Performance Assessments From the Perspective of Generalizability Theory

Robert L. Brennan
University of Iowa

Relevant aspects of generalizability theory relating to performance assessments (PAs) are reviewed. The role of various facets in assessing the generalizability of PAs is examined, and some popular estimates of reliability for PAs are considered from the perspective of generalizability theory.

Index terms: generalizability theory, occasions, performance assessments, raters, reliability, tasks.

Historically, reliability issues in psychology and education have been addressed principally using classical test theory (CTT), which postulates that an observed score can be decomposed into a "true" score and a single, undifferentiated random error term, E (Feldt & Brennan, 1989). Generalizability (G) theory liberalizes CTT 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 analysis of variance (ANOVA) methods.

The purposes of this paper are: (1) to provide an overview of those aspects of the conceptual framework and methods of G theory that are particularly relevant to performance assessments (PAs), (2) to evaluate some of the more prevalent estimates of reliability for PAs from the perspective of G theory, and (3) to discuss the role of various facets in assessing the generalizability of PAs.

Basic Concepts of G Theory

CTT and ANOVA can be viewed as the parents of G theory. However, G theory has a unique conceptual framework. Among the concepts in this framework are universes of admissible observations and G studies, as well as universes of generalization and decision (D) studies.

In-depth explications of the concepts and methods of G theory have been provided by Cronbach, Gleser, Nanda, & Rajaratnam (1972) and Brennan (in press). Brennan (1992a) provided a somewhat less-detailed treatment. Overviews of essential features of G theory were provided by Feldt & Brennan (1989) and Shavelson & Webb (1991). An introduction was provided by Brennan (1992b), a history by Brennan (1997), and a treatment of misconceptions about the theory by Brennan (2000). Brennan & Johnson (1995); Cronbach, Linn, Brennan, & Haertel (1997); and Gao, Shavelson, & Baxter (1994) covered topics that, in part, overlap those treated here.

The concepts and methods of G theory are explained and illustrated using an example from the PA literature reported by Shavelson, Baxter, & Gao (1993):

> The California Assessment Program (CAP) conducted a voluntary statewide science assessment in 1989-1990 .... Students were posed five independent tasks. More specifically, students rotated through a series of five self-contained stations at timed intervals (about 15 mins.). At one station, students were asked to complete a problem solving task (determine which of these materials may serve as a conductor). At the next station, students were asked

> to develop a classification system for leaves and then to explain any adjustments necessary to include a new mystery leaf in the system. At yet another, students were asked to conduct tests with rocks and then use the results to determine the identity of an unknown rock. At the fourth station, students were asked to estimate and measure various characteristics of water (e.g., temperature, volume). And at the fifth station, students were asked to conduct a series of tests on samples of lake water to discover why fish are dying (e.g., is the water too acidic?). At each station, students were provided with the necessary materials and asked to respond to a series of questions in a specified format (e.g., fill in a table).
>
> A predetermined scoring rubric developed by teams of teachers in California was used to evaluate the quality of students' written responses (California State Department of Education, 1990) to each of the tasks. Each rubric was used to score performance on a scale from 0 to 4 (0 = no attempt, 1 = serious flaws, 2 = satisfactory, 3 = competent, 4 = outstanding). All tasks were scored by three raters. (p. 222)

Universe of Admissible Observations and G Study Considerations

For the CAP example, the universe of admissible observations (UAO) consisted of two facets: tasks (t) and raters (r). Because, in principle, any task could be evaluated by any rater, these facets were crossed in the UAO, and this crossing is denoted $t \times r$. Persons (p) or students were not viewed as part of the UAO; rather, they constituted the population.

As reported by Shavelson et al. (1993), the G study design for the CAP example consisted of taking a sample of five tasks from the UAO, administering them to a sample of persons, and then having three raters evaluate all products/results produced by all persons. This constitutes a fully crossed G study $p \times t \times r$ design. For this example, the G study is a random effects G study because the authors assumed that the potential set of tasks and raters in the UAO were both indefinitely large, with the actual tasks and raters viewed as samples from the UAO.

G study variance components (VCs). The principal results of a G study are estimated VCs for each of the effects in its design. These estimates are obtained using ANOVA procedures. For the CAP example, the estimated VCs are reported in Table 1. For example, the estimated VC for persons is $\hat{\sigma}^2(p) = .298$, and the estimated VC for the interaction of persons and tasks is $\hat{\sigma}^2(pt) = .493$.

Table 1
CAP Generalizability Analyses

	D Study Estimated Variance Components		
G Study Estimated Variance Components		$n'_t = 5$ $n'_r = 3$	$n'_t = 10$ $n'_r = 1$
$\hat{\sigma}^2(p) = .298$	$\hat{\sigma}^2(p)$	.298	.298
$\hat{\sigma}^2(t) = .092$	$\hat{\sigma}^2(T) = \hat{\sigma}^2(t)/n'_t$	.018	.009
$\hat{\sigma}^2(r) = .003$	$\hat{\sigma}^2(R) = \hat{\sigma}^2(r)/n'_r$	.001	.003
$\hat{\sigma}^2(pt) = .493$	$\hat{\sigma}^2(pT) = \hat{\sigma}^2(pt)/n'_t$	.099	.049
$\hat{\sigma}^2(pr) = .000$	$\hat{\sigma}^2(pR) = \hat{\sigma}^2(pr)/n'_r$	.000	.000
$\hat{\sigma}^2(tr) = .002$	$\hat{\sigma}^2(TR) = \hat{\sigma}^2(tr)/(n'_t n'_r)$	.000	.000
$\hat{\sigma}^2(ptr) = .148$	$\hat{\sigma}^2(pTR) = \hat{\sigma}^2(ptr)/(n'_t n'_r)$	.010	.015
	$\hat{\sigma}^2(\tau)$	.30	.30
	$\hat{\sigma}^2(\delta)$	.11	.06
	$\hat{\sigma}^2(\Delta)$	.13	.08
	$\boldsymbol{E}\hat{\rho}^2$	.73	.82
	$\hat{\Phi}$	.70	.80

The VC for persons can be interpreted as follows: Suppose an investigator could obtain each person's mean (or expected value) over all tasks and raters in the UAO. The variance of these scores would be $\sigma^2(p)$, which is estimated to be $\hat{\sigma}^2(p) = .298$ for the CAP data. Similarly, $\hat{\sigma}^2(r) = .003$ is the estimated variance of rater mean scores, where each mean (or expected value) is over all persons in the population and all tasks in the UAO. The estimated variance of task mean scores in the UAO is then $\hat{\sigma}^2(t) = .092$, which suggests that tasks differ somewhat in difficulty.

