
Generalizability Theory

Richard J. Shavelson
Noreen M. Webb
Glenn L. Rowley

University of California, Santa Barbara
University of California, Los Angeles
Monash University

ABSTRACT: *Generalizability Theory (GT) provides a flexible, practical framework for examining the dependability of behavioral measurements. GT extends classical theory by (a) estimating the magnitude of multiple sources of measurement error, (b) modeling the use of a measurement for both norm-referenced and domain-referenced decisions, (c) providing reliability (generalizability) coefficients tailored to the proposed uses of the measurement, and (d) isolating major sources of error so that a cost-efficient measurement design can be built. Unfortunately, GT has not been readily accessible to psychological researchers, perhaps because its development and presentation have been largely technical. GT's inaccessibility may explain why classical theory remains the preferred method for estimating reliability. The purpose of this article is to present GT and its wide applicability to a broad audience. Our intent is to demystify GT and provide a useful tool to psychological researchers and test developers.*

Generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) provides a framework for examining the dependability of behavioral measurements. Generalizability theory (GT) extends classical reliability theory most notably by recognizing and estimating the magnitude of the multiple sources of measurement error. In this article, we describe GT for an audience of psychologists, not psychometricians. To this end, we develop the theory's central concepts concretely and simply, linking the new theory to the classical theory (CT). We then describe some recent advances and conclude with a concrete application.

Motivation for Generalizability Theory

CT partitions observed-score variance into two parts: that which is thought to be systematic (called true-score variance) and that which is thought to be random (called error variance). The results of this partition are elegant and satisfying.

CT's usefulness, however, depends on the researcher's ability to estimate true-score and error variances from data. With practical application of CT, we find that error variance is not a monolithic construct; error arises from multiple sources. Consequently, error estimates and estimates of reliability (the ratio of true- to observed-score variance) vary according to the data-collection design.

Test-retest reliability counts day-to-day variation in performance as error, but not variation due to item sampling. An internal-consistency coefficient counts variation due to item sampling as error, but not day-to-day variation. Alternate-forms reliability counts both sources as error. CT, then, sits precariously on shifting definitions of true- and error-scores.

GT, in contrast, recognizes that there may be multiple definitions of true- and error-scores. Multiple sources of error define the universe of generalization, a fundamental notion in GT. A measurement taken on a person records a sample of behavior. A score

is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. (Cronbach et al., 1972, p. 15)

A score's usefulness, then, largely depends on the extent to which it allows us to generalize accurately to behavior in some wider set of situations, a *universe of generalization*.

That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalizes from sample to universe. (Cronbach et al., 1972, p. 15)

The concept of *reliability*, so fundamental to CT, is replaced by the broader and more flexible notion of *generalizability*. Instead of asking how accurately observed scores reflect their corresponding true scores, GT asks how accurately observed scores permit us to generalize about persons' behavior in a defined universe of situations. "The question of 'reliability' thus resolves into a question of accuracy of generalization, or generalizability" (Cronbach et al., 1972, p. 15).

A Framework for Estimating Error and Generalizability

GT can be described and illustrated within the analysis of variance (ANOVA) framework. GT is to measurement

what the ANOVA is to substantive research. Just as the researcher attempts to identify and estimate the effects of potentially important independent variables, G theory attempts to identify and estimate the magnitude of the potentially important sources of error in a measurement. Just as the substantive researcher may manipulate some independent variables, control others, and ignore still others, so can the researcher concerned about measurement quality systematically manipulate certain facets of error, control others, and ignore still others.

The connection between GT and CT is paralleled by the connection between factorial and simple ANOVA. A researcher, using simple ANOVA, partitions variance into two parts, usually labeled "between" and "within." The former is thought of as systematic variance and is associated with the design factor that distinguishes groups from one another. The latter is thought of as random and treated as error because it is variance that obscures the researcher's focus on group differences. In the same way, CT partitions variance into true-score and error variance. The former is thought of as systematic variance, associated with differences between the objects of measurement (persons), whereas the latter is treated as random variance, unrelated to true-score variation.

By applying factorial instead of simple ANOVA, the researcher acknowledges multiple factors contributing to variance in data by partitioning data into segments corresponding to each factor, to the interactions among them, and to "random error." Simple ANOVA asks more limited questions, and addresses them with less efficiency, than does factorial ANOVA. Similarly, the extension of CT to GT acknowledges multiple influences on measurement variance. Whereas CT partitions variance into only two sources, GT partitions variance into many sources, corresponding to systematic variance among the objects of measurement, to multiple error sources, and to their interactions.

GT's multifaceted approach to error can be illustrated by a study of the dependability of measures of disaster survivors' psychological improvement (Gleser, Green, & Winget, 1978). Twenty adult survivors (S) were interviewed independently by two interviewers (I). Each interview was rated independently by two raters (R) as to the extent of each survivor's psychiatric impairment on a number of subscales (e.g., anxiety) of the Psychiatric Evaluation Form (Gleser et al., 1978). In differentiating survivors with respect to the extent of impairment, error in the measurement may arise from inconsistencies associated with interviewers, with raters, and with other unidentified sources. GT incorporates these potential sources of error into a measurement model and estimates the components of variance associated with each source of variation in the $20 \times 2 \times 2$ ($S \times I \times R$) design.

Shavelson presented a version of this article as a Division 5 invited address at the annual meeting of the American Psychological Association, August 1987.

Correspondence concerning this article should be addressed to Richard J. Shavelson, Graduate School of Education, University of California, Santa Barbara, CA 93106.

