular source of variation. Thus, the Hoyt reliability coefficient is consistent with the classical test theory notion of reliability, as reliability is defined as the ratio of true score variance (variance due to individuals with the influence of random error removed) divided by the variance due to individuals.

The results of the Hoyt method using the second set of example data are presented in Table 5. Based on the results presented in Table 5, the reliability coefficient for the example data can be calculated as

$$r_{xx} = \frac{.2133 - .1967}{.2133}$$

which generates a reliability coefficient of .0781. Note that this value is a reliability coefficient, as it is the ratio of two variances and is thus in a squared metric. Based on the calculations provided by the Hoyt method, the researcher is able to say that the scores of the six students on the spelling exam

are very unreliable.

Insert Table 5 About Here

The results presented in Table 5 (Hoyt method) can then be compared with a classical test theory coefficient alpha (Cronbach, 1951). Both analyses examine score variability due to the influence of the test items. In the Hoyt (1941) method, the variance associated with the test items was partitioned out of the total score variance. In calculating a coefficient alpha, the ratio of item score variance to test score variance is evaluated. Both statistics generate an estimate of score variability due to the inconsistencies in items. As is illustrated in Table 6, the reliability estimates generated by both statistics are exactly identical. Although the Hoyt method provided the groundwork for contemporaneous G theory, the manner in which it was used only allowed researchers to examine one source of measurement error at time. By examining Table 5 once again, the keen researcher may notice that the SOS associated with the residual (error) term constitutes over half of the total score variance (SOS_{tot}). This variance ambiguously dubbed 'residual' in the Hoyt method can be partitioned further to account for other influences of measurement error. Further partitioning the error term would potentially reduce the $MS_{residual}$ and thus provide a larger reliability coefficient.

Insert Table 6 About Here

Logic and Mechanics of G Theory

Cronbach, Glaser, Nanda, and Rajarantam (1972) extended the work of Hoyt (1941) and others to provide researchers with a modicum of partitioning measurement error into constituent components and interaction effects. Jaeger (1991, p. ix) stated,

> ...Cronbach and his associates... effectively demonstrated that it was no longer necessary to restrict decompensation of variation in individual's observed test scores to two components--variation attributed to true differences among individuals, and variation attributed to a conglomeration of systematic and random sources.... Indeed, this latter component of variation could be dissected further to gain an understanding of the systematic sources of variation that contributed to what we heretofore considered an undifferentiable mass, simply "error."

Thus, G theory recognizes that multiple sources of error are present in behavioral observations and provides a manner in which researchers can further partition error variance into smaller components that account for systematic and unsystematic sources of variation.

In G theory, unlike classical test theory, a behavioral observation is considered only one sample of behavior from an infinite universe of admissible observations. The universe, therefore, is comprised of all the potential measurements that

would be a direct substitute for the observation under investigation. The particular sample or observed score is not the focus of the generalizability analysis; rather, the analysis focuses on how well a particular sample of behavior generalizes to the larger universe of admissible observations. As stated by Cronbach et al. (1972, p. 15),

> The score... is only one of many scores that might serve the same purpose. The [researcher] is almost never interested in the response given to particular stimulus objects or questions, to the particular tester, at the particular moment of the testing. Some... of these conditions of measurement could be altered without making the score any less acceptable....

The focus of the dependability analysis is subsequently altered and, rather than being interested in how well observed scores represent respondent true scores, researchers are interested in assessing how well observed scores generalize to a "defined universe of situations" (Shavelson, Webb & Rowley, 1989, p. 922). Cronbach et al., (1972, p. 15, emphasis in original) summarized this notion by stating,

> The ideal datum... would be something like the person's mean score over all acceptable observations, which we shall call his [sic] "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he [sic] generalizes from sample to universe. *The question of "reliability" thus resolves into a*

question of accuracy of generalization, or generalizability. [Emphasis in original]

Measurement Objects and Facets. Generalizability studies first involve defining the score population of interest, called the universe of admissible observations. The universe of admissible observations involves both the "object of measurement" and the "facets of measurement." The object of measurement in a G study is a source of variability that the researcher deems real and legitimate, and which arises from true, systematic differences. It is also the score variance about which the researcher wishes to generalize to the larger universe. In the social sciences individuals are usually employed as the objects of measurement since researchers typically believe that people are different and that differences across people are real and systematic (Eason, 1991). As noted by Thompson (1991), however, people do not have to be the objects of measurement. Thompson (1991, p. 1073) offers an excellent example of a situation in which people are not the focus of the analysis:

> ...person need not be the object of measurement in a [G] study. Often people constitute the object of measurement, because our model of reality says to us that people really are different, so that observed differences in people represent true variance.... One example... presumes the use of mice, so popular with our colleagues in biology and medicine, that have been inbreed over many generations so that they are genetically almost identical. Our model of

> reality says that any observed differences in these mice as individuals represents measurement error and not systematic variance. We could then, for example, declare occasion as our object of measurement....

