Multivariate Generalizability Theory in Educational Measurement: An Empirical Study

Albert Nußbaum
RWTH Aachen, German Federal Republic

Multivariate generalizability theory was applied to the assessment of student achievement in art education. Twenty-five art students rated the paintings of 60 fourth-grade students with regard to three criteria. Paintings were made on four different topics. The results indicate that generalizability is low with respect to different raters and moderate with respect to different topics. The three ratings a rater gave on a single painting were moderately correlated. As indicated by the results for the covariance components, nearly half of the covariance between the three criteria was because the three ratings were from the same rater. Expected values for $\sigma^2(\Delta)$ are reported for different D study designs.

Some years ago, Cronbach (1976) predicted for generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) a fate similar to that of decision theory (Cronbach & Gleser, 1965):

> . . . I suggest that G theory will have a similar history. It too is going to be more important for its themes and the by-products they inspire than for its algorithms. Its formulas for estimating components of variance and covariance will see more use than the formulae of D theory, but they too have a hunger for data that often cannot be satisfied. (p. 200)

Cronbach's prophecy became only partly true. There have been statistical developments and the theoretical adaptation of generalizability theory to different measurement problems (Brennan & Kane, 1977a, 1977b; Kane & Brennan, 1980; Shavelson & Webb, 1981). In addition, a number of univariate generalizability studies have been conducted in different areas of research: (1) in research on teaching (Erlich & Borich, 1979; Erlich & Shavelson, 1978; Shavelson & Atwood-Russo, 1977; Shavelson & Dempsey-Atwood, 1976), (2) on student evaluations of instruction (Gillmore, Kane, & Naccarato, 1978; Kane & Brennan, 1977; Kane, Gillmore, & Crooks, 1976; Smith, 1979), (3) in the early stages of criterion-referenced measurement (Hively, Patterson, & Page, 1968), and (4) in some areas of psychological measurement (Endler & Hunt, 1966, 1969; Levy, 1974).

However, at least for multivariate generalizability theory, Cronbach's prophecy became true. One of the few multivariate generalizability studies is reported by Webb & Shavelson (1981).

In the first section of this article the basic concepts of generalizability theory are roughly outlined (for a comprehensive introduction to generalizability theory see Brennan, 1983). Then a multivariate generalizability study on the assessment of student achievement in art education is reported. The results

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 2, Spring 1984, pp. 219–230
© Copyright 1984 Applied Psychological Measurement Inc.
0146-6216/84/020219-12$1.85

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

are discussed with regard to their implications for validity and reliability, and conclusions are drawn concerning the optimal design for decision studies.

Generalizability Theory

"An investigator asks about the precision or reliability of a measurement because he wishes to generalize from the observation in hand to some class of observations to which it belongs" (Cronbach, Rajaratnam, & Gleser, 1963, p. 144). Even today, questions of reliability are usually treated within the framework of classical test theory. Here, the class of observations that an investigator is allowed to generalize consists of parallel measurements with equal means, equal variances, and equal covariances. However, there are many instances in educational measurement in which the equivalence requirements cannot be met. This is especially true for measures based on teacher judgments. Different teachers will vary substantially with respect to the mean and variance of their judgments. Generalizability theory, introduced by Cronbach et al. (1963) as a "liberalization of reliability theory," abandons the assumptions of equivalence and extends classical test theory to the case of nonparallel measurements. As will be shown below, its concepts are also important for the examination of the validity of a given measure.

The Universe of Admissible Observations

In generalizability theory an attribute is defined by sets of conditions under which it may be observed. Each set of conditions is called a facet. A student's level of concentration may be measured with different test forms and on different occasions. Test forms and occasions may be conceived as two different facets, each containing several conditions. Two or more facets can be combined in various ways—along with a definition of the objects of measurement, the persons—in order to define a set of observations, called the universe of admissible observations for a given attribute. The combination of facets follows the same logic as the combination of factors in factorial designs introduced by Fisher (1925); conditions of a facet are regarded as different levels of a factor. The set of persons is not called a facet but is treated in the same way.

This paper deals only with one- or two-facet designs in which facets and persons are completely crossed. The construct discussed is the ability of elementary school children to paint with watercolors. Assume that a teacher wants to assess these abilities. He or she may ask the students to make watercolor paintings on a particular topic, and rate them afterwards. If the teacher does not want to insist on a particular topic or rater, then he or she will accept as an admissible observation any rating given by a competent teacher of any painting on a topic appropriate to the age of the students. Accordingly, the universe of admissible observations can be defined by completely crossing two facets: topics (facet i) and teachers (facet j). Each combination of conditions of the facets i and j specifies one single observation of the ability of a given student p.

Let X_{pij} be the score that results from a single observation. According to the analysis of variance model, this score may be decomposed into an overall mean and several effects.

$$X_{pij} = \mu + \pi_p + \alpha_i + \beta_j + (\pi\alpha)_{pi} + (\pi\beta)_{pj} + (\alpha\beta)_{ij} + (\pi\alpha\beta,\varepsilon)_{pij} \tag{1}$$

where π_p is the person effect,

α_i is the effect of topic i on the observed score, and

β_j is the effect of teacher j.

The terms in parentheses are the interaction effects.

