An NCME Instructional Module on

Generalizability Theory

Robert L. Brennan
American College Testing

Generalizability theory consists of a conceptual framework and a methodology that enable an investigator to disentangle multiple sources of error in a measurement procedure. The roots of generalizability theory can be found in classical test theory and analysis of variance (ANOVA), but generalizability theory is not simply the conjunction of classical theory and ANOVA. In particular, the conceptual framework in generalizability theory is unique. This framework and the procedures of generalizability theory are introduced and illustrated in this instructional module using a hypothetical scenario involving writing proficiency.

Historically, in psychology and education, measurement issues have been addressed principally using classical test theory, which postulates that an observed score can be decomposed into a "true" score and a single, undifferentiated random error term, E. Generalizability theory liberalizes classical theory by providing models and methods that allow an investigator to disentangle multiple sources of error that contribute to E. This is accomplished in part through the application of certain ANOVA methods.

In a sense, then, classical test theory and ANOVA can be viewed as the parents of generalizability theory. However, this analogy limps somewhat, because the ANOVA issues emphasized in generalizability theory are different from those that predominate in experimental design and ANOVA texts. More importantly, however, generalizability theory has a unique conceptual framework. Among the concepts in this framework are *universes of admissible observations* and *G* (Generalizability) *studies*, as well as *universes of generalization* and *D* (Decision) *studies*. These concepts and the methods of generalizability theory are introduced here using a hypothetical scenario involving the measurement of writing proficiency. As illustrated by this scenario, generalizability analyses are useful not only for understanding the relative importance of various sources of error but also for designing efficient measurement procedures.

Universe of Admissible Observations and G Study Considerations

Suppose an investigator, Mary Smith, decides that she wants to construct one or more measurement procedures for evaluating writing proficiency. She might proceed as follows. First she might identify, or otherwise characterize, essay prompts that she would consider using, as well as potential raters of writing proficiency. At this point, Smith is not committing herself to actually using, in a particular measurement procedure, any specific items or raters—or, for that matter, any specific number of items or raters. She is merely characterizing the facets of measurement that might interest her or other investigators. A facet is simply a set of similar conditions of measurement. Specifically, Smith is saying that any one of the essay prompts constitutes an admissible (i.e., acceptable to her) condition of measurement for her essay-prompt facet. Similarly, any one of the raters constitutes an admissible condition of measurement for her rater facet. We say that Smith's universe of admissible observations contains an essay-prompt facet and a rater facet.

Furthermore, suppose Smith would accept as meaningful to her a pairing of any rater (r) with any prompt (t). If so, Smith's universe of admissible observations would be described as being crossed, and it would be denoted $t \times r$, where the "$\times$" is read "crossed with." Specifically, if there were N_t prompts and N_r raters in Smith's universe, then it would be described as crossed if any one of the $N_t N_r$ combinations of conditions from the two facets would be admissible for Smith. Here, it will be assumed that N_t and N_r are both very large—approaching infinity, at least theoretically.

Note that it is the particular investigator, Smith, who decides which prompts and which raters constitute the universe of conditions for the prompt and rater facets, respectively. Generalizability theory does not presume that there is some particular definition of prompt and rater facets that all investigators would accept. For example, Smith might characterize the potential raters as college instructors with a PhD in English, whereas another investigator might be concerned about a rater facet consisting of high school teachers of English. If so, Smith's universe of admissible observations may be of little interest to the other investigator. This does not invalidate Smith's universe, but it does suggest that other investigators need to pay careful attention to Smith's statements about facets if they are to judge the relevance of Smith's universe of admissible observations to their own concerns.

In the above scenario, no explicit reference has been made to persons who respond to the essay prompts in the universe of admissible observations. However, Smith's ability to specify a meaningful universe of prompts and raters is surely, in some sense, dependent upon her ideas about a population of examinees for whom the prompts and raters would be appropriate. Without some such notion, any characterization of prompts and raters as "admissible" seems vague at best. Even so, in generalizability theory the word *universe* is reserved for conditions of measurement (prompts and raters, in the scenario),

Robert L. Brennan is a Distinguished Research Scientist in the Measurement and Statistical Research Area at American College Testing, 2201 North Dodge St., P.O. Box 168, Iowa City, IA 52243.

Series Information
ITEMS is a series of units designed to facilitate instruction in education measurement. These units are published by the National Council on Measurement in Education.

whereas the word *population* is used for the objects of measurement (persons, in this scenario).

Presumably, Smith would accept as admissible the response of any person in the population to any prompt in the universe evaluated by any rater in the universe. If so, the population and universe of admissible observations are crossed, which is represented $p \times (t \times r)$, or simply $p \times t \times r$. For this situation, any observable score for a single essay prompt evaluated by a single rater can be represented as:

$$X_{ptr} = \mu + \nu_p + \nu_t + \nu_r + \nu_{pt} + \nu_{pr} + \nu_{tr} + \nu_{ptr}, \quad (1)$$

where μ is the grand mean in the population and universe and ν designates any one of the seven uncorrelated effects, or components, for this design. (Actually, the effect *ptr* is a residual effect involving the triple interaction and all other sources of error not explicitly represented in the universe of admissible observations.)

This population and universe can also be represented in terms of the Venn diagram in Figure 1. In this diagram, the three circles represent persons, essay prompts, and raters; circle-overlap areas represent interactions; and the seven distinct areas correspond to the seven effects.

The variance of the scores given by Equation 1, over the population of persons and the conditions in the universe of admissible observations is:

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr). \quad (2)$$

That is, the total observed score variance can be decomposed into seven independent variance components. It is assumed here that the population and both facets in the universe of admissible observations are infinite. Under these assumptions, the variance components in Equation 2 are called *random effects* variance components. It is important to note that they are for *single* person-prompt-rater combinations, as opposed to average scores over prompts and/or raters. Average scores are considered in D studies.

Now that Smith has specified her population and universe of admissible observations, she needs to collect and analyze data to estimate the variance components in Equation 2. To do so, Smith conducts a study in which, let us suppose, she has a sample of n_r raters use a particular scoring procedure to evaluate each of the responses by a sample of n_p persons to a sample of n_t essay prompts. Such a study is called a *G* (Generalizability) *study*. The design of this particular study (i.e., the G study design) is denoted $p \times t \times r$. We say this is a two-facet design because the objects of measurement (persons) are not usually called a "facet." Given this design, the usual procedure for estimating the variance components in Equation 2 employs the expected mean square (EMS) equations in Table 1. The resulting estimators of these variance components, in terms of mean squares, are also provided in Table 1. These estimators are for a random effects model.