It is critical to understand that although researchers are typically interested in the individual performance differences of people, there are other sensible alternatives for the object of measurement (namely anything that is believed to generate systematic variance such as schools, businesses, or occasions).

The universe is further composed of sources of error variation or facets (Shavelson & Webb, 1991). Each facet in turn is composed of a defined number of conditions, or levels of the facet. Facets can include any source of measurement error variance, but forms, items, occasions, and raters are typically the most popular facets utilized in the social sciences. If a researcher was interested in examining the internal consistency of test items using G theory, for example, the universe of admissible observations would include all possible individual test items. If the test items were defined as a facet, or source of measurement error variation, and the number of conditions or levels in the item facet would be constrained only by the number of test items that the researcher desired to include in the analysis. Thus, the number of facets included in a G study are only those sources of measurement error variance of interest to the researcher, but could potentially number as many or as few as desired.

By definition the object of measurement cannot be a facet.

Facets contain measurement error variance and objects of measurement contain real or systematic variance. By employing G theory in an analysis, a researcher is attempting to partition the score variance associated with a facet (error) from the score variance accounted for by the object of measurement. The object of measurement, therefore, is analogous to the concept of true score variance in classical test theory as both represents the variance that the researcher considers real and about which the researcher wishes to make statements.

In a G study researchers are primarily interested in examining and thus controlling for any source of variation that might affect score variance in an unpredictable or unreplicable manner. The focus of the analysis is an individual's universe score, or the individual's score over all possible items, forms, occasions, or raters. Obviously, it is impossible to calculate the universe score, so it must be estimated in the G study. As stated by Webb, Rowley and Shavelson (1988, p. 83), the researcher

> would like to know an individual's average score over all combinations of conditions (all possible forms, all possible occasions, all possible items).... Unfortunately, this ideal datum cannot be known; it can only be estimated from a sample of items from a sample of forms on a sample of occasions. This is where error, because of the particular form, occasion, and items chosen, occurs.

Absolute Versus Relative Decisions. Another important distinction between G theory and classical test theory is that G

theory differentiates between relative and absolute decisions. Relative decisions are those decisions which involve the relative standing of individuals in regard to one another and which are based solely on considering the stability of a person's rank within a group. This is the only type of decision rendered in classical test theory and is influenced only by the position of the objects of measurement in respect to one another. A relative decision involves making statements such as "We are giving exactly one scholarship each year to the person scoring highest each year on a scholarship examination, and we do not care that across years the highest score may vary quite a lot." In rendering relative decisions, the focus of the decision is only on the ranking of individuals within a group and no attention is given to scores in relation to a cutoff score or percentage of items correct. A Generalizability coefficient is the reliability coefficient used to evaluate scores in a relative decision context.

Absolute decisions, in contrast, involve the consistency of placement in relation to a standard or cut-off score. An example of an absolute decision involves admission into a university. Students entering most universities are required to attain a minimum score on an entrance examination (usually the Scholastic Aptitude Test) to gain admission. A given respondent will be offered admission into the chosen university only if the observed score surpasses the minimum cutoff score established by that university regardless of how well other students performed on the same examination. Thus, in rendering absolute decisions, the focus

of the decision is on the specified standard or cutoff score and the ranking of individuals within a group.

In generalizability theory, absolute decisions are examined by computing a phi coefficient. A phi coefficient evaluates the stability of an observed score in relation to a fixed cutoff or standard. The phi coefficient is calculated by comparing the systematic variance (i.e., the variance associated with the object of measurement) to all relevant main effect error components and their interaction effects. Similarly to g coefficients, phi coefficients are in a squared metric and range from 0 to 1.00. A noteworthy distinction between G theory and classical test theory involves absolute decisions and the phi coefficient: Classical test theory does not evaluate absolute decisions.