Interaction VCs are more difficult to interpret. Consider, for example, $\hat{\sigma}^2(pt)$ in the CAP example. $\hat{\sigma}^2(pt) = .493$ is considerably greater than zero, which suggests that there was a considerably different rank ordering of person mean scores for each of the various tasks in the UAO. By contrast, the fact that $\hat{\sigma}^2(pr) = 0$ means that the various raters rank ordered persons similarly. Also, $\hat{\sigma}^2(tr) = .002$ suggests that the various raters rank ordered the difficulty of the tasks similarly. The last VC, $\hat{\sigma}^2(ptr) = .148$, is a residual VC that includes the triple-order interaction and all other unexplained sources of variation.

Because $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(pr)$, and $\hat{\sigma}^2(tr)$ are all close to zero, the rater facet does not appear to contribute much to variability in observed scores. By contrast, $\hat{\sigma}^2(pt)$ is quite large, suggesting that the task facet contributed substantially to score variability.

The G study VCs provide a decomposition of the variance over p, t, and r of single person-task-rater scores:

$$\sigma^2(X_{ptr}) = \sigma^2(p) + \sigma^2(t) + \sigma^2(r) + \sigma^2(pt) + \sigma^2(pr) + \sigma^2(tr) + \sigma^2(ptr) \ , \tag{1}$$

which is usually called "total variance" in G theory literature, because it is analogous to "total" sums of squares in ANOVA. In G theory, the phrase "total variance" refers to the sum of the G study VCs. Table 1 indicates that the largest contributors to total variance are persons and person-task interactions.

Other examples. The CAP assessment results are typical of applications of G theory for many programs involving PAs. The rater facet often contributes relatively little to total variance for single person-task-rater observations, as evidenced by the fact that $\hat{\sigma}^2(r)$, $\hat{\sigma}^2(pr)$, and $\hat{\sigma}^2(tr)$ are frequently quite small. $\hat{\sigma}^2(t)$ is often small as well, suggesting that tasks are quite similar in average difficulty for many programs. As discussed more fully by Brennan (1996), other published studies using the $p \times t \times r$ design gave similar results, including science and math assessments discussed by Shavelson et al. (1993), and tests of listening and writing discussed by Brennan, Gao, & Colton (1995).

For simplicity, the $p \times t \times r$ design is the primary design used here. For this design, each rater evaluates all tasks performed by all persons. However, many published studies of PAs use designs that are more complicated in the sense that they involve nesting (and sometimes matrix sampling). Often, for example, different raters evaluate each task. One such design is denoted $p\times$ $(r{:}t)$. It has five VCs: $\sigma^2(p)$, $\sigma^2(t)$, $\sigma^2(r{:}t)$, $\sigma^2(pt)$, and $\sigma^2(pr{:}t)$. Analyses of the EXPLORE Program (ACT, 1994) and the Iowa Writing Assessment (see Brennan, 1998, pp. 324–325; Hoover & Bray, 1995) have used this design.

For almost all PAs reported in the literature, $\hat{\sigma}^2(pt)$ tends to be quite large, suggesting that there is only a limited degree of across-task generalizability. This conclusion has been widely reported [e.g., Dunbar, Koretz, & Hoover (1991); Lane, Liu, Ankenmann, & Stone (1996); Linn (1993); Linn & Burton (1994); van der Vleuten & Swanson (1990); Welch (1991)].

Infinite Universe of Generalization and D Study Considerations

Equation 1 states that G study VCs provide a decomposition of the total observed score variance for *single* person-task-rater scores. Let n_p, n_t, and n_r be the G study sample sizes for persons, tasks, and raters, respectively. For the CAP example, the total observed score variance is then the variance of the $n_p n_t n_r$ observed scores, each of which could be 0, 1, 2, 3, or 4. In practice, of course, decisions about persons (the objects of measurement) are not likely to be made based on persons' scores for a *single* task evaluated by a *single* rater. Rather, decisions are likely made based on average scores over multiple tasks and/or raters. A typical D study consideration is to identify values of n'_t and n'_r (which do not need to equal n_t and n_r, respectively) that result in acceptably small error variance and/or acceptably large reliability-like coefficients.

Another D study consideration is the specification of a universe of generalization (UG), which is the universe to which a decision maker wants to generalize. In this section, it will be assumed that the UG mirrors the UAO in the sense that the task and rater facets are both infinite. Strictly speaking, this means that the UG is a universe of *randomly* parallel forms of the measurement procedure, in which each such form consists of a *different* sample of tasks and a *different* sample of raters. (In a subsequent section, a restricted UG is considered that is smaller than the infinite UG.) A person's universe score is his/her mean (or expected) score over all randomly parallel forms of the measurement procedure in the UG. As such, the universe score is analogous to the true score in CTT.

An additional D study consideration is its design structure. It is assumed that the D study design mirrors the G study design in the sense that both are fully crossed—all persons respond to the same tasks, and the responses/products of all persons to all tasks are evaluated by the same raters.