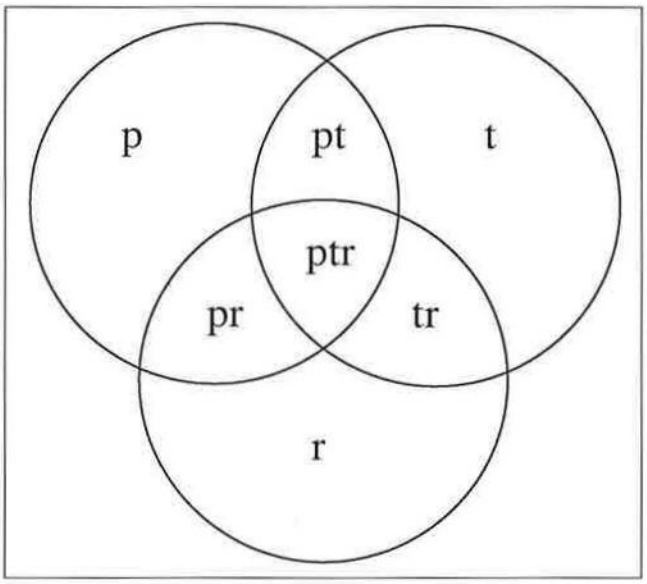

FIGURE 1. *Venn diagram for p × t × r design*

Suppose the following estimated variance components are obtained from Smith's G study:

$$\hat{\sigma}^2(p) = .25, \quad \hat{\sigma}^2(t) = .06, \quad \hat{\sigma}^2(r) = .02,$$

$$\hat{\sigma}^2(pt) = .15, \quad \hat{\sigma}^2(pr) = .04, \quad \hat{\sigma}^2(tr) = .00,$$

$$\text{and} \quad \hat{\sigma}^2(ptr) = .12. \quad (3)$$

These are estimates of the actual variances (parameters) in Equation 2. For example, $\hat{\sigma}^2(p)$ is an estimate of the variance component $\sigma^2(p)$, which can be interpreted roughly in the following manner. Suppose that, for each person in the population, Smith could obtain each person's mean score (technically, "expected" score) over all N_t essay prompts and all N_r raters in the universe of admissible observations. The variance of these mean scores (over the population of persons) is $\sigma^2(p)$. The other "main effect" variance components for the prompt and rater facets can be interpreted in a similar manner. Note that for Smith's universe of admissible observations the estimated variance attributable to essay prompts, $\hat{\sigma}^2(t) = .06$, is three times as large as the estimated variance attributable to raters, $\hat{\sigma}^2(r) = .02$. This suggests that prompts differ much more in average difficulty than raters differ in average stringency.

Interaction variance components are more difficult to interpret verbally, but approximate statements can be made. For example, $\hat{\sigma}^2(pt)$ estimates the extent to which the relative ordering of persons differs by essay prompt, and $\hat{\sigma}^2(pr)$ estimates the extent to which persons are rank ordered differently by different raters. For the illustration considered here, it is especially important to note that $\hat{\sigma}^2(pt) = .15$ is almost four times as large as $\hat{\sigma}^2(pr) = .04$. This fact, combined with the previous observation that $\hat{\sigma}^2(t)$ is three times as large as $\hat{\sigma}^2(r)$, suggests that prompts are a considerably greater source of variability in persons' scores than are raters. The implication and importance of these facts will become evident in subsequent sections.

D Study Considerations for the $p \times T \times R$ Design and an Infinite Universe of Generalization

The purpose of a G study is to obtain estimates of variance components associated with a universe of admissible observations. These estimates can be used to design efficient measurement procedures for operational use and to provide information for making substantive decisions about objects of measurement (usually persons) in various *D* (Decision) *studies*. Broadly speaking, D studies emphasize the estimation, use, and interpretation of variance components for decision-making with well-specified measurement procedures.

Perhaps the most important D study consideration is the specification of a universe of generalization, which is the universe to which a decision-maker wants to generalize based on the results of a D study with a particular measurement procedure. To understand the concept of a universe of generalization, it is helpful to consider certain D study design issues, first.

D Study p × T × R Design

Let us suppose that Smith decides to design her measurement procedure such that each person will respond to n'_t essay

Table 1
Expected Mean Squares and Estimators of Variance Components for the G Study $p \times t \times r$ Design

Effect (α)	EMS(α)	$\hat{\sigma}^2(\alpha)$
p	$\sigma^2(ptr) + n_t\sigma^2(pr) + n_r\sigma^2(pt) + n_tn_r\sigma^2(p)$	$[MS(p) - MS(pt) - MS(pr) + MS(ptr)]/n_tn_r$
t	$\sigma^2(ptr) + n_p\sigma^2(tr) + n_r\sigma^2(pt) + n_pn_r\sigma^2(t)$	$[MS(t) - MS(pt) - MS(tr) + MS(ptr)]/n_pn_r$
r	$\sigma^2(ptr) + n_p\sigma^2(tr) + n_t\sigma^2(pr) + n_pn_t\sigma^2(r)$	$[MS(r) - MS(pr) - MS(tr) + MS(ptr)]/n_pn_t$
pt	$\sigma^2(ptr) + n_r\sigma^2(pt)$	$[MS(pt) - MS(ptr)]/n_r$
pr	$\sigma^2(ptr) + n_t\sigma^2(pr)$	$[MS(pr) - MS(ptr)]/n_t$
tr	$\sigma^2(ptr) + n_p\sigma^2(tr)$	$[MS(tr) - MS(ptr)]/n_p$
ptr	$\sigma^2(ptr)$	$MS(ptr)$

Note. α represents any one of the effects.

prompts, with each response to every prompt evaluated by the same n_r' raters. Furthermore, assume that decisions about a person will be based on his or her mean score over the $n_t'n_r'$ observations associated with the person. This is a verbal description of the $p \times T \times R$ design for a D study. It appears to be much like the $p \times t \times r$ design for Smith's G study, but there are two important differences.