Fixed Versus Random Facets. An important concept in G theory involves distinctions between fixed and random facets. [The object of measurement is always considered a random effect.] When deciding whether a given facet is random or fixed, it is necessary to be familiar with the notion of exchangability (Webb, Rowley & Shavelson, 1988). A facet is considered random if the researcher is willing to exchange the conditions in the sample with any other conditions in the universe of interest. Suppose for example, that an instructor is conducting a G study and is interested in examining the items on a statistics exam. The instructor has generated a sizable item pool through 20 years of teaching and extracts 40 items for the exam from an item pool that totals 300. The items facet in this type of G study would be considered random

if the instructor is readily willing to exchange the 40 chosen items with any other set of 40 items in the item pool. A facet is considered random, therefore, if and only if the chosen sample of observations can be exchanged with another equally sized sample of observations without a loss of information.

A fixed facet, however, exhausts the conditions available in the universe of interest, or at least includes all the conditions of interest to the researcher. In the case of a fixed facet, the researcher does not desire to generalize beyond the conditions included in the G study. Using the example presented in the previous paragraph, the items facet would be considered fixed if the researcher was not willing to exchange another sample of 40 test items with any other set from the item pool. Webb, Rowley and Shavelson (1988) have recommended that if a question arises as to whether a facet is fixed or random, the researcher should first conduct a G study with the questionable facet considered random. After examining the variance component for the questionable facet, the researcher should consider the facet fixed only if the variance component is very large (thus indicating that performance is differential across conditions of the facet).

Crossed Versus Nested Designs. Another important concept in G theory is the notion of crossed and nested designs. A design is considered to be crossed when all of the conditions in one facet are observed with all conditions of another source of variation. For example, a crossed design would occur in a persons-by-occasions-by-items design when all persons in the G study respond

to the all of the same items on all of the same occasions. Thus, all of the conditions in each facet are combined with every condition of another facet.

Conversely, a design is considered nested when two or more conditions of the nested facet appear within one and only one condition of another facet. For example, a generalizability study might be done with 20 people as the object of measurement. Each respondent completes 10 English and 10 History short-answer questions on each of two occasions. This is the items-nested-within-subjects-by-time design. That is, subjects and occasions are "crossed," since both subjects are evaluated on both occasions. However, the two sets of 10 items are respectively "nested" within the two subject areas (i.e., the 10 English items are never administered as part of the History test, nor are the 10 History items ever administered as part of the English test). G theory can be employed with both nested and crossed facets or any combination of crossed and nested facets, however, there are statistical advantages to using crossed designs whenever possible.

Partitioning Variance in G Theory. A critical aspect of G theory is the manner in which score variance is partitioned. G theory invokes an ANOVA framework in partitioning the score variance into its constituent components. An observed measurement is decomposed into one component representing the variance due to the object of measurement (universe score) and one or more components representing the variance due to measurement and random error. Thus, G theory utilizes the exact logic employed in ANOVA

to partition variance. Just as a multi-way ANOVA examines the influence of one or more independent variables and their unique and cumulative interaction effects on a dependent variable, G theory can examine the influence of one or more sources of measurement error and their unique and cumulative interaction effects.

The partitioning of score variance in G theory can further be illustrated using the logic invoked in simple and factorial ANOVA. In a simple one-way ANOVA, variance on the dependent variable is partitioned into two components, between and within. The between component represents the portion of variance on the dependent variable that the researcher believes is systematic while the within component represents the portion of variance on the dependant variable that is due to error and will most likely not replicate in future analyses. This concept of partitioning variance into only two components is analogous to the decomposition of score variance in classical test theory, as variance is partitioned into true score (systematic) variance and error variance.

Factorial ANOVA, however, allows researchers the opportunity to further partition the variance on the dependent variable. In factorial ANOVA, researchers recognize that multiple factors and the unique and cumulative effects of their interactions influence the variance evidenced on the dependent variable. Factorial ANOVA decomposes the error component of simple ANOVA into variance that can be attributed to other factors and the interaction effects of

these factors with one other. G theory operates in much the same manner by partitioning the error variance manifested in classical test theory into constituent components associated with sources of measurement error and their interaction effects.