Table 1 presents the sources of variation and the estimated variance components for the anxiety subscale. Three components are large relative to the others. The first, for survivors, is analogous to true-score variance in CT and indicates that substantial differences among the survivors were detected, regardless of who interviewed them or who rated their interviews. The second, the survivor by interviewer interaction, represents one source of measurement error and is due to inconsistencies between the two interviewers in obtaining information from different survivors (e.g., the two interviewers may have provided protocols that led raters to rank order survivors differently). The third is the residual term representing the $S \times I \times R$ interaction and unidentified sources of measurement error. The lower variance components for R and I indicate that there were less systematic differences between raters in the overall level of the ratings that they gave and between interviewers in the data that they gathered. R did not interact substantially with S or with I, indicating that the differences among survivors and between interviewers were reported consistently by the two raters.

Conceptual Underpinnings of Generalizability Theory

GT extends CT in several important ways. First, it recognizes multiple sources of measurement error, estimates each source separately, and provides a mechanism for optimizing the reliability. Second, although GT provides a reliability coefficient, called a "generalizability coefficient," the theory self-consciously focuses on variance components that index the magnitude of each source of error. Third, GT distinguishes between *relative decisions*, where interest attaches to the dependability of the differences among individuals, and *absolute decisions*, where scores are themselves interpretable without reference to others' performance. Fourth, GT distinguishes between *generalizability* (G) and *decision* (D) studies. G studies estimate the magnitude of as many potential sources of measurement error as possible. D studies use information from a G study to design a measurement that minimizes error for a *particular* purpose.

Table 1
Generalizability of Measures of Disaster Survivors' Psychiatric Impairment

Source of variation	Estimated variance components
Survivors (S)	1.84
Raters (R)	0.21
Interviewers (I)	0.49
$S \times R$	0.27
$S \times I$	1.82
$R \times I$	0.03
Residual ($S \times R \times I$, error)	1.58

Multiple Sources of Measurement Error

It was recognized early that the measurement error estimated in two applications of a test (retest method) was not necessarily the same as the error estimated with alternative forms of the same test (e.g., Goodenough, 1936; Thorndike, 1947). CT, however, cannot estimate the separate sources of error simultaneously. Hence, researchers who measure the achievement of five students on a four-item test on two test occasions have to decide whether to estimate the retest reliability of total scores or the internal consistency of the test on each of the two test occasions. Regardless, they cannot simultaneously estimate the impact of *test occasion* (o), *item homogeneity* (i), and their combination (o $\times$ i) on the reliability of the measurement. Without being able to estimate each potential source of error, how are they to decide, in order to achieve a desired level of reliability, whether to increase the number of occasions, the number of items, or some combination of the two?

GT addresses this problem by recognizing that behavioral measurements may contain multiple sources of error. The sources of error (e.g., occasions, raters, and items) are called facets, and the levels of the facets are called conditions. Hence, GT replaces CT's simple representation of a person's (p) observed score on a particular test (i), X_{pi} , as the sum of his or her true score (T_p) plus an undifferentiated error (E_{pi}) term— $X_{pi} = T_p + E_{pi}$ —with a representation that includes multiple sources of error, such as occasions (o) and items (i) in our achievement example (Table 2).

Notice that in GT there is a variance associated with each component of an observed score—except for the constant μ . For example, the person effect, analogous to the true score in CT, has a variance over people: $E_p(\mu_p - \mu)^2 = \sigma_p^2$. GT, then, enables us to estimate the magnitude of the variance in observed scores due to universe-score

variance and to multiple sources of error—*occasions* and *items* in our example.

Estimated Variance Components

Variance component estimation is not unique to GT; it also underlies CT's approach to determining reliability. For example, a randomized blocks ANOVA can be used to estimate variance components and then calculate reliability. Only a single source of error (either occasions or items) can be considered at a time. If we consider occasions as introducing error, the score for each person on each occasion would be summed over items. The ANOVA provides mean squares (MS) for persons and the residual (person $\times$ occasion interaction confounded with random error). From the mean squares, variance components can be estimated. Table 3 provides the ANOVA summary information plus the expressions used to compute variance components for the persons-by-occasions (retest reliability) design. The reliability of the test is given by $(MS_p - MS_{res})/MS_p = \text{true-score variance/observed-score variance}$. The procedures for determining internal consistency reliability (persons $\times$ items design) are analogous.

Suppose we expand the ANOVA approach of CT to include both error *facets* in the achievement example: occasions *and* items. We could then estimate variance components for seven sources of variation: persons (p), occasions (o), items (i), the two-way interactions po, pi and oi, and the three-way interaction confounded with error (poi, e). The magnitude of the variance components would tell us how much each facet contributes to measurement error (see Table 4).

From these estimated variance components, a *generalizability coefficient*, analogous to a reliability coefficient, can be calculated by dividing the estimated person variance component by an estimated observed-score variance. A more formal development of the G coefficient, however, must await a treatment of relative and absolute decisions and their implication for the definition of what constitutes measurement error.

Relative and Absolute Decisions and the Generalizability Coefficient

CT is a theory of individual differences. In a $p \times o$ design, for example, there are three sources of variation (Table 3): p, o, and the $p \times o$ interaction confounded with error (we call the highest order interaction the residual). A reliability coefficient is calculated with two of the three sources of variation, persons and residual:

$$r_{xx'} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{res}^2/n_o}$$

The occasion effect should be zero because CT assumes strictly parallel measurements (but see Lord & Novick, 1968). Only the person and residual variation give rise to differences among individuals; the occasion effect is constant for all individuals in the $p \times o$ design.