First, the sample sizes for the D study (n_t' and n_r') need not be the same as the sample sizes for the G study (n_t and n_r). This distinction is highlighted by the use of primes with D study sample sizes. Second, for the D study, interest focuses on *mean* scores for persons, rather than single person-prompt-rater observations that are the focus of G study estimated variance components. This emphasis on mean scores is highlighted by the use of upper-case letters for the facets in Smith's D study $p \times T \times R$ design.

Relating Smith's D Study and an Infinite Universe of Generalization

The universe of generalization can be conceptualized as a universe of measurement procedures each employing the specified D study sample sizes and design structure. In generalizability theory these measurement procedures are described as "randomly parallel," and it is assumed that any particular measurement procedure consists of a random sample of conditions for *at least one* facet (e.g., essay prompts, raters, or both). Randomly parallel measurements need not have equal means, which is an assumption for classically parallel measurements.

Here, let us suppose that Smith decides that, in theory, any one of the randomly parallel instances of her measurement procedure would involve a *different* sample of n_t' essay prompts and a *different* sample of n_r' raters from her universe of admissible observations. As such, replications of her measurement procedure would span a universe that theoretically includes all the prompts and raters in her universe of admissible observations. Under these circumstances, we would describe Smith's universe of generalization as being *infinite*. More specifically, for Smith's universe of generalization, the rater and item facets are both infinite. In analysis of variance terminology, this model is described as *random*. (For this reason, it is sometimes stated that prompt and rater facets are random.) In short, under this scenario, Smith wants to generalize persons' scores based on the specific prompts and raters in her measurement procedures to their scores for a universe of generalization that involves many other prompts and raters.

Universe Scores

In principal, for any person, Smith can conceive of obtaining the person's mean score for every instance of the measurement procedure in her universe of generalization. For any such person, the expected value of these mean scores is defined as the person's *universe score*.

The variance of universe scores over all persons in the population is called *universe score variance*. It has conceptual similarities with true score variance in classical test theory.

D Study Random Effects Variance Components

For Smith's D study $p \times T \times R$ design the linear model for an observable mean score over n_t' essay prompts and n_r' raters can be represented as:

$$X_{pTR} = \mu + \nu_p + \nu_T + \nu_R + \nu_{pT} + \nu_{pR} + \nu_{TR} + \nu_{pTR}. \quad (4)$$

The variances of the score effects in Equation 4 are called *D study variance components*. When it is assumed that the population and all facets in the universe of generalization are infinite, these variances components are *random effects* variance components. They can be estimated using the G study estimated variance components in Equation Set 3.

For example, suppose Smith wants to consider using the sample sizes $n_t' = 3$ and $n_r' = 2$ for her measurement procedure. If so, the estimated D study random effects variance components are

$$\hat{\sigma}^2(p) = .25, \quad \hat{\sigma}^2(T) = .02, \quad \hat{\sigma}^2(R) = .01,$$
$$\hat{\sigma}^2(pT) = .05, \quad \hat{\sigma}^2(pR) = .02, \quad \hat{\sigma}^2(TR) = .00, \quad (5)$$
$$\text{and } \hat{\sigma}^2(pTR) = .02.$$

These estimated variance components are for person *mean scores* over $n_t' = 3$ essay prompts and $n_r' = 2$ raters.

Rule. Obtaining these results is simple. Let $\hat{\sigma}^2(\alpha)$ be any one of the G study estimated variance components. To get the estimated D study variance components, one simply divides $\hat{\sigma}^2(\alpha)$ by n_t' if α contains t but not r, by n_r' if α contains r but not t, and by $n_t'n_r'$ if α contains both t and r.

The estimated variance component $\hat{\sigma}^2(p) = .25$ is particularly important because it is the estimated universe score variance in this scenario. In terms of parameters, when prompts and raters are both random, universe score is defined as

$$\mu_p = E_T E_R X_{pTR} = \mu + \nu_p, \quad (6)$$

where E stands for expected value. The variance of universe scores (i.e., universe score variance) is denoted generically $\sigma^2(\tau)$, and it is simply $\sigma^2(p)$, here.

Error Variances

Given Smith's infinite universe of generalization, variance components other than $\sigma^2(p)$ contribute to one or more different types of error variance. Considered below are "absolute" and "relative" error variances.

Absolute error variance, $\sigma^2(\Delta)$. Absolute error is simply the difference between a person's observed and universe scores:

$$\Delta_p = X_{pTR} - \mu_p. \tag{7}$$

For this scenario, given Equations 4 and 6,

$$\Delta_p = \nu_T + \nu_R + \nu_{pT} + \nu_{pR} + \nu_{TR} + \nu_{pTR}. \tag{8}$$

Consequently, the variance of the absolute errors, $\sigma^2(\Delta)$, is the sum of all the variance components except $\sigma^2(p)$. This result is also provided in Table 2 under the column headed "*T*, *R* random."

Given the estimated D study variance components in Equation Set 5, the estimate of $\sigma^2(\Delta)$ for three prompts and two raters is:

$$\hat{\sigma}^2(\Delta) = .02 + .01 + .05 + .02 + .00 + .02 = .12,$$

and its square root is $\hat{\sigma}(\Delta) = .35$, which is interpretable as an estimate of the "absolute" standard error of measurement. Consequently, with the usual caveats, $X_{pTR} \pm .35$ constitutes a 68% confidence interval for persons' universe scores.

Suppose Smith judged $\hat{\sigma}(\Delta) = .35$ to be unacceptably large for her purposes, or suppose she decided that time constraints preclude using three prompts. For either of these reasons, or other reasons, she may want to estimate $\hat{\sigma}(\Delta)$ for a number of different values of n_t' and/or n_r'. Doing so is simple. Smith merely uses the rule following Equation Set 5 to estimate the D study variance components for any pair of D study sample sizes of interest to her. Then, as indicated in Table 2, she sums all the estimated variance components except $\hat{\sigma}^2(p)$, and takes the square root.

Figure 2 illustrates results for both n_t' and n_r' ranging from one to four. It is evident from Figure 2 that increasing n_t' and/or n_r' leads to a decrease in $\hat{\sigma}(\Delta)$. This result is sensible, because averaging over more conditions of measurement should reduce error. Figure 2 also suggests that using more than three raters leads to only a very slight reduction in $\hat{\sigma}(\Delta)$. Consequently, probably it would be unnecessary to use more than three raters (and perhaps only two) for an actual measurement procedure. In addition, Figure 2 indicates that using additional prompts decreases $\hat{\sigma}(\Delta)$ quicker than using additional raters. This is a direct result of the fact that $\hat{\sigma}^2(t) = .06$ is bigger than $\hat{\sigma}^2(r) = .02$, and $\hat{\sigma}^2(pt) = .15$ is bigger than $\hat{\sigma}^2(pr) = .04$. Consequently, for this example, all other things being equal, it would seem desirable to use as many prompts as possible.