The variance partitions in G theory are unique, separate and perfectly uncorrelated, just as they are in classical ANOVA. When partitioning score variance in G theory, the researcher partitions out the variance associated with the object of measurement. The remaining error component can then be subdivided into components corresponding to the influence of items, occasions and forms. For example, suppose a psychometrician is interested in examining the influence of test items (10), forms (3), and occasions (3) on the performance of a group of individuals. The score variance could be first partitioned into two components, one corresponding to the individuals (p) and one which is simply labeled error. The second partition would subdivide the error component into the pooled main effect sources of error variance and the pooled interaction sources of error variance. The third partition would further subdivide the pooled main effect component into the score variance associated with the three main effects of items (i), forms (f) and occasions (o). The pooled interaction variance would be subdivided into two-way, three-way, and four-way effects, and then these would be further partitioned into their specific elements: (a) pi, pf, po, if, io, fo, for two-way partitions, (b) pif, pio, pfo, ifo for three-way partitions, and finally (c) the quadruple interaction of persons with items, forms and occasions compounded

with random error (pifo,e). Notice that in G theory, unlike factorial ANOVA, the highest order interaction effect is confounded with random error.

Generalizability Studies

As noted earlier, a G study is intended to estimate the variance associated with the sources of measurement error that are included in the universe of admissible observations defined by a given researcher. Once the universe is defined and the sources of error are identified, it is possible to estimate the relative contribution of each source of error by generating variance components. Variance components are the building blocks of both ANOVA and G theory, but due to an increased emphasis on using ANOVA for statistical significance testing (Thompson, 1989), variance components in ANOVA were abandoned for mean squares (Eason, 1991; Guilford, 1950). Since statistical significance testing is not the focus of a G study, F tests have little utility in this approach; thus, mean squares as used to estimate the variance components rather than to test statistical significance (Brennan, 1983).

There is a variance component for each main effect and each interaction source of measurement error variance. Using the example described in the classical test theory section of the present paper, the derivation of the variance components can be illustrated. Recall that the psychometrician in the previous example was interested in developing an instrument that assessed a respondent's degree of happiness. This researcher intended to

develop a short questionnaire of five items so that the administration and scoring of the instrument would be relatively painless. Recall also that after developing the instrument, the psychometrician administered the same form of five items to a group of 10 individuals on four separate occasions. This data was presented in Table 1.

Using G theory, it is possible to examine the multiple sources of error variance simultaneously. The researcher has defined the object of measurement as the group of individuals (10), and the universe of admissible observations as the items on the test (5) and the occasions of administration (4). Thus, this design can be labeled a fully crossed pxixo (10x5x4) random effects generalizability analysis. The facets are considered random in this example because the researcher believes that the occasions and items could be exchanged with another sample of occasions and items in the universe without a substantial loss of information.

The formulas for the relevant calculations will not be described in the present paper, but interested readers are referred to Cronbach et al. (1972), Brennan (1983) and Thompson and Crowley (1994) for more complete descriptions. The method of calculating variance components in the present paper is extracted from Thompson and Crowley (1994). Since the score variance in the example has been partitioned into three main effect (p, o, i) and four interaction components (po, pi, oi, poi,e), it is possible to compute the SOS for each source of variation.

The SOS values for each source of variation are then transformed into mean squares (MS) by dividing the SOS by the relevant degrees of freedom (calculated as k-1 where k = number of items, occasions or persons for main effects and as the product the of involved main effects for interaction effects). The mean squares for each source of variation is then transformed into a variance component for scores by invoking a series of additive and divisive properties that remove the influence of the interaction mean squares from the main effects and which removes the error component from the interaction mean squares. The results of these calculations for the example data are presented in Table 7.

Insert Table 7 About Here

An important point to note prior to discussing the remainder of the calculations is that computed variance components in G studies can be negative even though such an occurrence is impossible conceptually since variance is a squared statistic and should always be positive. Due to the manner in which the variance components are calculated, however, it is possible for variance components in G studies to be less than zero. As illustrated in Tables 7 and 8, several of the score variance components are in fact negative.

A negative variance component typically indicates a problem in either the size of the sample utilized in the G study or in a misspecification of the model employed. Shavelson, Webb, and Rowley (1989, p. 927) proclaimed that, "estimating variance

components [is] the Achilles heel of GT and sampling theories in general. Especially with small sample sizes, ANOVA estimates of variance components are unstable and may even be negative." The magnitude of the negative variance component is critical as very small negative variance components (which typically indicate sample size problem) are less disconcerting than very large negative variance components (which typically indicate a model misspecification error). Since the three negative variance components in the present example are small in magnitude, the researcher can attribute the cause to an unusually small sample size which was chosen here only for heuristic value.