Suppose, however, that to pass the written portion of a driving test one must answer at least 85% of the questions correctly. A passing score does not depend on

Table 2
Decomposition of an Observed Score in Classical and Generalizability Theory

Decomposition of observed score	Definition
Classical theory	
$X_{pi} = T_p + E_{pi}$	True score for person p Error for person p on item i
Generalizability theory	
$X_{poi} = \mu$	Grand mean
+ $\mu_p - \mu$	Person effect
+ $\mu_o - \mu$	Occasion effect
+ $\mu_i - \mu$	Item effect
+ $\mu_{po} - \mu_p - \mu_o + \mu$	Person $\times$ occasion effect
+ $\mu_{pi} - \mu_p - \mu_i + \mu$	Person $\times$ item effect
+ $\mu_{oi} - \mu_o - \mu_i + \mu$	Occasion $\times$ item effect
+ residual	$p \times o \times i$, error

Note. p = person; i = item; o = occasions; μ = constant.

Table 3
Expected Mean Squares for a Random-Effects Persons by Occasions Design

Source of variation	MS	Expected MS	Estimated variance component
Persons (p)	MS_p	$\sigma^2_{res} + n_o\sigma^2_p$	$\sigma^2_p = (MS_p - MS_{res})/n_o$
Occasion (o)	MS_o	$\sigma^2_{res} + n_p\sigma^2_o$	$\sigma^2_o = (MS_o - MS_{res})/n_p$
$p \times o, e$	MS_{res}	$\sigma^2_{po,e}$	$\sigma^2_{po,e} = MS_{res}$

Note. n_o = number of occasions; n_p = number of persons; e = error; res = residual.

the performance of others, and even if one ranks highly, a score of 84% will not be a passing grade. A large *occasion* effect suggests that one's score on the driving test depends on, for example, health or fatigue on the day one takes the test. When attention focuses on the actual value of the score and not one's standing among a group of individuals, as in driving tests, typing tests, or domain referenced tests, all three sources of variation in a $p \times o$ design (persons, occasions, and the residual) must be considered. Occasions can only be ignored if its variance component is very small.

GT distinguishes between decisions based on the relative standing of individuals (e.g., college admission) and decisions based on the absolute value of a score (e.g., driving license). Error in *relative decisions* arises from all nonzero variance components associated with rank ordering of individuals, other than the component for the object of measurement (persons). Specifically, variance components associated with the interaction of persons and each facet or combination of facets define error. In a $p \times o \times i$ design, for example, the variance components that contribute to error for relative decisions (σ^2_{Rel}) are σ^2_{po} , σ^2_{pi} , and $\sigma^2_{poi,e}$.

Decisions based on the level of the observed score, without regard to the performance of others, are called *absolute decisions*. All variance components associated with that score, except the component for the object of measurement, are defined as error. Error is associated not only with the ranking of individuals but also with constant effects associated with the facets. The difficulty

of a particular test form or the stringency of a particular interviewer will influence an individual's observed score and, hence, absolute decisions about that individual. In a $p \times o \times i$ design, then, the variance components that contribute to error variance for absolute decisions (σ^2_{Abs}) are σ^2_o , σ^2_i , σ^2_{po} , σ^2_{pi} , σ^2_{oi} , and $\sigma^2_{poi,e}$.

While stressing the importance of variance components and errors such as σ^2_{Rel} , GT also provides a coefficient analogous to the reliability coefficient in CT: true-score variance divided by observed score variance: $r_{xx'} = \sigma^2_t/\sigma^2_x$. GT's *generalizability coefficient*, E_p^2 , is the ratio of the universe-score variance to the expected observed-score variance, that is, an intraclass correlation:

$$E_p^2_{Rel} = \frac{\sigma^2_p}{E\sigma^2_x} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{Rel}}$$

(For absolute decisions, a generalizability coefficient can be defined in an analogous manner, using σ^2_{Abs} .) The expected observed-score variance is used because G theory assumes only random sampling of facet levels; observed-score variance may change from one application of the design to another. Sample estimates of the parameters in the G coefficient formula are used to estimate the G coefficient:

$$E_p^2_{Rel} = \frac{\hat{\sigma}^2_p}{\hat{\sigma}^2_p + \hat{\sigma}^2_{Rel}}$$

E_p^2 is a biased but consistent estimator of E_p^2 .

Generalizability (G) and Decision (D) Studies

A generalizability study is designed to estimate variance components underlying a measurement process by defining the universe of admissible observations as broadly as possible. Once data have been collected, a large set of variance components can be estimated. By optimizing the number of conditions of each facet, much in the way the Spearman-Brown prophecy formula does in CT, variance components can be reduced in magnitude and the measurement can be tailored to a particular purpose. Substantive studies that tailor the measurement to make decisions are known as D (for decision) studies. 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.

By design, then, the G study systematically treats all the facets that are likely to enter into the generalizations

Table 4
Generalizability Theory Applied to the Persons by Occasions by Items Design

Source of variation	Estimated variance component
Persons (p)	.06
Occasions (o)	.00
Items (i)	.03
$p \times o$	.01
$p \times i$	.00
$o \times i$	.00
Residual ($p \times o \times i$, error)	.24

Note. The Generalizability coefficient for two occasions, four items = .67, $\sigma^2_p / [\sigma^2_p + (\sigma^2_{po}/n_o) + (\sigma^2_{pi}/n_i) + (\sigma^2_{poi}/n_o n_i)]$.

of a diverse group of users. Hence, the G study will incorporate many more facets and variance component estimates than any one user would employ. To estimate all possible variance components, crossed designs are typically used.¹

In planning a D study, the decision maker (a) defines the universe of generalization and (b) specifies the proposed interpretation of the measurement. These plans determine (c) the questions to be asked of the G study data in order to optimize the D study's measurement design.

Defining the universe of generalization. G theory recognizes that the decision maker will generalize to a universe made up of some subset of the facets in the G study. This universe of generalization may be defined by reducing the levels of a facet (e.g., creating a fixed facet; cf. fixed factor in ANOVA), by selecting and thereby controlling one level of a facet, or by ignoring a facet.