Relative error variance, $\sigma^2(\delta)$. Relative error is defined as the difference between a person's observed deviation score and his or her universe deviation score:

$$\delta_p = (X_{pTR} - \mu_{TR}) - (\mu_p - \mu), \tag{9}$$

where μ_{TR} is the expected score over persons of the observed scores, X_{pTR}. For the $p \times T \times R$ design and an infinite universe of generalization, it can be shown that

$$\delta_p = \nu_{pT} + \nu_{pR} + \nu_{pTR}, \tag{10}$$

and the variance of these relative errors is the sum of the variance components for the three effects in Equation 10. This result is also given in Table 2, under the column headed "*T*, *R* random." Relative error variance is similar to error variance in classical theory.

For the example introduced previously, if $n_t' = 3$ and $n_r' = 2$, then

$$\hat{\sigma}^2(\delta) = .05 + .02 + .02 = .09,$$

Table 2

Estimated Random Effects Variance Components That Enter $\hat{\sigma}^2(\tau)$, $\hat{\sigma}^2(\delta)$, and $\hat{\sigma}^2(\Delta)$ for the $p \times T \times R$ Design

	T, *R* random	*T* fixed
$\hat{\sigma}(p)$	τ	τ
$\hat{\sigma}^2(T)$	Δ	
$\hat{\sigma}^2(R)$	Δ	Δ
$\hat{\sigma}^2(pT)$	Δ, δ	τ
$\hat{\sigma}^2(pR)$	Δ, δ	Δ, δ
$\hat{\sigma}^2(TR)$	Δ	Δ
$\hat{\sigma}^2(pTR)$	Δ, δ	Δ, δ

Note: τ is universe score.

and its square root is $\hat{\sigma}(\delta) = .30$, which is interpretable as an estimate of the "relative" standard error of measurement. Note that this value of $\hat{\sigma}(\delta)$ is smaller than $\hat{\sigma}(\Delta) = .35$ for the same pair of sample sizes. In general, $\hat{\sigma}(\delta)$ is less than $\hat{\sigma}(\Delta)$ because, as indicated in Table 2, $\hat{\sigma}^2(\delta)$ involves fewer variance components than $\hat{\sigma}^2(\Delta)$. In short, relative interpretations about persons' scores are less error prone than absolute interpretations.

Coefficients

Two types of reliability-like coefficients are available in generalizability theory. One coefficient is called a "generalizability coefficient" and denoted here as ρ^2. The other coefficient is an "index of dependability" that is denoted Φ.

Generalizability coefficient, ρ^2. A generalizability coefficient is defined as

$$\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)}. \tag{11}$$

It is the analogue of a reliability coefficient in classical theory. For the example considered here, with $n_t' = 3$ and $n_r' = 2$,

$$\hat{\rho}^2 = .25/[.25 + .09] = .74.$$

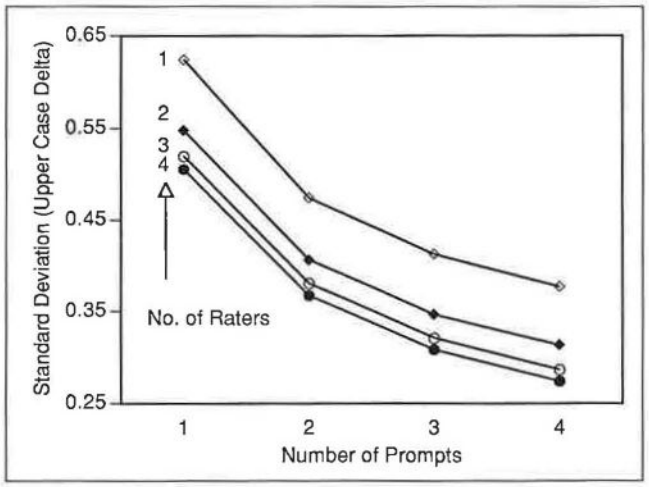

FIGURE 2. *$\hat{\sigma}(\Delta)$ for the $p \times T \times R$ design and an infinite universe of generalization, with the number of prompts and the number of raters ranging from one to four*

Figure 3 provides a graph of $\hat{\rho}^2$ for values of n'_t and n'_r ranging from one to four. As observed in the discussion of Figure 2, little is gained by having more than three raters, and using a relatively large number of prompts seems highly desirable.

Index of dependability, Φ. An index of dependability is defined as:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)}. \qquad (12)$$

Φ differs from ρ^2 in that Φ involves $\sigma^2(\Delta)$, whereas ρ^2 involves $\sigma^2(\delta)$. Consequently, Φ is generally less than ρ^2. The index Φ is appropriate when scores are given "absolute" interpretations, as in domain-referenced or criterion-referenced situations. For the example considered here, with $n'_t = 3$ and $n'_r = 2$,

$$\hat{\Phi} = .25/[.25 + .12] = .68.$$

D Study Considerations for Different Designs and/or Universes of Generalization

The previous section assumed that the D study employed a $p \times T \times R$ design and the universe of generalization was infinite, consisting of two random facets, T and R. Recall that the G study also employed a fully crossed design ($p \times t \times r$) for an infinite universe of admissible observations. In short, to this point, it has been assumed that both designs are fully crossed and the size or "extent" of both universes is essentially the same. This need not be the case, however. For example, the universe of generalization may be narrower than the universe of admissible observations. Also, the structure of the D study can be different from that employed to estimate variance components in the G study. Generalizability theory does not merely permit such differences—it effectively encourages investigators to give serious consideration to the consequences of employing different D study designs and to assumptions about a universe of generalization. This is illustrated below using two examples.

The $p \times T \times R$ Design With a Fixed Facet

Returning to the previously introduced scenario, suppose another investigator, Sam Jones, has access to Smith's G study estimated variance components in Equation Set 3. However, Jones is not interested in generalizing over essay prompts. Rather, if he were to replicate his measurement procedure, he would use different raters but the *same* prompts. If so, we would say that Jones' universe of generalization is "restricted" in that it contains a *fixed facet, T*. Consequently, Jones's universe of generalization is narrower than Smith's infinite universe of generalization. (In ANOVA terminology, the context here is essentially that of a mixed model.)