D study VCs. G study VCs are used to estimate D study VCs for average scores over n'_t tasks and n'_r raters. For a random effects model, the process is simple. Let $\sigma^2(\alpha)$ be the G study VC for α [e.g., for $\sigma^2(pt)$, $\alpha = pt$], and let $\sigma^2(\overline{\alpha})$ be the corresponding D study VC. If α contains t, then $\sigma^2(\overline{\alpha}) = \sigma^2(\alpha)/n'_t$; if α contains r, then $\sigma^2(\overline{\alpha}) = \sigma^2(\alpha)/n'_r$; and if α contains t and r, then $\sigma^2(\overline{\alpha}) = \sigma^2(\alpha)/(n'_t n'_r)$. The resulting equations for the estimated D study VCs are provided in Table 1. Note that they are often designated using upper-case subscripts, to emphasize that the VCs are for mean scores.

The D study VC $\sigma^2(p)$ is the variance of persons' universe scores for the infinite UG. As such, it is called *universe score variance,* which is analogous to true score variance in CTT. D study VCs are used to estimate relative and absolute error variances, as well as two types of reliability-like coefficients called generalizability coefficients and dependability coefficients.

Absolute error variance. Absolute error is the difference between a person's observed score and universe score. The variance over persons of absolute errors is

$$\begin{aligned}\sigma^2(\Delta) &= \sigma^2(T) + \sigma^2(R) + \sigma^2(pT) + \sigma^2(pR) + \sigma^2(TR) + \sigma^2(pTR) \\ &= \frac{\sigma^2(t)}{n'_t} + \frac{\sigma^2(r)}{n'_r} + \frac{\sigma^2(pt)}{n'_t} + \frac{\sigma^2(pr)}{n'_r} + \frac{\sigma^2(tr)}{n'_t n'_r} + \frac{\sigma^2(ptr)}{n'_t n'_r} \,. \qquad (2)\end{aligned}$$

Absolute error variance is the sum of all the D study VCs except for universe score variance. For the CAP data with $n'_t = 5$ and $n'_r = 3$, Table 1 reports that $\hat{\sigma}^2(\Delta) = .13$. The square root is $\hat{\sigma}(\Delta) = .36$, which is the Δ-type, or absolute standard error of measurement (SEM). Under normality assumptions, adding and subtracting .36 to persons' observed scores over 5 tasks and 3 raters provides approximate 68% confidence intervals for their universe scores.

Relative error variance. Relative error is the difference between a person's observed deviation score and his/her universe deviation score. The variance over persons of relative errors is

$$\sigma^2(\delta) = \sigma^2(pT) + \sigma^2(pR) + \sigma^2(pTR) = \frac{\sigma^2(pt)}{n'_t} + \frac{\sigma^2(pr)}{n'_r} + \frac{\sigma^2(ptr)}{n'_t n'_r} \,. \qquad (3)$$

Relative error variance is the sum of the D study VCs that include the index p and at least one other index. The square root of relative error variance is analogous to the SEM in CTT. For the CAP data with $n'_t = 5$ and $n'_r = 3$, relative error variance is $\hat{\sigma}^2(\delta) = .11$, and the square root is $\hat{\sigma}(\delta) = .33$—which is the δ-type, or relative, SEM. Note that $\sigma^2(\delta) \leq \sigma^2(\Delta)$, because $\sigma^2(\delta)$ does not contain $\sigma^2(T)$, $\sigma^2(R)$, or $\sigma^2(TR)$.

Generalizability coefficient. A generalizability coefficient is defined as

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} , \tag{4}$$

where $\sigma^2(\tau)$ is a generic notation for universe score variance. A generalizability coefficient is the ratio of universe score variance to itself plus relative error variance. As such, a generalizability coefficient is analogous to a reliability coefficient in CTT. For the CAP example, $\sigma^2(\tau) = \sigma^2(p)$, and $\sigma^2(\delta)$ is given by Equation 3. Therefore,

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \left[\dfrac{\sigma^2(pt)}{n'_t} + \dfrac{\sigma^2(pr)}{n'_r} + \dfrac{\sigma^2(ptr)}{n'_t n'_r}\right]} . \tag{5}$$

For the CAP data with $n'_t = 5$ and $n'_r = 3$, $\boldsymbol{E}\hat{\rho}^2 = .73$ (as shown in Table 1).

Dependability coefficient. A dependability coefficient is defined as

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} . \tag{6}$$

That is, a dependability coefficient is the ratio of universe score variance to itself plus absolute error variance. For the CAP data, $\sigma^2(\tau) = \sigma^2(p)$, and $\sigma^2(\Delta)$ is given by Equation 2. Therefore,

$$\Phi = \frac{\sigma^2(p)}{\sigma^2(p) + \left[\frac{\sigma^2(t)}{n'_t} + \dfrac{\sigma^2(r)}{n'_r} + \dfrac{\sigma^2(pt)}{n'_t} + \dfrac{\sigma^2(pr)}{n'_r} + \dfrac{\sigma^2(tr)}{n'_t n'_r} + \dfrac{\sigma^2(ptr)}{n'_t n'_r}\right]} . \tag{7}$$