Recognizing that there may be different universes of generalization allows us to resolve a paradox that is difficult to resolve with CT. Suppose that a performance is rated by several judges and that there are inconsistencies among them. The reliability of a composite score from all of the judges' ratings will be reduced by their disagreements and may even be increased by discarding all but one judge. However, even though the scores may be more reliable, who would prefer to use the score from a single judge rather than a composite score from many judges? Viewed from the framework of G theory, the choice is clear. The score from a single judge may be highly generalizable to a narrow universe: ratings given by that particular judge, with all of his or her idiosyncracies. A score obtained by combining ratings from several judges may be generalizable with less precision to a more interesting and important universe: ratings given by a population of judges.

Optimizing the measurement design. G studies provide information needed to optimize the measurement design for a D study. For example, in a study of hands-on performance of Navy machinist mates (Webb, Shavelson, Kim, & Chen, in press), variance components for a persons (machinist mates) $\times$ observers $\times$ tasks design were estimated. One question asked of the G study was whether the reliability of the measurement could be increased more efficiently by adding observers or tasks (Table 5). By adding tasks, we could reduce error variance much more than by adding observers, and the G coefficient would increase correspondingly. This is because the observer and persons $\times$ observer effects are remarkably small—a desirable measurement property if it arises from successful observer training and not from consultation among them.

D studies can be tailored not only by examining the trade-off between the number of conditions of different

Table 5
Estimated Variance Components for the Navy's Hands-on Performance Tests of Machinist Mates: Engine Room

Source of variation	$N_{obs} = 1$ $N_{tasks} = 1$	$N_o = 2$ $N_t = 11$	$N_o = 1$ $N_t = 18$
Persons (p)	.00626	.00626	.00626
Observers (o)	.00000	.00000	.00000
Tasks (t)	.00979	.00088	.00054
po	.00000	.00000	.00000
pt	.02584	.00235	.00144
ot	.00003	.00000	.00000
po, e	.00146	.00007	.00008
Generalizability coefficients			
Relative	.19	.72	.80
Absolute	.14	.66	.75

facets, but also by considering a wide variety of designs, including crossed, partially nested, and completely nested designs. All facets in the D design may be random (cf. random-model ANOVA), or only some may be random (cf. mixed-model ANOVA). Nested designs are often used for convenience, for increasing sample size, or for both. In general, the G study should be as fully crossed as possible, so that all sources of variation in the design can be estimated. From the results of a fully crossed G study, the generalizability for a wide variety of nested D study designs can be estimated.

Assumption of Steady State Behavior

Most measurement approaches, including CT and GT, assume that behavior remains constant over observations. Assessment of reliability is more complicated when behavior changes over time. Rogosa, Floden, and Willett (1984) showed that when behavior changes systematically over time (whether linearly or nonlinearly), the estimate of an individual's universe score will be time dependent. Inconsistent individual differences over time will arise whenever the trends for different individuals (whether they are due to unsystematic or systematic factors) vary.

A generalizability analysis with occasions as a facet can only identify some of these trends. Only when the variance components for occasions and the interaction between persons and occasions are both small can changes over time be ruled out. When one or the other component (or both) is relatively large, it is impossible to determine whether the changes over time are due to unsystematic variation in persons' behavior over time (the usual interpretation) or to different systematic trends over occasions for different people.

The best way to meet the steady state assumption is to restrict the time interval so that the observations of individuals' behavior represent the same phenomenon. If, even within a short time period, behavior cannot be assumed to be in a steady state, a generalizability analysis may not be appropriate. Rogosa et al. (1984) described

¹ The only exception would be D studies in which a particular nesting will definitely be used; in that case, the G study should incorporate the same nesting to provide the most precise estimates of the variance components corresponding to the nested facets; see Cronbach et al. (1972, p. 45).

ways to specify the process by which behavior changes over time and, consequently, to determine whether changes in behavior are systematic or unsystematic.

Rowley (in press) has pointed out that, even in the absence of any systematic time trend, there still may be dependencies among observations that may influence estimates of error variance and, hence, estimates of generalizability. When successive observations are correlated (e.g., there may be a nonzero correlation between the amount of time that an elementary school student spends reading from one occasion to the next; see Leinhardt, Zigmond, & Cooley, 1981), the usual methods of estimating error variance will produce underestimates (Gottman, 1981), resulting in inflated estimated generalizability coefficients. This potential problem can be avoided by carrying out nonsuccessive observations (e.g., every third day rather than on consecutive days; see Rowley, in press).

New Developments in Generalizability Theory

A variety of new developments in GT have emerged since the publication of Cronbach et al.'s (1972) monograph. Here we touch on only a few: (a) a new procedure for estimating variance components, (b) treatment of fixed facets in D-study designs, (c) symmetry in G-study designs, and (d) generalizability of a multivariate profile.

Estimated Variance Components

Shavelson and Webb (1981) characterized problems in estimating variance components as 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 (e.g., Cronbach et al., 1972; Lindquist, 1953; Shavelson & Webb, 1981; Smith, 1978).

Several methods have been proposed to treat negative estimates. Cronbach et al. (1972) recommended setting negative estimates to zero and carrying the zero through wherever that variance component enters the expected mean squares of another variance component. This produces biased estimates. Brennan (1983) recommended setting negative estimates to zero but using the negative estimate wherever that variance component enters the expected mean squares for another component. This procedure produces unbiased estimates. However, in neither case is it statistically comfortable to change estimates to zero.

Shavelson and Webb (1981) reviewed a Bayesian approach that sets a lower bound of zero on the estimated variance component and thereby produces nonnegative estimates. However, they concluded that this approach is not well enough developed to have widespread applicability.