Suppose, also, that Jones decides to use the same D study design structure as Smith: namely, the $p \times T \times R$ design. Under these circumstances, the last column of Table 2 indicates which of the estimated random effects D study variance components need to be summed to obtain estimated universe score variance, $\hat{\sigma}^2(\tau)$, as well as $\hat{\sigma}^2(\Delta)$ and $\hat{\sigma}^2(\delta)$.

For example, if $n_t = n'_t = 3$ and $n'_r = 2$, then the estimated random effects D study variance components are given by Equation Set 5, and using the last column in Table 2

$$\begin{aligned}
\hat{\sigma}^2(\tau) &= \hat{\sigma}^2(p) + \hat{\sigma}^2(pT) \\
&= .25 + .05 = .30, \\
\hat{\sigma}^2(\Delta) &= \hat{\sigma}^2(R) + \hat{\sigma}^2(pR) + \hat{\sigma}^2(TR) + \hat{\sigma}^2(pTR) \\
&= .01 + .02 + .00 + .02 = .05, \text{ and} \\
\hat{\sigma}^2(\delta) &= \hat{\sigma}^2(pR) + \hat{\sigma}^2(pTR) \\
&= .02 + .02 = .04.
\end{aligned}$$

It is particularly important to note that, with prompts fixed, $\sigma^2(pT)$ contributes to universe score variance, not either error variance. Consequently, for a restricted universe of generalization with T fixed, universe score variance is larger than it is for an infinite universe of generalization in which both T and R are random.

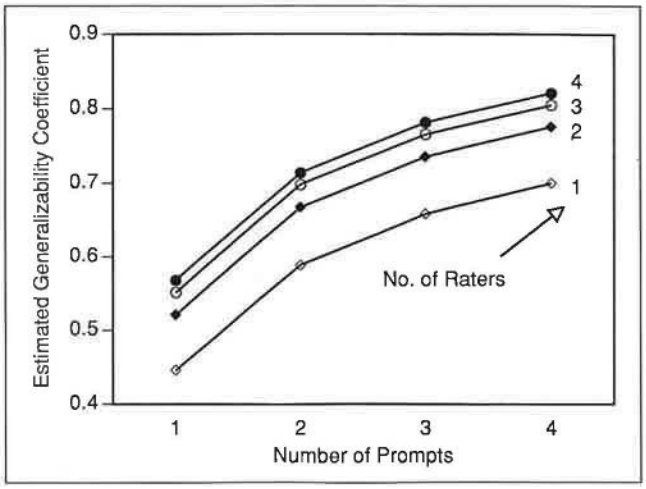

FIGURE 3. *$\hat{\rho}^2$ for the $p \times T \times R$ design and an infinite universe of generalization, with the number of prompts and the number of raters ranging from one to four*

Given these results, it follows from Equation 11 that

$$\hat{\rho}^2 = .30/[.30 + .04] = .88.$$

Recall that, for these sample sizes ($n'_t = 3$ and $n'_r = 2$), when prompts were considered random, Smith obtained $\hat{\rho}^2 = .74$. The estimated generalizability coefficient $\hat{\rho}^2$ is larger when prompts are considered fixed because a universe of generalization with a fixed facet is narrower than a universe of generalization with both facets random. That is, generalizations to narrow universes are less error prone than generalizations to broader universes. It is important to note, however, this does not necessarily mean that narrow universes are to be preferred, because restricting a universe also restricts the extent to which an investigator can generalize. For example, when prompts are considered fixed, an investigator cannot logically draw inferences about what would happen if different prompts were used.

The D Study $p \times (R{:}T)$ Design

To expand our scenario even further, consider a third investigator, Ann Hall, who decides that practical constraints preclude her from having all raters evaluate all responses of all persons to all prompts. Rather, she decides that, for each prompt, a different set of raters will evaluate persons' responses. This is a verbal description of the D study $p \times (R{:}T)$ design, where ":" is read "nested within." Figure 4 provides a Venn diagram representation of this design. In this Venn diagram, the nesting of R within T is represented by the inclusion of one entire circle within another circle.

As suggested by the five distinct areas in Figure 4, for the $p \times (R{:}T)$ design, the total variance is the sum of five independent variance components, i.e.,

$$\begin{aligned}
\sigma^2(X_{pR:T}) = \sigma^2(p) + \sigma^2(T) + \sigma^2(R{:}T) \\
+ \sigma^2(pT) + \sigma^2(pR{:}T). \qquad (13)
\end{aligned}$$

For a random effects model, these variance components can be estimated using Smith's estimated G study variance components, even though Smith's G study design is fully crossed, whereas Hall's D study design is partially nested. The process of doing so involves two steps.

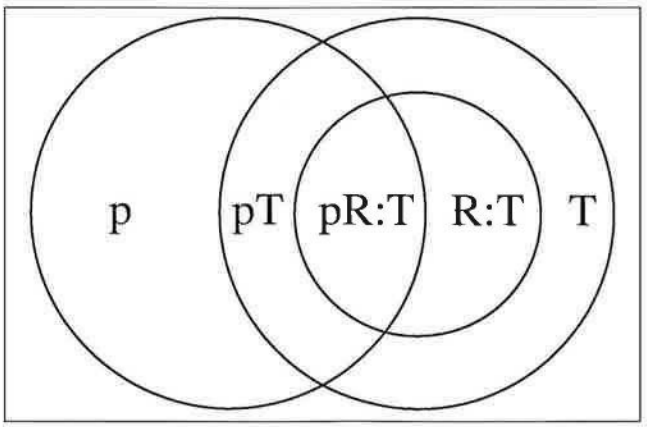

FIGURE 4. *Venn diagram for p × (R:T) design*

First, the G study variance components for the $p \times (r{:}t)$ design are estimated using the results in Equation Set 3 for the $p \times t \times r$ design. For both designs, $\hat{\sigma}^2(p) = .25$, $\hat{\sigma}^2(t) = .06$, and $\hat{\sigma}^2(pt) = .15$, and it can be shown that estimates of the remaining two G study variance components for the $p \times (r{:}t)$ design are:

$$\hat{\sigma}^2(r{:}t) = \hat{\sigma}^2(r) + \hat{\sigma}^2(tr) \quad \text{and} \tag{14}$$

$$\hat{\sigma}^2(pr{:}t) = \hat{\sigma}^2(pr) + \hat{\sigma}^2(ptr). \tag{15}$$

Using Equations 14 and 15,

$$\hat{\sigma}^2(r{:}t) = .02 + .00 = .02 \quad \text{and}$$

$$\hat{\sigma}^2(pr{:}t) = .04 + .12 = .16.$$

Second, the rule following Equation Set 5 is applied to the estimated G study variance components for the $p \times (r{:}t)$ design. Assuming $n'_t = 3$ and $n'_r = 2$, the results are:

$$\hat{\sigma}^2(p) = .25, \quad \hat{\sigma}^2(T) = .02, \quad \hat{\sigma}^2(pT) = .05,$$

$$\hat{\sigma}^2(R{:}T) = .003 \quad \text{and} \quad \hat{\sigma}^2(pR{:}T) = .027. \tag{16}$$

The second column in Table 3 specifies how to combine the estimates in Equation Set 16 to obtain $\hat{\sigma}^2(\tau)$, $\hat{\sigma}^2(\Delta)$, and $\hat{\sigma}^2(\delta)$ for an infinite universe of generalization in which both T and R are random. The third column applies when prompts are fixed. Once $\hat{\sigma}^2(\tau)$, $\hat{\sigma}^2(\Delta)$, and $\hat{\sigma}^2(\delta)$ are obtained, $\hat{\rho}^2$ and $\hat{\Phi}$ can be obtained using Equations 11 and 12, respectively.

Suppose, for example, that Hall decides to generalize to a universe in which both T and R are considered random. For this universe of generalization, given the results in Equation Set 16, and using Table 3,

$$\hat{\sigma}^2(\tau) = \hat{\sigma}^2(p) = .25,$$

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(T) + \hat{\sigma}^2(pT) + \hat{\sigma}^2(R{:}T) + \hat{\sigma}^2(pR{:}T)$$

$$= .02 + .05 + .003 + .027 = .10, \text{ and}$$

$$\hat{\sigma}^2(\delta) = \hat{\sigma}^2(pT) + \hat{\sigma}^2(pR{:}T)$$

$$= .05 + .027 = .077.$$

It follows that

$$\hat{\rho}^2 = \hat{\sigma}^2(\tau)/[\hat{\sigma}^2(\tau) + \hat{\sigma}^2(\delta)]$$

$$= .25/[.25 + .077] = .76.$$

Recall that for the $p \times T \times R$ design with T and R random, Smith obtained $\hat{\rho}^2 = .74$, which is somewhat different from $\hat{\rho}^2 = .76$ for the *same* universe using the $p \times (R{:}T)$ design. The difference in these two results is *not* rounding error. Rather, it is attributable to the fact that $\hat{\sigma}^2(\delta) = .09$ for the $p \times T \times R$ design is larger than $\hat{\sigma}^2(\delta) = .077$ for the $p \times (R{:}T)$ design. This illustrates that reliability, or generalizability, is affected by design structure. Recall that it has been demonstrated previously that reliability, or generalizability, is also affected by sample sizes and the size or "extent" of a universe of generalization. These results illustrate an important fact: namely, it can be very misleading to refer to *the* reliability or *the* error variance of a measurement procedure without considerable explanation and qualification.

Other Issues

All other things being equal, the power of generalizability theory is most likely to be realized when a G study employs a fully crossed design and a large sample of conditions for each facet in the universe of admissible observations. A large sample of conditions is beneficial because it leads to more stable estimates of G study variance components. A crossed design is advantageous because it maximizes the number of design structures that can be considered for one or more subsequent D studies.

It is important to note, however, that any design structure can be used in a G study. For example, the scenario discussed previously could have used a G study $p \times (r{:}t)$ design. However, under these circumstances an investigator could not estimate results for a D study $p \times T \times R$ design. This limitation occurs because independent estimates of $\sigma^2(r)$ and $\sigma^2(tr)$ are needed for the D study, but they are completely confounded in the $\sigma^2(r{:}t)$ G study variance component, and independent estimates of $\sigma^2(pr)$ and $\sigma^2(ptr)$ are completely confounded in $\sigma^2(pr{:}t)$.

It often happens that the distinction between a G and D study is blurred, usually because the only available data are for an operational administration of an actual measurement procedure. In this case, the methods discussed can still be followed to estimate parameters such as error variances and generalizability coefficients, but obviously under these circumstances an investigator cannot take advantage of all aspects of generalizability theory.

In most applications of generalizability theory, examinees or persons are the objects of measurement. Occasionally, however, some other collection of conditions plays the role of objects of measurement. For example, in evaluation studies, classes are often the objects of measurement with persons and other facets being associated with the universe of generalization. It is straightforward to apply generalizability theory in such cases.

Summary

Classical test theory and ANOVA can be viewed as the parents of generalizability theory in the sense that generalizability theory employs ANOVA procedures with models that are

Table 3
Estimated Random Effects Variance Components That Enter $\hat{\sigma}^2(\tau)$, $\hat{\sigma}^2(\delta)$, and $\hat{\sigma}^2(\Delta)$ for the $p \times (R{:}T)$ Design

	T, R random	*R* fixed
$\hat{\sigma}^2(p)$	τ	τ
$\hat{\sigma}^2(T)$	Δ	
$\hat{\sigma}^2(R{:}T)$	Δ	Δ
$\hat{\sigma}^2(pT)$	Δ, δ	τ
$\hat{\sigma}^2(pR{:}T)$	Δ, δ	Δ, δ

Note. τ is universe score.

extensions of the model used in classical theory. However, generalizability theory is *not* simply a conjunction of classical theory and ANOVA. For example, classical theory has an undifferentiated error term, whereas the models and methods used in generalizability theory allow an investigator to systematically distinguish among multiple sources of error. Also, generalizability theory emphasizes the estimation of variance components, rather than F-tests, which predominate in most experimental design and ANOVA texts. Further, generalizability theory has a conceptual framework that is not part of either classical theory or ANOVA.

A generalizability analysis begins with the specification of a universe of admissible observations. A G study is employed to estimate variance components for this universe and a relevant population. These G study estimated variance components are used to estimate results (error variances, generalizability coefficients, etc.) for one or more D studies associated with a prespecified universe of generalization. D studies may differ in terms of sample sizes and/or design structure. Specifying a universe of generalization requires identifying which facets are random and which are fixed.