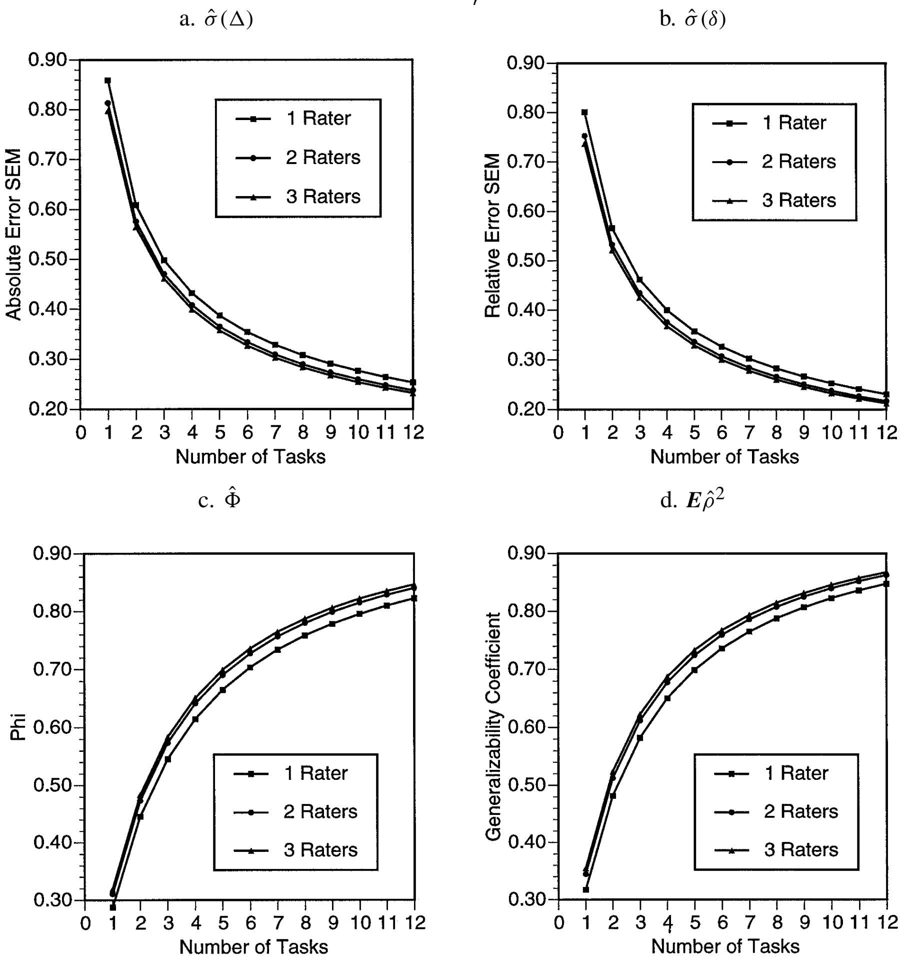
The principal difference between $\boldsymbol{E}\rho^2$ and Φ is that $\boldsymbol{E}\rho^2$ uses $\sigma^2(\delta)$ as error variance, whereas Φ uses $\sigma^2(\Delta)$. Because $\sigma^2(\Delta) \geq \sigma^2(\delta)$, it follows that $\Phi \leq \boldsymbol{E}\rho^2$. From Table 1, $\Phi = .70$ for $n'_t = 5$ and $n'_r = 3$, which is less than $\boldsymbol{E}\hat{\rho}^2 = .73$.

Different D study sample sizes. Figure 1 displays values of $\hat{\sigma}(\Delta)$, $\hat{\sigma}(\delta)$, Φ, and $\boldsymbol{E}\hat{\rho}^2$ for the CAP example with n'_t ranging from 1 to 12 and n'_r ranging from 1 to 3. These results could be employed to examine the consequences of using various numbers of tasks and raters in a D study.

Figure 1 shows that the number of raters had little influence on the magnitude of SEMs and coefficients. This is a direct result of the fact that VCs involving the rater facet are quite small, presumably because the scoring rubrics (i.e., criteria used for assigning scores to examinee responses to each task) were well defined and the raters were well trained. Consequently, there probably is no compelling psychometric reason to employ more than two raters; one rater might provide adequate results.

Figure 1 also shows that increasing the number of tasks had a major effect on lowering SEMs and increasing the values of coefficients. With one rater, nine tasks were required for $\boldsymbol{E}\hat{\rho}^2 \geq .80$, and ten tasks were needed for $\hat{\Phi} \geq .80$ (see also the last column of Table 1). Developing, administering, and scoring so many tasks obviously would not be a trivial undertaking.

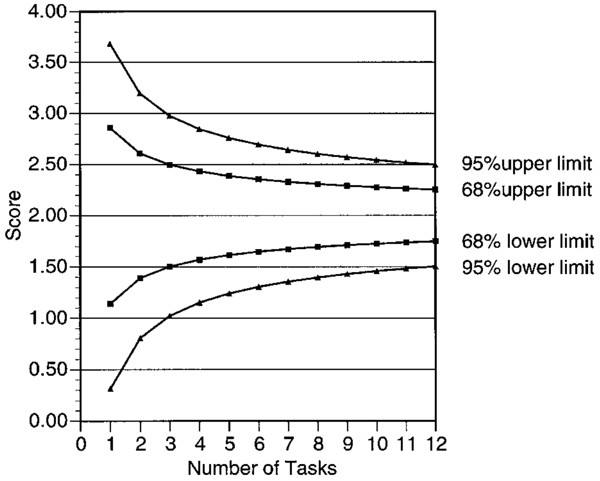
Figure 1
Generalizability Statistics for CAP Data With n'_t
From 1 to 12 and n'_r From 1 to 3


a. $\hat{\sigma}(\Delta)$
b. $\hat{\sigma}(\delta)$
c. $\hat{\Phi}$
d. $\boldsymbol{E}\hat{\rho}^2$

On the other hand, in some circumstances, an investigator might argue that it is not sensible to select n'_t on the basis of the resulting magnitude of $\boldsymbol{E}\hat{\rho}^2$ or $\hat{\Phi}$, because both depend on universe score variance. Such an investigator might be satisfied with a somewhat low value for a coefficient, provided that individual persons were accurately measured. If so, the investigator would be more interested in the magnitude of SEMs than coefficients. Recall that CAP scores for a single task ranged from 0 to 4, which means that average scores over n'_t tasks had the same range, but with fractional scores frequently occurring. Assuming normally distributed observed scores, if the investigator wanted to be 95% certain that persons' observed and universe scores differed by no more than 2 points, only 3 tasks would be required with 1 rater. However, to be 95% certain that observed and universe scores differed by no more than 1 point, 12 tasks would be required with 1 rater, and approximately 10 tasks would be required with 3 raters. [Assuming normally distributed observed scores, a 95% confidence interval covers a width of approximately 4 SEMs. Consequently, for the difference between the upper and lower limits to be 2 points, $\hat{\sigma}(\Delta) = 2/4 = .50$. For the difference to be 1 point, $\hat{\sigma}(\Delta) = 1/4 = .25$. These values can be used in Figure 1 to obtain the required numbers of tasks and raters.]