Marcoulides (1987) examined another approach to estimating variance components that also restricts the estimate to zero or above: *restricted maximum likelihood estimation* (RMLE or REML now available in BMDP3V; see Genrich & Sampson, 1976). In all of his simulations, the RMLE variance component estimates were closer to the

true parameters than those from ANOVA. In contrast to RMLE, the ANOVA was sensitive to distributional form, incorrectly estimated variance components for unbalanced designs, and produced negative estimates. Only with balanced data sets from normal distributions and balanced one-way dichotomous distributions did the two methods perform similarly. Compared to RMLE, ANOVA tended to underestimate universe score variance and overestimate error variances. So, on average, the RMLE method produced higher estimated G coefficients than ANOVA. With the real datasets, the two methods provided very similar estimates of variance components and G coefficients.

Fixed Facets in G Theory

A fixed facet in GT is analogous to a fixed factor in ANOVA—the conditions of the facet (levels of the factor) in the G study exhaust the possible conditions of interest. A fixed facet may arise because the entire universe of conditions is small and is included in the measurement design or because the decision maker is not interested in generalizing beyond the conditions observed in the G study.

Statistically, G theory treats a fixed facet by averaging over the conditions of the facet and examining the generalizability of these averages over the random facets (Cronbach et al., 1972, p. 60; see Erlich & Shavelson, 1976, for a proof). Averaging provides the best score for an individual because it represents the individual's universe score over all conditions of the fixed facet. The statistical analysis of data from G studies with a fixed facet leads to conclusions about scores averaged over the conditions of the fixed facet. The treatment of fixed facets shows that G theory is essentially a random facet measurement theory or, more specifically, a random-effects ANOVA theory (Rubin, 1974).

When are facets fixed and when are they random?

A question immediately arises: Do we really randomly sample test items or test forms or judges (etc.)? In fact, we almost invariably take a convenient sample or include all possible levels of the facet (e.g., days of the week). Don't all psychometric procedures really produce fixed facets (Loevinger, 1965)?

There are several answers to this question. Cornfield and Tukey (1956) provided one answer: The inference is to the population from which the convenient sample was drawn. A second answer, perhaps somewhat less limiting, invokes the Bayesian notion of *exchangeability* (Novick, 1976): Even though conditions of a facet have not been sampled randomly, the facet may be considered to be random if conditions not observed in the G study are equally acceptable to, or exchangeable with, the observed conditions (see Shavelson & Webb, 1981, for details).

Procedures for handling fixed facets. Sometimes it may not make sense to average over the conditions of a fixed facet. For example, Erlich and Shavelson (1978) calculated the generalizability of measures of teachers' behavior by averaging over measures taken during mathematics and reading instruction. This seemed to be av-

eraging apples and oranges because of differences in the two types of instruction. If it does not make conceptual sense to average over the conditions of a fixed facet, or if conclusions about such averages are of little interest, a better strategy is to conduct a separate G study within each condition of the fixed facet. This permits, for example, different conclusions to be reached concerning the teaching of mathematics and reading.

An alternative strategy is to conduct a preliminary analysis, treating the fixed facet as random. If the variance components associated with that facet are small, we may (if it makes conceptual sense) average over the conditions of the fixed facet (see Shavelson & Webb, 1981, for details). If the component is large, we may conduct separate G studies for each condition of the fixed facet.

Subtests as Random Representatives of Their Domains

Cronbach (in press) has suggested a new treatment of reliability for test batteries in which several subtests measure a dimension. In the Stanford-Binet Intelligence Test, there are 15 subtests that index four dimensions: short-term memory, and verbal, quantitative, and abstract/visual reasoning. These dimensions are combined to create a composite or total standard age score. To determine the CT reliability of the composite, we would most likely calculate the reliability of each subtest and then use a version of the Jackson-Ferguson (1941) procedure to evaluate the reliability of the composite.

But if subtests are indicators of a construct (e.g., verbal reasoning), the analysis might better view subtests as random and evaluate the adequacy of the test score as a representation of the domain of verbal reasoning subtests. It seems to me that an interpretation that a pupil is better in verbal than in Abstract Reasoning [sic] is a statement about the domain, not the fixed subtests. (Cronbach, personal communication, July 15, 1987)

Cronbach (in press) treated "subtests as random representatives of their domains, to evaluate how well each section [subtest] score measures its postulated latent trait" (p. 6). In estimating reliability, he considered each dimension separately (see the section on fixed facets presented earlier) and calculated an internal consistency coefficient treating variability among subtests within a dimension as measurement error.

In a G study of the Stanford-Binet, then, subjects would constitute the object of measurement (random). Dimensions would be considered a fixed facet covering the universe of dimensions of intelligence, or at least the universe to which the decision maker is interested in generalizing. Subtests would be considered a random facet, nested within their respective dimensions (cf. Cronbach et al., 1972, pp. 234-256).

Symmetry of Behavioral Measurements

The purpose of behavioral measurement has typically been to differentiate *individuals* and, more recently, to locate their scores within some subject matter or skill domain. In CT, we customarily think about persons as the objects of measurement and judge the suitability of a measure in terms of how well it enables us to distinguish

between persons. All other facets (test items, occasions, raters, etc.) are treated as error. Cardinet and his colleagues (Cardinet & Tourneur, 1985; Cardinet, Tourneur, & Allal, 1976, 1981) recognized, however, that the *object of measurement* may change. Particularly in educational research, interest might attach to rates of success on a particular test item as an indicator of change, as has been the practice with the National Assessment of Educational Progress. Alternatively, classrooms, rather than individual students, might be the focus, even though we have a test score for each student in the class. In these cases, individual differences may give rise to measurement error (e.g., Gillmore, Kane, & Naccarato, 1978; Kane & Brennan, 1977).