Generalizability theory is a very broadly defined measurement model, and different applications of generalizability theory tend to involve somewhat different mixes of conceptual and statistical concerns. Consequently, conducting a generalizability analysis is often a nontrivial exercise. The process of doing so, however, reveals the importance and consequences of various sources of measurement error, and aids an investigator in better understanding measurement itself.

Self-Test

1. Suppose an investigator obtains the following mean squares for a G study $p \times t \times r$ design using $n_p = 100$ persons, $n_t = 5$ essay items, and $n_r = 6$ raters:

   $MS(p) = 6.20$ $MS(t) = 57.60$, $MS(r) = 28.26$,
   $MS(pt) = 1.60$, $MS(pr) = .26$, $MS(tr) = 8.16$,
   and $MS(ptr) = .16$.

   a. Estimate the G study variance components assuming both t and r are infinite in the universe of admissible observations.
   b. Estimate the D study random effects variance components for a D study $p \times T \times R$ design with $n'_t = 4$, $n'_r = 2$, and persons as the objects of measurement.
   c. For the D study design and sample sizes in 1b, above, estimate absolute error variance, relative error variance, a generalizability coefficient, and an index of dependability.
   d. Estimate $\sigma(\Delta)$ if an investigator decides to use the D study $p \times (R{:}T)$ design with $n'_t = 3$ and $n'_r = 2$, assuming T and R are both random in the universe of generalization.
2. Suppose an investigator specifies that a universe of generalization consists of only two facets and both are fixed. From the perspective of generalizability theory, why is this nonsensical?
3. Brennan (1992, p. 65) states that "the Spearman-Brown Formula does *not* apply when one generalizes over *more* than one facet."
   a. Illustrate this fact using the example in the instructional module for the $p \times T \times R$ design with T and R random, i.e., $\hat{\rho}^2 = .74$ with three prompts and two raters. Assume the number of prompts is doubled, in which case the Spearman-Brown Formula is 2 * rel/(1 + rel), where "rel" is reliability.
   b. Explain why the two procedures give different results.

Answers to Self-Test

1. a. Using the formulas in Table 1

$$\hat{\sigma}^2(p) = [MS(p) - MS(pt) - MS(pr) + MS(ptr)]/n_t n_r$$
$$= (6.20 - 1.60 - .26 + .16)/(5 \times 6)$$
$$= .15,$$

$\hat{\sigma}^2(t) = .08$, $\hat{\sigma}^2(r) = .04$, $\hat{\sigma}^2(pt) = .24$

$\hat{\sigma}^2(pr) = .02$, $\hat{\sigma}^2(tr) = .08$, and $\hat{\sigma}^2(ptr) = .16$

   b. Because persons are the objects of measurement, $\hat{\sigma}^2(p) = .15$ is unchanged, and using the rule following Equation Set 5:

$\hat{\sigma}^2(T) = .02$, $\hat{\sigma}^2(R) = .02$, $\hat{\sigma}^2(pT) = .06$

$\hat{\sigma}^2(pR) = .01$, $\hat{\sigma}^2(TR) = .01$, and $\hat{\sigma}^2(pTR) = .02$.

   c. Absolute error variance: from the "T, R random" column of Table 2,

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(T) + \hat{\sigma}^2(R) + \hat{\sigma}^2(pT) + \hat{\sigma}^2(pR) + \hat{\sigma}^2(TR) + \hat{\sigma}^2(pTR)$$
$$= .02 + .02 + .06 + .01 + .01 + .02 = .14$$

   Relative error variance: from the "T, R random" column of Table 2,

$$\hat{\sigma}^2(\delta) = \hat{\sigma}^2(pT) + \hat{\sigma}^2(\mathrm{pR}) + \hat{\sigma}^2(pTR)$$
$$= .06 + .01 + .02 + .09$$

   Generalizability coefficient: from Equation 11, because $\hat{\sigma}^2(\tau) = \hat{\sigma}^2(p)$ when T and R are both random,

$$\hat{\rho}^2 = \hat{\sigma}^2(p)/[\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)]$$
$$= .15/(.15 + .09) = .63$$

   Index of dependability: from Equation 12, because $\hat{\sigma}^2(\tau) = \hat{\sigma}^2(p)$ when T and R are both random,

$$\hat{\Phi} = \hat{\sigma}^2(p)/[\hat{\sigma}^2(p) + \hat{\sigma}^2(\Delta)]$$
$$= .15/(.15 + .14) = .52.$$

   d. First, we need to estimate G study variance components for the $p \times (r{:}t)$ design given the results in 1a for the $p \times t \times r$ design. Under these circumstances, $\hat{\sigma}^2(p) = .15$, $\hat{\sigma}^2(t) = .08$, and $\hat{\sigma}^2(pt) = .24$ are unchanged. Using Equation 14,

$$\hat{\sigma}^2(r{:}t) = \hat{\sigma}^2(r) + \hat{\sigma}^2(tr) = .04 + .08 = .12,$$

   and using Equation 15,

$$\hat{\sigma}^2(pr{:}t) = \hat{\sigma}^2(pr) + \hat{\sigma}^2(ptr) = .02 + .16 = .18.$$

   Second, using the rule following Equation Set 5, for $n'_t = 3$ and $n'_r = 2$, the estimated random effects D study variance components are

$\hat{\sigma}^2(p) = .15$, $\hat{\sigma}^2(T) = .027$, $\hat{\sigma}^2(R{:}T) = .02$,

$\hat{\sigma}^2(pT) = .08$ and $\hat{\sigma}^2(pR{:}T) = .03$.

   Third, because T and R are random, we use the next to the last column in Table 2 to obtain

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(T) + \hat{\sigma}^2(R{:}T) + \hat{\sigma}^2(pT) + \hat{\sigma}^2(pR{:}T)$$
$$= .027 + .02 + .08 + .03 = .157.$$

   The square root is $\hat{\sigma}(\Delta) = .40$.
2. If there are only two facets and both facets are considered fixed, then every instance of a measurement

procedure would involve the same conditions. Under these circumstances, there is no generalization to a broader universe of conditions of measurement. Logically, therefore, all error variances are zero, by definition. No measurement procedure is *that* precise! To avoid this problem, at least one of the facets in a universe of generalization must be viewed as variable across instances of the measurement procedure.