Figure 2 provides 68% and 95% Δ-type confidence intervals when $n'_r = 1$ for the CAP data for examinees with a universe score of $\tau = 2$, assuming that observed scores were normally distributed around τ. For example, when $n'_t = 3$, there is a 68% probability that examinees with true scores of 2 will obtain observed scores between 1.5 and 2.5. The limits of the confidence intervals in Figure 2 can also be used as minimum passing and maximum failing scores, as discussed by Linn & Burton (1994). For example, from Figure 2, with $n'_t = 8$, a minimum passing score of 2.6 and a maximum failing score of 1.4 are required for 95% confidence that correct decisions will be made when the standard is set at 2.

Figure 2
Confidence Intervals for CAP Data Based on $\hat{\sigma}^2(\Delta)$, Using a Single Rater and Various Numbers of Tasks, and Assuming a Universe Score of 2



Fixed facets. It has been assumed that raters and tasks are both random. This assumption is almost always preferable to treating either facet as fixed. In theory, however, tasks or raters could be treated as fixed (see Brennan, 1992b). For example, if tasks were considered fixed, then

$$\sigma^2(\tau) = \sigma^2(p) + \sigma^2(pT) \; , \tag{8}$$

$$\sigma^2(\delta) = \sigma^2(pR) + \sigma^2(pTR) \; , \tag{9}$$

and

$$\sigma^2(\Delta) = \sigma^2(R) + \sigma^2(TR) + \sigma^2(pR) + \sigma^2(pTR) \; . \tag{10}$$

As shown below, tasks are effectively fixed in some traditional treatments of the reliability of PAs, although investigators are not always aware of this fact.

Contributions of Different Facets

Raters, tasks, and other facets can influence the measurement precision of PAs. For expository purposes, each facet is discussed separately in this section. This is done to isolate issues associated with the various facets and to facilitate relating traditional results to the types of generalizability results discussed in the previous section. In practice, G analyses involving many facets simultaneously are preferred.

Raters

The previous section suggested that, for current PAs, the rater facet often does not contribute much to variability in observed scores. Although these results have been found in a wide range of current PAs, it should not be assumed that such results are necessarily inevitable. For example, Linn (1993) reported that the classic studies of Starch & Elliot (1912, 1913) on the grading of high school work in English and mathematics demonstrated a wide range of grades assigned to written essays and extended responses in geometry. Indeed, Starch and Elliot provided considerable support for the increased use of objective testing. Undoubtedly, lack of trained raters and agreed-on scoring rubrics contributed to their results.

Until recently, for any measurement procedure involving subjective scoring, the most frequently discussed measurement issue was inter-rater reliability (see, e.g., Dunbar et al., 1991). Indeed, if inter-rater reliability were high, it was often assumed that there were no other reliability issues of consequence. Although this narrow perspective no longer predominates, inter-rater reliability is still a very important issue. High inter-rater reliability is viewed by most investigators as a necessary—although not sufficient—condition for using a PA.

There are at least two general measurement perspectives on inter-rater reliability. One perspective typically involves inter-rater reliability coefficients. The other perspective considers error variances or SEMs. Both perspectives involve differences in ratings, but the two perspectives are not isomorphic.

Inter-rater reliability coefficients. In the PA literature, two general conclusions about inter-rater reliability seem to predominate. First, when tasks are the same for all students and scoring procedures are well defined, inter-rater reliability tends to be high (e.g., Dunbar et al., 1991). Second, when different students respond to different tasks, select their own tasks (e.g., select their own essay topics), or produce unique products, inter-rater reliability tends be relatively low. This latter case appears to be especially true for portfolio assessments (e.g., Koretz, Klein, McCaffrey, & Stecher, 1994). That is, when tasks are standardized, inter-rater reliability tends to be high; when tasks are not standardized, it tends to be low.

These conclusions are to be expected based on the manner in which VCs enter the standardized and nonstandardized inter-rater reliability estimates. The standardized estimate is typically obtained by correlating the ratings of two different raters to the responses of a group of persons to the same task (Brennan & Johnson, 1995). (Throughout this section, it is assumed that the same two raters are used for all persons.) In terms of the VCs previously introduced, this correlation is approximately equal to the generalizability coefficient

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p) + \sigma^2(pt)}{\sigma^2(p) + \sigma^2(pt) + \left[\sigma^2(pr) + \sigma^2(ptr)\right]} \ . \tag{11}$$

The denominator of Equation 11 is identical to the denominator of $\boldsymbol{E}\rho^2$ in Equation 5 when $n'_t = n'_r = 1$. For the case considered in this section, $n'_t = 1$ because only one task is involved in the correlation, and $n'_r = 1$ because a correlation between two raters gives an estimate of reliability for a single rater.

The numerators of the generalizability coefficients in Equations 5 and 11 differ in that Equation 11 includes not only $\sigma^2(p)$, but also $\sigma^2(pt)$. When all persons respond to the same single task, the single task is effectively hidden and fixed for all persons. Statistically, this leads to $\sigma^2(pt)$ being included with $\sigma^2(p)$ in the numerator of Equation 11. Because $\sigma^2(pt)$ is usually quite large in PA, the numerator of $\boldsymbol{E}\rho^2$ in Equation 11 is likely to be large, resulting in a relatively high value of $\boldsymbol{E}\rho^2$.

Effectively, $\sigma^2(pt)$ is part of universe score variance in Equation 11, whereas $\sigma^2(pt)$ is almost always more properly viewed as part of error variance (see Equations 2 and 3). For this reason (and another reason considered below), inter-rater reliability coefficients for standardized situations are almost always too large, relative to more appropriate estimates of generalizability for persons' scores on a PA (see Equation 5).

The nonstandardized estimate of inter-rater reliability is typically obtained by correlating two ratings of a different task or product for each person. In such cases, the design has tasks or products nested within persons and crossed with raters, $(t{:}p) \times r$, and the correlation is approximately equal to the generalizability coefficient

$$\boldsymbol{E}\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \left[\sigma^2(t{:}p) + \sigma^2(pr) + \sigma^2(tr{:}p)\right]}\ , \tag{12}$$

where
$\sigma^2(t{:}p) = \sigma^2(t) + \sigma^2(pt)$, and
$\sigma^2(tr{:}p) = \sigma^2(tr) + \sigma^2(ptr)$.