Cardinet, Tourneur, and Allal (1976, 1981; Cardinet & Tourneur, 1985) proposed the *principle of symmetry* to enable G theory to address the situation in which the object of measurement might change. This principle asserts that any facet in a design may be regarded as the object of measurement. We may ask how well data enable us to distinguish among classes or schools or among instructional objectives. Such questions can be handled within GT: Differences among persons, instead of being the object of the measurements, obscure the distinctions being sought and are therefore "error." Variance components may be computed from a dataset regardless of the measurement design; the treatment of the variance components becomes necessarily asymmetrical only when a decision is made about the object of measurement. "Until the two kinds of problems (ANOVA estimates, measurement) were clearly disentangled, no multipurpose measurement was possible" (Cardinet, personal communication, May 9, 1980).

Generalizability of a Multivariate Profile

Behavioral measurements often involve multiple scores describing individuals' aptitudes or skills. As noted earlier, the Stanford-Binet intelligence test, for example, uses 15 subtests to measure four correlated dimensions. Although multiple scores may be conceived as vectors and thus should be treated simultaneously, the majority of G and D studies have not done so. Rather, each variable has been treated separately, perhaps because the multivariate literature is not easily comprehended (e.g., Bock, 1963, 1966; Cronbach et al., 1972). An attempt to elucidate multivariate G theory has been made elsewhere (e.g., Shavelson, 1985; Webb & Shavelson, 1981; Webb, Shavelson, & Maddahian, 1983), but several recent developments may be important to future work.

G coefficient for a multivariate composite. In forming a composite for a multivariate measure such as the Stanford-Binet, a number of alternatives present themselves: Take the sum of the four dimensions, take a weighted sum based on expert judgment from theory, or take a weighted sum based on the variance of dimension scores. From a GT perspective, we might weight the dimension scores to maximize the G coefficient (Joe & Woodward, 1976; Woodward & Joe, 1973) and report a G coefficient for the multivariate composite. One problem

with weightings based on variances or on maximizing a G coefficient is that the weights are "blind" to the theory that originally gave rise to the profile of measures. Statistical criteria for weighting measures do not necessarily make conceptual sense.

A second option, one based on the conceptual underpinnings of the profile, is to determine the weights by expert judgment. However, even expert judges may have differing views about the relative importance of elements in a composite. There may be no reason for preferring one judge's weights over another, and there is no guarantee that an average set of weights is in any sense the "correct" set. Another problem is that this choice of weights ignores the statistical properties of the data and therefore results in a composite whose psychometric properties may be poor.

A third alternative is to posit a conceptual model of the profile, fit this model statistically with confirmatory factor analysis, and then weight each variable in the composite according to its loading from the factor analysis. In this way, expert judgment in setting forth the conceptual model is combined with statistical accuracy for obtaining the appropriate weights to fit the model.

Short, Webb, and Shavelson (1986) compared these approaches in analyses of addition, subtraction, multiplication, and division subtests (Table 6). As expected, the multivariate G coefficient was highest with the weights derived for the coefficient itself. However, the weights derived from the confirmatory factor analysis produced nearly as high a multivariate G coefficient. Expert judgment resulted in a considerably lower G coefficient. If this finding is replicated, confirmatory factor analysis and G theory might be merged to handle multivariate composites.

Illustrative Application

One of GT's greatest strengths lies in its ability to model a remarkably wide array of behavioral measurement applications. Here we present only one example, a study of the general educational development ratings contained in the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1965). We hope it gives a flavor of the flexibility of the theory.

The U.S. Department of Labor (1972) developed

the General Educational Development (GED) scale to characterize the amount of *reasoning*, *mathematics*, and *language* ability needed to perform various jobs. GED ratings are used in several employment and training situations, such as (a) estimating the time required to learn job skills; (b) deciding whether to refer persons to specific employers, job training programs, or remedial education programs; and (c) equating jobs with similar educational requirements.

In a G study of GED ratings (Webb & Shavelson, 1981; Webb, Shavelson, Shea, & Morello, 1981), job analysts were given written descriptions of jobs published in the *Dictionary of Occupational Titles* (DOT) and were asked to rate the jobs on three dimensions of the GED scale: reasoning development, mathematics development, and language development. Each of 71 raters from 11 field centers across the United States evaluated the three dimensions of a sample of jobs on two occasions. Different centers had from 2 to 12 analysts. Hence, the G-study design was a partially nested, unbalanced design with different numbers of raters nested within centers.

Because we were concerned with estimating the general educational development required to perform each job, *jobs* is the *object of measurement*, and the variance component for jobs, σ_j^2 , is interpreted as the universe-score variance. All other variance components were considered measurement error because *absolute* decisions are made regarding GED requirements for each job. Measurement error, then, was defined as raters nested within center, centers, occasions, and all interactions.

Univariate G study. Due to the unbalanced nature of the design, variance components were estimated in two ways (a) by dropping analysts randomly until we had two analysts per center, thus creating a balanced design; and (b) by Rao's MIVQUE procedure for estimating variance components in unbalanced designs (Hartley, Rao, & La Motte, 1978). We report the estimates in Table 7 for each dimension separately because we considered *dimension* to be a fixed facet, with the three ratings interpreted separately by decision makers.

The estimated variance components for jobs (universe-score variance) differ across GED ratings and are greatest for language. This indicates that jobs were seen to differ more in their demands for language than in their demands for mathematics and reasoning. The patterns of variance components contributing to error were consistent across dimensions: The interaction of raters with jobs accounted for most of the error variation, and occasions and centers accounted for little. This pattern suggested that, by taking the average ratings given by four raters, measurement error could be reduced by about 75% and the G coefficient correspondingly increased to 0.86 for reasoning, 0.79 for mathematics, and 0.85 for language.