There are two general ways to handle negative variance components: One is to immediately convert the negative variance component to zero and to use the value of zero in all subsequent calculations (Cronbach et al., 1972); the other option is to convert the variance component to zero but to use the negative value in all subsequent computations (Brennan, 1983). There are, of course, pros and cons to using each strategy. Research has indicated that using zero in place of a negative variance component will produce a biased estimate of the g or phi coefficient (Webb, Rowley, & Shavelson, 1988). Using the negative variance component does produce an unbiased estimate, but as stated by Webb, Rowley and Shavelson (1988), "in neither case is it statistically comfortable to change estimates to zero" (p. 927). Researchers must use their discretion when converting negative variance components to zero and remain cognizant that

doing so may bias G study results.

After computing the variance components for scores, the variance components for means must be calculated by multiplying each variance component by the frequency with which it occurs in the analysis. The results of these transformations are presented in Table 8. Notice that the variance component for items is multiplied by five since there are five items on the instrument. Similarly, the score variance component for occasions is multiplied by four since there were four occasions. The object of measurement variance component is slightly different, however, as each person is believed to truly vary and thus the frequency for the persons variance component is set to 1. The frequency for the relevant interaction variance components are derived by simply multiplying the frequency of the associated sources of error (just as would be done to calculate the degrees of freedom in a factorial ANOVA).

Insert Table 8 About Here

After transforming the score variance components into mean variance components it is possible to examine the percentage of score variance that is accounted for by the objects of measurement, facets and interaction effects. These calculations are presented in Table 9. For a given study to demonstrate generalizable scores, the object of measurement variance component must be very large and the measurement error variance components must be very small.

Insert Table 9 About Here

Notice that the G coefficient and the phi coefficient utilize different variance components in their computations. The G coefficient includes only those sources of measurement error that involve the object of measurement (p, po, pi, and poi,e). Conversely, the phi coefficient includes all relevant main effect and interaction sources of measurement error variance. It is important to note, therefore, that the phi coefficient will always be exactly equal to or smaller than the G coefficient since more sources of measurement error are included in the denominator of its computation. Unfortunately in the present example both coefficients are zero and it is impossible to illustrate magnitude differences in the two coefficients.

After examining the previous three summary tables, the ardent and hopeful researcher is frustrated and disconcerted by the attained results. The negative variance components for persons, occasions and items possibly indicates that the sample size was too small to generate useful information. The researcher also noticed that the variance components for the persons-by-occasions interaction (.1784) was fairly large (62.66%), indicating that the participants performed differentially across occasions in which the instrument was administered. Unfortunately for this researcher, the G study has uncovered some serious flaws in the measurement approach. If the researcher had failed to employ G

theory, however, the researcher might have been tempted to report the acceptable coefficient alphas generated on occasions three and four (see Table 2). The researcher does have one avenue of recourse, however, as it is now possible to employ a D study to explore how changes in the measurement protocol will affect the generalizability coefficients.

Decision Studies

A D study involves "what if" type analyses to discern the most efficient or cost effective measurement protocol. D studies are another powerful feature of G theory, as researchers are able to discern whether or not it will be more cost effective (in terms of the consequential effect on the G or phi coefficient) to administer a different number of items or forms on a different number of occasions. As stated by Eason (1991, p. 94), "D studies use variance components information from the G study to facilitate the design of a measurement protocol that both minimizes error variance and is most efficient, i.e., that yields the most reliable scores with the least effort."

It is important to note in D studies that only the facets included in the G study may be used in the D study. In the previous example where items and occasions were facets, it would not be possible to examine the influence of a forms facet unless that facet was included in the original G study. Similarly, it is also important to note that all of the facets included in the G study do not have to be used in the D study.

In a D study a researcher alters the influence of a facet by

changing the number of conditions in the facet of interest. For instance, in the previous example five items were utilized in the study. A D study analysis allows the researcher to examine the effects on the G and phi coefficients if the items are increased to 100 or decreased to two. Similarly, a D study analysis might reveal that reducing the number of occasions from four to two will increase the coefficients of interest to the desired level. Thus, it is possible for a given researcher to state a desired level of generalizability (e.g., .75) and then alter the number of conditions in the facet until the desired level is attained.