The nonstandardized coefficient in Equation 12 will be less than the standardized coefficient in Equation 11 for two reasons: (1) the numerator of the nonstandardized coefficient is smaller than the numerator of the standardized coefficient by $\sigma^2(pt)$; and (2) the denominator of the nonstandardized coefficient is larger than the denominator of the standardized coefficient by $\sigma^2(t) + \sigma^2(tr)$, which contributes to the relative error variance $\sigma^2(\delta)$ for the nonstandardized coefficient.

The VC $\sigma^2(pt)$ does not appear in the numerator of the nonstandardized coefficient in Equation 12 because tasks vary over persons, and therefore person-task interaction does not contribute to universe score variance. Rather, $\sigma^2(pt)$ is included in $\sigma^2(t{:}p)$ and contributes to relative error variance. The different role that $\sigma^2(pt)$ plays in Equations 11 and 12 is often the principal reason that the nonstandardized inter-rater reliability coefficient is smaller than the standardized coefficient. In addition, Equation 12 tends to be smaller than Equation 11 if tasks are unequal in average difficulty [i.e., $\sigma^2(t)$ is relatively large], and/or raters rank order persons differently on different tasks [i.e., $\sigma^2(pt)$ is relatively large]. Either or both of these conditions are more likely to characterize portfolio assessments than other types of PAs.

The generalizability coefficients in Equations 11 and 12 approximate inter-rater reliability coefficients often reported in the literature. Note, however, that they are appropriate for making decisions based on only one rating of only one task. This is evident from the fact that the VCs in Equations 11 and 12 containing r and/or t are all divided by 1. Frequently, inter-rater reliability coefficients are incorrectly interpreted as estimates of reliability when persons' scores are the sum or average of two ratings. Such estimates are easily obtained from Equations 11 and 12 by halving VCs that contain r'. The resulting estimates of generalizability for two ratings are necessarily larger than those for a single rating.

Conditional SEMs. In the PA literature, issues of rater reliability are most frequently discussed in terms of inter-rater reliability coefficients. However, sometimes such issues are treated (perhaps more appropriately) from the perspective of differences in the actual ratings. It is intuitively clear that such differences reflect error in some sense. SEMs such as those given by the square roots of Equations 2 and 3 capture information about such differences for a "typical" person. Alternatively, Brennan (1998) provided an extensive discussion of *conditional* SEMs that capture information about errors for *particular* persons.

As a simple example, Brennan (1998) showed that, when data are available for only two ratings and decisions are based on the mean over these two ratings, the Δ-type conditional SEM for a given person is

$$\hat{\sigma}(\Delta_p) = \frac{|X_{p1} - X_{p2}|}{2} \; . \tag{13}$$

The average over persons of these conditional SEMs is the overall Δ-type error variance.

Equation 13 is appropriate if each person responds to the same task(s)—i.e., tasks are standardized—or if persons respond to different tasks. If the two ratings are based on the same task, then the task facet is fixed and generalization is over the rater facet only. If the two ratings are based on different tasks, then generalization is over both raters and tasks (with $n'_t = 1$). In general, standard errors for generalizing over raters only are likely to be smaller than standard errors for generalizing over both tasks and raters.

Adjudication and other issues. In PAs, it is relatively common practice to obtain a third rating if two ratings differ by more than one rating-scale point. There are various ways in which this third rating might be used, but very often the net effect is that the final two ratings for a person differ by no more than a point. Under these circumstances, for $n'_r = 2$, $\hat{\sigma}(\Delta_p)$ is either .5 or 0.0. Furthermore, if k_1 is the proportion of persons with a difference score of 1, the overall value of $\hat{\sigma}(\Delta)$ will be $.5\sqrt{k_1}$, which is likely to be quite small. For example, if $k_1 = .36$, then $\hat{\sigma}(\Delta) = .30$ of a rating-scale point. (The logic here assumes that the two ratings that result after adjudication are uncorrelated. If this assumption is violated, then these illustrative results would be questionable.)

Whenever a process is used to adjudicate ratings, SEMs are likely to be relatively small; in this sense, the ratings will appear to be quite reliable. However, an adjudication process is beneficial only if it does not distort the intended construct being rated. Such distortion could occur, for example, if the adjudicators were systematically influenced by the original ratings, or if the adjudicators were experts untrained in the scoring rubric.

Carefully constructed scoring rubrics, an intensive training session for raters, and an adjudication process usually produce ratings with small error variance. However, small error variance does not guarantee that inter-rater reliability coefficients will be high. This follows from universe score variance, $\sigma^2(p)$, being in the numerator of any inter-rater reliability coefficient, whereas $\sigma^2(p)$ is absent from error variance. Consequently, if $\sigma^2(p)$ is small relative to error variance, then an inter-rater reliability coefficient could be small even if error variance is small. This illustrates that an inter-rater reliability coefficient encapsulates information about the magnitude of true differences among persons (i.e., universe score variance) relative to errors.

If true differences among persons are of no consequence—as might be the case in a fully criterion-referenced context—then inter-rater reliability coefficients might be of little value. Real-world testing contexts are seldom so clear cut, however. Error variances and inter-rater reliability coefficients capture overlapping, but still different, types of information. Therefore, it is frequently sensible to report both. In any case, it is always advisable to report estimated VCs.

Tasks

The use of well-constructed scoring rubrics with well-trained raters can substantially reduce errors attributable to raters. However, virtually all available research on PAs suggests that generalizing over tasks is an error-prone activity, no matter how well the tasks are designed.