Multivariate G study. For the multivariate G study, a random effects, multivariate ANOVA was used to estimate variance and *covariance* components with reasoning, mathematics, and language ratings as a vector of scores. Due to the limited capacity of computer programs available to perform the multivariate analysis and because

Table 6
Generalizability (G) Coefficients Resulting
From Four Weighting Schemes

Scheme	Weights			G coefficient	
	A/S ^a	M ^b	D ^c	n _o = 1 ^d	n _o = 2 ^d
Multivariate G	0.11	0.07	0.35	0.71	0.83
Decision maker	2.00	3.00	1.00	0.59	0.74
Unit weights	1.00	1.00	1.00	0.65	0.79
Factor loadings	2.05	2.09	4.87	0.70	0.82

^a Addition/subtraction. ^b Multiplication. ^c Division. ^d Number of occasions.

Table 7
Univariate Generalizability Study of GED Ratings

Source of variation	Estimated variance component					
	Reasoning ^a	Math ^a	Language ^a	Reasoning ^b	Math ^b	Language ^b
Jobs (j)	.74	.63	1.01	.74	.63	1.01
Occasions (o)	.00	.00	.00	.00	.00	.00
Centers (c)	.00	.02	.05	.00	.02	.05
Raters (r:c)	.06	.02	.00	.02	.00	.00
jo	.00	.01	.01	.00	.01	.01
jc	.00	.00	.00	.00	.00	.00
jr:c	.13	.16	.14	.03	.04	.04
oc	.01	.00	.00	.01	.00	.00
or:c	.00	.09	.07	.00	.02	.02
joc	.00	.02	.01	.00	.02	.01
jo:r:c	.22	.25	.22	.06	.06	.05
Absolute error	.42	.57	.50	.11	.17	.18
$\hat{\rho}^2$ (Absolute G coefficient)	.64	.53	.67	.86	.79	.85

Note. Generalizability (G) study design is raters nested within centers crossed with jobs and occasions. GED = General Educational Development; n_r denotes number of raters; n_o is the number of occasions; n_c is the number of centers.

^a $n_r = n_o = n_c = 1$. ^b $n_r = 4$, $n_o = 1$, $n_c = 1$.

geographic center contributed almost nothing to variability among job ratings, geographic center was disregarded in this analysis. (The reader may note that the case for dropping occasions was equally strong; this is reflected in the results that follow). The design, then, was raters crossed with jobs and occasions.

For each source of variation, variance and covariance component matrices were computed from the mean product matrices. One matrix, for example, contained estimated universe-score variances and covariances. All negative estimated variance components (diagonal values) were set equal to zero in further estimation. For this analysis, the variance component matrices, multivariate G coefficients, and canonical weights corresponding to each G coefficient were computed.

The estimated variance and covariance matrices are presented in Table 8.² Because centers were dropped, the variance components in Table 8 are similar to but do not equal those in the univariate analysis (Table 7). The components of covariance, however, provide new information. The large components for jobs reflect the underlying (dis-attenuated) covariances among the GED components. Jobs that require high reasoning ability are seen by raters to require high mathematics and language ability. The nonzero components of variance for raters indicate that some raters give generally higher ratings than others. The positive covariance components for raters indicate that the raters who give higher ratings on one GED component are likely to give higher ratings on the other GED components. The positive variance and covariance components for the job by rater interaction suggest that not only

Table 8
Estimated Variance-Covariance Components for the Multivariate Generalizability Study of GED Ratings

Source of variation	Reasoning	Math	Language
Jobs (j)			
Reasoning	.75		
Math	.64	.66	
Language	.88	.74	1.09
Occasions (o)			
Reasoning	.00		
Math	.00	.00	
Language	.00	.00	.00
Raters (r)			
Reasoning	.03		
Math	.03	.09	
Language	.03	.05	.05
jo			
Reasoning	.00		
Math	.00	.00	
Language	.00	.00	.00
jr			
Reasoning	.12		
Math	.11	.13	
Language	.09	.07	.11
or			
Reasoning	.00		
Math	.01	.01	
Language	.00	.00	.01
jo, e			
Reasoning	.21		
Math	.07	.29	
Language	.11	.10	.26

Note. The G study design is raters nested within centers and crossed with jobs and occasions with scores on reasoning, mathematics, and language.

² Only the components for one rater and one occasion are included. To obtain results for four raters, the components corresponding to the rater effect and rater interactions only need to be divided by four.

do raters disagree about which jobs require more ability but their disagreement is consistent across GED components. The nonzero components for error suggest that the unexplained factors that contribute to the variation of ratings also contribute to the covariation between ratings. As expected, the components of variance and covariance due to the occasion main effect and its interactions are negligible.

The GED composites with maximum generalizability are presented in Table 9. When multivariate generalizability was estimated for one rater and one occasion, a single dimension emerged with a coefficient exceeding .50. This dimension is a verbal composite of reasoning and language. The analysis with four raters and one occasion produced two dimensions with coefficients over .50. One composite was defined by reasoning and language (G coefficients .74 and .92 for one and four raters, respectively); the other by a mathematics–language contrast with a G coefficient of .62 for four raters and one occasion.

Summary

GT extends CT by recognizing and simultaneously estimating the magnitude of the multiple possible sources of error in a behavioral measurement, such as occasions, forms, raters, and items. The theory also goes beyond CT by distinguishing between the use of a measurement to make relative or “norm-referenced” decisions and absolute or “domain-referenced” decisions. Error is defined differently for decisions based on ordering of individuals (e.g., in selection situations) and on an individual’s observed scores (e.g., typing speed test). Consequently, reliability (*generalizability*) is defined differently for the two types of decisions. GT also recognizes that there is a potentially large number of reliability (G) coefficients associated with a particular measurement. Different G coefficients can be defined according to how broadly or narrowly each decision maker proposes to generalize from a particular application of the measurement procedure. Finally, GT enables decision makers to pinpoint the major sources of error entering into the measurement and determine the most cost-effective ways to apply the measurement procedure for their purpose.