Since the results of the example G study were less than desirable, the present author only performed one alteration of the measurement protocol to simply illustrate to the readers the power of D studies and the manner in which they are completed. By increasing the number of test items from five to 10 and increasing the number of occasions from four to seven, a G coefficient of .4041 and a phi coefficient of .4012 were attained. These values are still less than desirable, but the power of this approach is evident: it is possible to examine the costs and benefits of altering the number of items and occasions until the desired level of generalizability is achieved. Interested readers are referred to Shavelson and Webb (1991) and Cronbach et al. (1972) for more information on D studies.

Summary and Conclusions

The present paper has illustrated that G theory, which simultaneously considers multiple sources of measurement error

variance and their unique and cumulative interaction effects, is superior to classical test theory. G theory more closely honors the model of reality in which researchers desire to generalize their results. Since reality is complex and composed of multiple sources of error, the research and analytic methodology that is employed in conducting analyses must also be multifaceted and able to examine multiple sources of measurement error and their interaction effects. The power and far reaching applicability of G theory can be illustrated by Jaeger's (1991. p. x) statement, "Thousands of social science researchers will no longer be forced to rely on outmoded reliability estimation procedures when investigating the consistency of their measurements."

References

American Psychological Association. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Arnold, M.E. (1996). Influences on and limitations of classical test theory reliability estimates. Research in the Schools, 3(2), 61-74.

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Fort Worth: Harcourt, Brace, Johanovich.

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

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

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in educational research: Substantive findings, methodological developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

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.

Jaeger, R.M. (1991). Forward. In R.J. Shavelson & N.M. Webb,

Generalizability theory: A primer (pp. ix-x). Newbury Park: Sage.

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. (1982). Reliability of measurement. In H.E. Mitzel (Ed.), Encyclopedia of educational research (pp. 1589-1601). New York: Free Press.

Pedhauzur, E.J., & Schmelkin, L.P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Lawrence Erlbaum.

Reinhardt, B. (1996). Factors affecting coefficient alpha: A mini Monte Carlo study. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp. 3-20). Greenwich, CT: JAI Press.

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

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. Harvard Educational Review, 9, 99-103.

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

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

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

161-169.

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

Thompson, B. (1991). Review of Generalizability theory: A primer by R.J. Shavelson & N.W. Webb. Educational and Psychological Measurement, 51, 1069-1075.

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

Thompson, B., & Crowley, S. (1994). 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)

Thorndike, E.L. (1904). An introduction to the theory of mental and social measurements. New York: John Wiley.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

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.

Willson, V.L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.

Table 1

Heuristic Data Set

	Occasion 1					Occasion 2					Occasion 3					Occasion 4				
	Item					Item					Item					Item				
Person	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	4	4	3	5	5	4	2	4	3	3	4	3	5	2	2	2	5	3	4
2	5	6	6	4	6	2	6	4	4	4	5	3	4	3	2	1	2	3	1	1
3	6	2	4	2	5	0	6	2	4	4	2	2	3	1	4	3	2	1	2	2
4	6	2	4	3	4	4	2	0	2	2	2	2	3	3	2	4	4	2	2	2
5	1	1	6	2	1	1	6	2	6	6	2	3	1	2	2	4	4	3	4	4
6	2	1	2	3	4	3	4	2	2	4	5	5	6	6	6	6	3	3	2	6
7	6	5	6	4	4	2	6	4	6	4	2	5	3	4	4	3	2	2	3	2
8	4	0	6	4	2	0	6	4	3	1	6	6	3	4	2	4	4	5	2	5
9	0	2	5	5	4	2	0	5	4	4	4	2	6	3	2	6	6	5	4	6
10	3	3	2	5	0	3	0	4	0	6	1	1	2	1	3	2	2	3	1	4

Table 2

Summary of Coefficient Alphas and Stability Coefficients for Table 1 Data

		Occasion			
Occasion	Alpha	One	Two	Three	Four
One	.4799	1.0000			
Two	-.2128	.3555 (.1265)	1.0000		
Three	.7974	-.3445 (.1187)	.3216 (.1034)	1.0000	
Four	.8393	-.5956 (.3547)	-.2656 (.0705)	.4341 (.1884)	1.0000

Note. Test-retest reliability coefficients for administrations one through four are included in parentheses below the reliability indexes.