Dunbar et al. (1991) reviewed a number of studies of direct assessments of performance, primarily in the area of writing. To compare the influence of raters and tasks on reliability, they computed average reliability due to raters and what they called "score reliability." Dunbar et al.

employed a procedure discussed by Gulliksen (1950, pp. 212–214) to estimate reliability due to scores (Gulliksen referred to this as "content reliability"). Gulliksen described his coefficient as the "reliability of an essay test corrected for attenuation due to the inaccuracy of reading" (p. 214). It can be shown that Gulliksen's content reliability is $\sigma^2(p)/[\sigma^2(p)+\sigma^2(pt)]$, which is referred to as score reliability by Dunbar et al. The Dunbar et al. reliability due to raters is approximated by Equation 11 (provided that all persons are rated by the same raters), and

$$\text{score reliability} \approx \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(pt)} . \tag{14}$$

Because Equation 14 involves $\sigma^2(pt)$, this coefficient can be referred to as "task reliability." It is equivalent to the generalizability coefficient in Equation 5 for a single task and an infinite number of raters. Another interpretation of Equation 14 is that it is a generalizability coefficient under the assumption that raters are perfectly consistent and the only source of error is person-task interactions.

Equations 11 and 14 provide some sense of the relative influence of raters and tasks on generalizability. Note that multiplying the right-hand sides of Equations 11 and 14 gives the generalizability coefficient in Equation 5 for a single task and a single rater. (See Kane, 1982, pp. 145–146, for a discussion of a similar result in terms of the reliability-validity paradox.)

For the studies they reviewed, Dunbar et al. (1991) concluded that task reliability was relatively small, suggesting again that person-task interactions are a considerable source of error. However, they also concluded that there was considerable variability in rater reliability. In particular, rater reliability tended to be lower for the older studies (see Dunbar et al., p. 293).

Often, methods can be found to increase rater reliability. However, comparable methods do not exist for increasing task reliability. Of course, tasks should be developed as carefully as possible, and if this is not done then it is likely that reliability will be adversely affected. Well-constructed tasks do not ensure high reliability, however, for two reasons. First, PAs usually include only a small number of tasks. Second, there tends to be considerable variability in examinee performance on different tasks—even for tasks in the same domain; that is, $\sigma^2(pt)$ tends to be large.

In principal, $\sigma^2(pt)$ can be reduced by narrowing the domain of tasks. For example, an investigator could define the tasks in a domain in such a way that each of them is simply a slight modification of the others. Doing so might well decrease $\sigma^2(pt)$ and, therefore, increase task reliability. However, restricting the domain of tasks in this way leads to a narrowing of the UG and, in this sense, a decrease in validity (see Kane, 1982). This is an example of the so-called "reliability-validity paradox." It is generally not advisable to take steps to increase reliability that lead to a decrease in validity.

One of the most important considerations in the development of a PA is a careful specification of the task facet in the UG. At a minimum, an investigator should be able to defend the set of tasks in a PA as a reasonable representation of the domain of tasks that might have been used. Otherwise, there is little basis for claiming that performance on the particular tasks in a PA can be generalized to a larger universe of tasks. (See Shavelson et al., 1993, p. 216, for an example of a specification of a domain of tasks.)

The importance of accurate specification of a subject-matter domain for PAs has been illustrated by Shavelson, Gao, & Baxter (1996). For the domain of elementary science, they demonstrated that an inappropriately broad specification of the domain leads to overestimating task variability and underestimating generalizability. Conversely, an inappropriately narrow specification of the domain will likely lead to underestimating task variability and overestimating generalizability. Note, however, that G theory per se does not tell an investigator how narrow or wide a domain should be. It is the investigator's responsibility to clearly specify the domain and defend that specification.

Other Facets

Shavelson et al. (1993) provided the following perspective on relevant facets for PAs:

> . . . we view a performance assessment as a sample of student performance drawn from a complex universe defined by a combination of all possible tasks, occasions, raters, and measurement methods. We view the task facet to be representative of the content in a subject-matter domain. The occasion facet includes all possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment. We view the rater facet as including all possible individuals who could be trained to score performance reliably. These three facets are, traditionally, thought of as sources of unreliability in a measurement.
>
> In addition, we incorporate a method facet into our definition of the universe of generalization. This formulation moves us beyond reliability into a sampling theory of validity (cf. Kane, 1982). Specifically, we view the method facet to be all possible methods (e.g., short answer, computer simulation) that a decision maker would be equally willing to interpret as bearing on student achievement. (p. 216)

Occasion as a facet. From a CTT perspective, sampling variability due to occasions most closely corresponds to the notion of test-retest reliability. Variability due to occasions is incorporated in traditional notions of intra-rater reliability, which reflects variability in ratings for the same raters on two occasions. Ideally, from the perspective of minimizing error variance, an investigator would prefer examinee PA products to be minimally changed over occasions during which no instruction occurred. Similarly, an investigator would prefer ratings for the same raters to be stable over occasions.

For at least two reasons, there are very few studies in the PA literature that incorporate more than one occasion. First, doing so is logistically difficult and costly. Second, in operational settings, collecting data on two occasions is usually not an intended part of the testing process. Nevertheless, it is highly desirable that at least small-scale G studies be conducted that involve occasion as a facet in order to examine the extent to which an investigator can legitimately claim that scores obtained on one occasion are generalizable to scores that might be obtained on different, but similar, occasions.

The stability of several types of science PAs was examined by Ruiz-Primo, Baxter, & Shavelson (1993); Shavelson et al. (1993); and Webb, Schlackman, & Sugrue (2000). They found that variance attributable to the interaction of persons, tasks, and occasions, $\sigma^2(pto)$ was very large—usually much larger than $\sigma^2(pt)$. However, variance attributable to persons and occasions, $\sigma^2(po)$, was quite small. This means that, over tasks, these various studies found little person-occasion interaction, but for single tasks, persons were ranked differently on different occasions. The obvious remedy for a large $\sigma^2(pto)$ is to use a large number of tasks and/or occasions in making decisions about examinees, but this might not be feasible in practice.

Usually, when occasion is considered as a facet in a PA, it is associated with the time when examinee performance occurred or products were created. However, there is actually a second occasion facet that could influence the generalizability of scores on PAs. This facet involves the occasion(s) on which the ratings are obtained. This second occasion facet would be important if judges' ratings were not stable over time. [For an example, see Wainer (1993, p. 15).] Actually, this is a very complicated situation, because changes in occasion are confounded with changes in judges (and perhaps subtle changes in rubrics or training procedures).