Table 9
Canonical Variates for Multivariate Generalizability Study of GED Ratings

GED component	Canonical coefficients		
	Dimension I ^a	Dimension I ^b	Dimension II ^b
Reasoning	.34	.38	.05
Mathematics	.06	.06	−1.95
Language	.51	.57	1.33
Multivariate G coefficient	.74	.92	.62

^a For Dimension I, $n_r = 1$ and $n_o = 1$ (n_r = number of raters and n_o = number of occasions). ^b For Dimensions I and II, $n_r = 4$ and $n_o = 1$.

GT, then, provides perhaps the most flexible measurement theory available to psychologists, one that fits a wide variety of practical and theoretical applications. We surmise that the theory is not widely applied in psychological research because of its fairly formidable mathematical development (Cronbach et al., 1972; see also Brennan, 1983). The purpose of this article was to make GT accessible to the potential user. We hope that we succeeded by drawing an analogy between the use of ANOVA to address research questions and the use of GT to address measurement questions, by contrasting CT and GT, and by providing a practical application that illustrates GT’s concepts and procedures.

REFERENCES

- Bock, R. D. (1963). Multivariate analysis of variance of repeated measurements. In C. W. Harris (Ed.), *Problems of measuring change* (pp. 85–103). Madison, WI: University of Wisconsin Press.
- Bock, R. D. (1966). Contributions of multivariate experimental designs to educational research. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 820–840). Chicago: Rand McNally.
- Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.
- Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure* [Dependability of the measure]. Bern, Switzerland: Lang.
- Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, 13, 119–135.
- Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, 18, 183–204.
- Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 27, 907–949.
- Cronbach, L. J. (in press). Stanford-Binet: Edition 4. In *Mental measurements yearbook*. Lincoln: University of Nebraska Press.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: Wiley.
- Erlach, O., & Shavelson, R. J. (1976). *The application of generalizability theory to the study of teaching* (Technical Report No. 76-9-1, Beginning Teacher Evaluation Study). San Francisco: Far West Laboratory.
- Erlach, O., & Shavelson, R. J. (1978). The search for correlations between measures of teacher behavior and student achievement: Measurement problem, conceptualization problem or both? *Journal of Educational Measurement*, 15, 77–89.
- Genrich, R. I., & Sampson, P. F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, 18, 11–17.
- Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of course and teacher effects. *Journal of Educational Measurement*, 15, 1–14.
- Gleser, G. C., Green, B. L., & Winget, C. N. (1978). Quantifying interview data on psychic impairment of disaster survivors. *The Journal of Nervous and Mental Diseases*, 166, 209–216.
- Goodenough, F. L. (1936). A critical note on the use of the term “reliability” in mental measurement. *Journal of Educational Psychology*, 27, 173–178.
- Gottman, J. M. (1981). *Time-series analysis: A comprehensive introduction for social scientists*. Cambridge, England: Cambridge University Press.
- Hartley, H. O., Rao, J. N. K., & La Motte, L. (1978). A simple synthesis-based method of variance component estimation. *Biometrics*, 34, 233–242.
- Jackson, R. W. B., & Ferguson, G. A. (1941). *Studies on the reliability of tests*. Toronto: University of Toronto Press.
- Joe, G. N., & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika*, 41, 205–217.

- Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, 47, 267-292.
- Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Educational Research Journal*, 18, 343-361.
- Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston: Houghton-Mifflin.
- Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, 72, 143-155.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Marcoulides, G. A. (1987). *An alternative method for variance component estimation: Applications to generalizability theory*. Unpublished doctoral dissertation, University of California, Los Angeles.
- Novick, M. R. (1976). Bayesian methods in educational testing: A third survey. In D. M. N. de Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational testing*. New York: Wiley.
- Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, 76, 1000-1027.
- Rowley, G. (in press). Assessing error in behavioral data: Problems of sequencing. *Journal of Educational Measurement*.
- Rubin, D. B. (Reviewer). (1974). The dependability of behavioral measurements: Theory of generalizability for scores and profiles [Review]. *Journal of the American Statistical Association*, 69, 1050.
- Shavelson, R. J. (1985). *Generalizability of military performance measurements: I. Individual performance*. Washington, DC: National Research Council, Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education.
- Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology*, 34, 133-166.
- Short, L., Webb, N. M., & Shavelson, R. J. (1986, April). *Issues in multivariate generalizability: Weighting schemes and dimensionality*. Paper presented at the meeting of the American Educational Research Association, San Francisco, CA.
- Smith, P. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Measurement*, 3, 319-346.
- Thorndike, R. L. (1947). *Research problems and techniques* (Report No. 3, AAF Aviation Psychology Program Research Reports). Washington, DC: U.S. Government Printing Office.
- U.S. Department of Labor. (1965). *Dictionary of occupational titles* (3rd ed., vol. 2). Washington, DC: U.S. Government Printing Office.
- U.S. Department of Labor. (1972). *Handbook for analyzing jobs*. Washington, DC: U.S. Government Printing Office.
- Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement*, 18, 13-22.
- Webb, N. M., Shavelson, R. J., Kim, K-S., & Chen, Z. (in press). Reliability (generalizability) of job performance measurements: Navy machinist mates. *Journal of Military Psychology*.
- Webb, N. M., Shavelson, R. J., & Maddahian, E. (1983). Multivariate generalizability theory. In L. J. Fyans, Jr. (Ed.), *Generalizability theory: Inferences and practical applications* (pp. 67-81). San Francisco: Jossey-Bass.
- Webb, N. M., Shavelson, R. J., Shea, J., & Morello, E. (1981). Generalizability of general educational development ratings of jobs in the U.S. *Journal of Applied Psychology*, 66, 186-191.
- Woodward, J. A., & Joe, G. W. (1973). Maximizing the coefficient of generalizability in multi-facet decision studies. *Psychometrika*, 38, 173-181.