3. a. According to the Spearman-Brown Formula, reliability for a measurement procedure with twice the number of prompts (i.e., with six prompts) is $2(.74)/(1 + .74) = .85$. According to generalizability theory, however,

$$\begin{aligned}\hat{\sigma}^2(\delta) &= \hat{\sigma}^2(pT) + \hat{\sigma}^2(pR) + \hat{\sigma}^2(pTR) \\ &= \frac{\hat{\sigma}^2(pt)}{n_t'} + \frac{\hat{\sigma}^2(pr)}{n_r'} + \frac{\hat{\sigma}^2(ptr)}{n_t' n_r'} \\ &= .15/6 + .04/2 + .12/12 \\ &= .055,\end{aligned}$$

and it follows that

$$\begin{aligned}\hat{\rho}^2 &= \hat{\sigma}^2(p)/[\hat{\sigma}^2(p) + \hat{\sigma}^2(\delta)] \\ &= .25/[.25 + .055] \\ &= .82.\end{aligned}$$

b. The explanation for this difference is that the term $\hat{\sigma}^2(pR) = \hat{\sigma}^2(pr)/n_r'$ in $\hat{\sigma}^2(\delta)$ is unaffected by doubling the number of prompts, whereas the Spearman-Brown procedure effectively divides $\hat{\sigma}^2(pr)/n_r'$ by two. This is an illustration of the fact that the error term is undifferentiated in classical theory, whereas generalizability theory can take into account the relative contributions of different numbers of prompts and raters to error variance.

Annotated References

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing.

This monograph is less complicated and less detailed than Cronbach, Gleser, Nanda, & Rajaratnam (1972). However, Brennan (1992) is still comprehensive enough to cover most of the major topics in generalizability theory, and it is coordinated with a computer program for generalizability analyses (see Crick & Brennan, 1983).

Crick, J. E., & Brennan, R. L. (1983). *GENOVA: A generalized analysis of variance system*. Iowa City, IA: American College Testing.

GENOVA is an ANSI Fortran V computer program for performing generalizability analyses. It is available for main frames, PCs, and Macintosh computers. The conventions and statistical procedures employed in GENOVA are those treated in Brennan (1992).

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

This book was the first comprehensive treatment of generalizability theory. Although it is 20 years old, Cronbach et al. (1972) is still the most authoritative and definitive treatment of generalizability theory. It is, however, very detailed and generally too difficult for the generalizability theory "beginner."

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), pp. 105–146. New York: Macmillan.

This chapter is the most extensive, recent treatment of reliability in educational measurement. Approximately one third of it is devoted to generalizability theory.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160.

This paper addresses a number of validity issues from the perspective of generalizability theory. In particular, by distinguishing between a universe of generalization and what he calls a "universe of allowable observations," Kane (1982) treats both reliability and validity issues associated with standardized measurement procedures.

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

As the title suggests, this reference is a monograph-length primer on generalizability theory. As such, it is less detailed than Brennan (1992) and less comprehensive than Cronbach et al. (1972). Shavelson and Webb (1992) cover most of the important issues in generalizability theory in a relatively simple manner.

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

This journal article provides a very readable and reasonably comprehensive overview of generalizability theory. After studying this ITEMS module, a generalizability theory "beginner" would be well advised to read Shavelson et al. (1989) as a subsequent step that does not require too much time.

Statement of Ownership, Management and Circulation
(Required by 39 U.S.C. 3685)

1A. Title of Publication: Educational Measurement: Issues and Practice	1B. PUBLICATION NO.: 6 8 0 8 9 0	2. Date of Filing: October 5, 1992
3. Frequency of Issue: Quarterly	3A. No. of Issues Published Annually: Four	3B. Annual Subscription Price: $10/$20/$25

4. Complete Mailing Address of Known Office of Publication (Street, City, County, State and ZIP+4 Code) (Not printers): 1230 17th Street, NW, Washington, DC 20036-3078

5. Complete Mailing Address of the Headquarters of General Business Offices of the Publisher (Not printer): 1230 17th Street, NW, Washington, DC 20036-3078

6. Full Names and Complete Mailing Address of Publisher, Editor, and Managing Editor (This item MUST NOT be blank)

Publisher (Name and Complete Mailing Address): National Council on Measurement in Education, 1230 17th Street, NW, Washington, DC 20036-3078

Editor (Name and Complete Mailing Address): John Fremer, Mail Stop 07-E, Educational Testing Service, Princeton, NJ, 08541

Managing Editor (Name and Complete Mailing Address): Thomas Campbell, National Council on Measurement in Education, 1230 17th Street, NW, Washington, DC 20036-3078

7. Owner (If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated.) (Item must be completed.)

Full Name	Complete Mailing Address
National Council on Measurement in Education	1230 17th Street, NW Washington, DC 20036-3078

8. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages or Other Securities (If there are none, so state)

Full Name	Complete Mailing Address
None	

9. For Completion by Nonprofit Organizations Authorized To Mail at Special Rates (DMM Section 424.12 only)
The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes (Check one)
(1) [X] Has Not Changed During Preceding 12 Months (2) [ ] Has Changed During Preceding 12 Months (If changed, publisher must submit explanation of change with this statement.)

10. Extent and Nature of Circulation (See instructions on reverse side)	Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
A. Total No. Copies (Net Press Run)	2,867	2,815
B. Paid and/or Requested Circulation 1. Sales through dealers and carriers, street vendors and counter sales	0	0
2. Mail Subscription (Paid and/or requested)	2,418	2,398
C. Total Paid and/or Requested Circulation (Sum of 10B1 and 10B2)	2,418	2,398
D. Free Distribution by Mail, Carrier or Other Means Samples, Complementary, and Other Free Copies	12	12
E. Total Distribution (Sum of C and D)	2,430	2,410
F. Copies Not Distributed 1. Office use, left over, unaccounted, spoiled after printing	437	408
2. Return from News Agents	0	0
G. TOTAL (Sum of E, F1 and 2—should equal net press run shown in A)	2,867	2,818

11. I certify that the statements made by me above are correct and complete. Signature and Title of Editor, Publisher, Business Manager, or Owner: Susan L. Maitland Director of Publications

PS Form 3526, January 1991 (See instructions on reverse)