It would be imprudent to use the results of previous studies as a basis for general conclusions about the extent to which scores on performance tasks are invariant over occasions. However, given those results and results with other modes of testing, it would be surprising if scores on performance tasks did not exhibit at least some variability over occasions of testing and/or rating.

Method as a facet. Cronbach et al. (1972), Brennan (1992a), and Kane (1982) all observed that G theory often blurs distinctions between reliability and validity. This is indeed the case if method or mode of testing (e.g., performance tasks, multiple-choice items, short-answer questions) is incorporated as a facet in the UG.

If results are invariant over modes of testing, then there is evidence of convergent validity, and a supportable argument can be made that different modes of testing provide exchangeable information. If results are not invariant over modes of testing, then different modes provide different types of information about student performance.

Shavelson et al. (1993) examined two science performance tasks and four modes or methods (observations of actual performance, notebook reports of steps employed, computer administration of tasks, and short-answer questions). They concluded that not all methods converged, and "certain methods may measure different aspects of achievement" (p. 229). Their conclusions could generalize to other settings. This does not mean, however, that one mode is preferable to another. Such a conclusion can be drawn only through a joint consideration of psychometric properties, content, context, logistical, and cost considerations.

Scoring rubrics/procedures. Wainer (1993, p. 15) suggested that PAs yield acceptable levels of accuracy only when scoring rubrics are rigidly defined. This might be one reason why most empirical analyses of PAs effectively consider scoring rubrics as fixed in the sense that only one rubric is used. In principal, however, there might be many rubrics that could be used, although it seems unlikely that all such rubrics are equally preferable. For example, the "ideal" rubric might be so costly to implement that a simpler rubric is adopted for operational use. If so, distinctions between reliability and validity become blurred.

In the language of G theory, the issue is the extent to which an investigator can generalize over rubrics, or the extent to which examinee scores are in some sense invariant over rubrics. If scores vary depending on the rubric, then it is not likely that scores can be interpreted meaningfully without a clear understanding of the specific rubric employed. This is one reason why the interpretation of PA scores is often more demanding than for traditional modes of testing. Decision makers must understand not only what is being tested but also the standards and procedures used to assign scores.

Computerized scoring procedures for PAs are currently receiving considerable attention (e.g., Bennett, 1999; Bennett, Morley, & Quardt, 2000). It is presumed that such procedures will lead to a reduction in error attributable to using human raters. This could happen, but it is not necessarily certain; judgments are required to develop the rules or protocols for the computerized scoring algorithms. Judgments made by different experts and/or different types of experts might lead to different computerized scoring algorithms (see Clauser, Swanson, & Clyman, 1999).

Conclusion

The nature of PAs forces investigation of numerous complicated conceptual, measurement, and statistical issues. The ultimate goal of such investigations is to provide evidence concerning the validity of inferences drawn about persons' scores. An important part of that goal is to assess the extent to which scores possess an acceptable degree of consistency for the intended purpose (see, e.g., Kane, Crooks, & Cohen, 1999). G theory appears to be particularly well suited for examining such issues in that it forces the investigator to give very explicit consideration to the intended universe of generalization—in particular, which facets are fixed and which are random. Doing so often reveals important insights into seemingly simple measures of consistency.

References

ACT, Inc. (1994). *EXPLORE technical manual.* Iowa City IA: Author.

Bennett, R. E. (1999). Using new technologies to improve assessment. *Educational Measurement: Issues and Practice, 18 (3),* 5–12.

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three response types for broadening the conception of mathematical problem solving in computerized tests. *Applied Psychological Measurement, 24,* 294–309.

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

Brennan, R. L. (1992b). Generalizability theory. *Educational Measurement: Issues and Practice, 11 (4),* 27–34.

Brennan, R. L. (1996). Generalizability of performance assessments. In G. W. Phillips (Ed.), *Technical issues in performance assessments.* Washington DC: National Center for Education Statistics.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16 (4),* 14–20.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement, 22,* 301–331.

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19 (1),* 5–10.

Brennan, R. L. (in press). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys Listening and Writing tests. *Educational and Psychological Measurement, 55,* 157–176.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14 (4),* 9–12.

Clauser, B. E., Swanson, D. B., & Clyman, S. G. (1999). A comparison of the generalizability of scores produced by expert raters and automated scoring systems. *Applied Measurement in Education, 12,* 281–299.

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.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57,* 373–399.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4,* 289–303.

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

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 7,* 323–343.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hoover, H. D., & Bray, G. B. (1995, April). *The research and development phase: Can a performance assessment be cost-effective?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

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

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18 (2),* 5–17.

Koretz, D., Klein, S., McCaffrey, D., & Stecher, B. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice, 13 (3),* 5–16.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement, 33,* 71– 92.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analyses, 15,* 1–16.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13 (1),* 5–15.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30,* 41–53.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.

Shavelson, R. J., Gao, X., & Baxter, G. (1996). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 131–141). Boston: Kluwer.

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

Starch, D., & Elliot, E. C. (1912). Reliability of grading high school work in English. *School Review, 20,* 442–457.

Starch, D., & Elliot, E. C. (1913). Reliability of grading high school work in mathematics. *School Review, 21,* 254–259.

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: The state of the art. *Teaching and Learning in Medicine, 2,* 58–76.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30,* 1–21.

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13,* 277–301.

Welch, C. (1991, April). *Estimating the reliability of a direct measure of writing through generalizability theory.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Acknowledgments

Parts of this paper are a revised version of Brennan (1996). G study VCs in Table 1 were provided by Xiaohong Gao.

Author's Address

Send requests for reprints or further information to Robert L. Brennan, Iowa Testing Programs, 334A Lindquist Center, University of Iowa, Iowa City IA 52242, U.S.A. Email: robert-brennan@uiowa.edu.