Table 3

Example Data for Hoyt Method of Reliability Through ANOVA

	Item					
Student	1	2	3	4	5	Total
1	1	1	1	1	0	4
2	1	0	1	1	1	4
3	1	1	1	1	1	5
4	1	0	1	0	0	2
5	0	1	1	0	1	3
6	1	1	1	1	0	4
Total	5	4	6	4	3	22

Table 4

Example Summary Table for Hoyt Reliability Through ANOVA Method

Source	SOS	df	MS (Variance)
Individuals	SOS_{ind}	k-1	SOS_{ind}/k-1
Items	SOS_{itm}	n-1	SOS_{itm}/n-1
Residual	SOS_{res}	(nk-1)-(n+k-2)	SOS_{res}/(nk-1)-(n+k-2)
Total	SOS_{tot}	nk-1	SOS_{tot} / nk-1

Note. n = number of items, k = number of individuals.

Table 5

ANOVA Summary Table for Table 3 Data

Source	SOS	df	MS (Variance)
Individuals	1.0667	5	.2134
Items	.8667	4	.2167
Residual	3.9333	20	.1967
Total	5.8667	29	.2023

Table 6

Results of Table 3 Data Using Coefficient Alpha

	Item 1	Item 2	Item 4	Item 5
Item 1	1.0000			
Item 2	-.3162	1.0000		
Item 4	.6325	.2500	1.0000	
Item 5	-.4472	.0000	.0000	1.0000

Alpha = .0781

Note. The variance of item 3 is equal to zero so it is not possible to compute its correlation with other variables.

Table 7

Summary Table for G Study with Table 1 Data

Source	SOS	df	MS	-	MS	-	MS	+	MS	=	Sum
p	48.7050	9	5.4117	-	5.4416	-	1.9415	+	1.8740	=-	.0974
o	3.3050	3	1.1017	-	5.4416	-	3.7225	+	1.8740	=-	6.1884
i	5.4800	4	1.3700	-	1.9415	-	3.7225	+	1.8740	=-	2.4200
po	146.9228	27	5.4416	-	1.8740					=	3.5676
pi	69.8907	36	1.9415	-	1.8740					=	.0675
oi	44.6700	12	3.7225	-	1.8740					=	1.8485
poi,e	202.3973	108	1.8740							=	1.8740

Table 8

Conversion of Score Variance Components to Mean Variance Components

Source	Sum	/	Product	=	Score Variance Component	/	Frequency	=	Mean Variance Component
p	- .0974	/	(o=4) (i=5)	=	-.0049	/	1	=	-.0049[a]
o	-6.1884	/	(p=10)(i=5)	=	-.1238	/	4	=	-.0309[a]
i	-2.4200	/	(p=10)(o=5)	=	-.0605	/	5	=	-.0605[a]
po	3.5676	/	(i=5)	=	.7135	/	4	=	.1784
pi	.0675	/	(o=4)	=	.0169	/	5	=	.0034
oi	1.8485	/	(p=10)	=	.1849	/	20	=	.0092
poi,e	1.8740			=	1.8740	/	20	=	.0937

Note. The mean variance components denoted with an 'a' indicate variance components that will be set to '0' for all future calculations as recommended by Cronbach et al., (1972).

Table 9

Generation of G and Phi Coefficients

Source	Mean Variance Component		G Coeff	Phi Coeff	% of Total
p	.0000		.0000	.0000	0.00%
o	.0000			.0000	0.00%
i	.0000			.0000	0.00%
po	.1784		.1784	.1784	62.66%
pi	.0034		.0034	.0034	1.19%
oi	.0092			.0092	3.23%
poi,e	.0937		.0937	.0937	32.91%
Sum	.2847	Value	.0000[a]	.0000[b]	100.00%

Note. a = the G coefficient is calculated by dividing the object of measurement variance component by the object of measurement variance component added to all the variance components that include the object of measurement (i.e., po, pi, poi,e). B = the Phi coefficient is computed by dividing the object of measurement variance component by the object of measurement variance component added to all the other variance components (i.e., o, i, po, pi, oi, poi,e).

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

TM028855

I. DOCUMENT IDENTIFICATION:

Title: Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate	
Author(s): Kevin M. Kieffer	
Corporate Source:	Publication Date: 4/11/98

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

Sample sticker to be affixed to document

Check here

Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

KEVIN M. KIEFFER

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: X	Position: RES ASSOCIATE
Printed Name: KEVIN M. KIEFFER	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409)845-1831
	Date: 4/25/98