Correspondingly, the total variance in the universe of admissible observations may be decomposed into several variance components.

$$\sigma^2(X_{pij}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(j) + \sigma^2(pi) + \sigma^2(pj) + \sigma^2(ij) + \sigma^2(pij,e) \quad . \tag{2}$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Knowledge of the magnitude of each variance component is of great importance for considerations concerning reliability as well as validity of the measurement process in question. The magnitude of the variance components can be estimated in a so-called generalizability study (G study) in which several conditions of each facet are sampled. The estimates of the variance components resulting from a G study are used to determine the generalizability of the outcomes of decision studies (D studies) in which the parameters for the objects of measurement are estimated.

The Universe of Generalization

In a D study, a universe of generalization is specified first. The universe of generalization contains those facets and conditions to which an investigator wants to generalize in a D study with a particular measurement procedure. Along with the specification of a universe of generalization, universe scores for the objects of measurement are defined. A universe score is the expected value of the observed scores for an object of measurement over all conditions in the universe of generalization. In many instances the universe of generalization is identical to the universe of admissible observations. In the previous example this will be the case if the investigator wants to generalize over all conditions of facets i and j.

Let a student's observed score be defined as his or her average score under two randomly selected sets of conditions I and J with sizes n_i and n_j

$$X_{pIJ} = \frac{1}{n_i n_j} \sum_{ij}^{n_i n_j} X_{pij} \quad . \tag{3}$$

Correspondingly, Equations 1 and 2 can be written as

$$X_{pIJ} = \mu + \pi_p + \alpha_I + \beta_J + (\pi\alpha)_{pI} + (\pi\beta)_{pJ} + (\alpha\beta)_{IJ} + (\pi\alpha\beta,\varepsilon)_{pIJ} \tag{4}$$

and

$$\sigma^2(X_{pIJ}) = \sigma^2(p) + \sigma^2(I) + \sigma^2(J) + \sigma^2(pI) + \sigma^2(pJ) + \sigma^2(IJ) + \sigma^2(pIJ,e) \tag{5}$$

where $\beta_I = \sum_i^{n_i} \beta_i$ etc. and $\sigma^2(I) = \sigma^2(i)/n_i$ etc.

Generalizing over both facets i and j, a student's universe score is defined as

$$\mu_p = \mathrm{E}_{IJ}\, X_{pIJ} \quad . \tag{6}$$

If the investigator is interested in the absolute value of a universe score, he or she may choose to estimate μ_p by X_{pIJ}. The error associated with this estimation is denoted

$$\Delta_{pIJ} = X_{pIJ} - \mu_p \quad . \tag{7}$$

Since the expectation of the effects in Equation 4 over any of its subscripts is zero,

$$\mu_p = \mu + \pi_p \quad , \tag{8}$$

all effects in Equation 4 except π_p are part of the error Δ_{pIJ}, that will be made when μ_p is to be estimated by X_{pIJ}, generalizing over all conditions of facets i and j. This can be expressed by rewriting Equation 4 into

$$X_{pIJ} = \mu + \pi_p + \Delta_{pIJ} = \mu_p + \Delta_{pIJ} \quad . \tag{9}$$

The comparison between Equation 6 and Equation 1 shows that generalizability theory is simply an extension of classical test theory, in which the error term is divided into several components attributable to different sources.

Correspondingly, the total variance in the universe of generalization may be divided into universe score variance (associated with the concept of true score variance in classical test theory) and error variance, the latter containing all variance components in Equation 5 except $\sigma^2(p)$,

$$\sigma^2(X_{pIJ}) = \sigma^2(p) + \sigma^2(\Delta) \quad . \tag{10}$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Sometimes investigators are interested in the relative sizes of universe scores expressed by the deviation of a universe score from the overall mean $(\mu_p - \mu)$. These values may be estimated by the deviation of the observed score from the sample mean $(X_{pIJ} - X_{IJ})$. The error associated with this estimation is denoted

$$\delta_{pIJ} = (X_{pIJ} - X_{IJ}) - (\mu_p - \mu) \quad . \tag{11}$$

Forming Equation 11 to

$$\delta_{pIJ} = (X_{pIJ} - \mu_p) - (X_{IJ} - \mu) \quad , \tag{12}$$

makes the difference from error Δ in Equation 7 more clear. δ will generally be smaller than Δ because, from all effects in Equation 1, only the interaction effects indexed with p enter into the computation of δ. This paper concentrates on the estimation of the absolute values of universe scores since it is of interest in criterion-referenced measurement. Hence the error δ will not be further discussed. However, the same logic developed here with respect to error Δ may be applied to the computation of δ.

An investigator may choose to specify the universe of generalization to be different from the universe of admissible observations. He or she may, for instance, not intend to generalize over teachers. A universe score is then defined as

$$\mu_p = \mathrm{E}_I \, X_{pIj*} \quad , \tag{13}$$

$j*$ indicating one particular teacher. In this case the variance components $\sigma^2(J)$ and $\sigma^2(pJ)$ will no longer enter into the error variance $\sigma^2(\Delta)$.

The Multivariate Case

In the multivariate case a student's ability is represented by a vector of universe scores on different variables V. Let ${}_vX_{pIJ}$ represent an observed score of person p on variable v and ${}_v\mu_p$ his or her corresponding universe score on that variable. In this example the variables may be different criteria by which a painting can be judged. A student's observed scores on the different criteria can be weighted and combined to a composite score

$$X_{pIJ} = \sum_{v=1}^{n_V} w_v \, {}_vX_{pIJ} \quad , \tag{14}$$

where w_v = weight assigned to variable V; for convenience, $\Sigma w_v = 1$ and $w_v \geqslant 0$ for all v.

The corresponding universe score is

$$\mu_p = \sum_{v=1}^{n_V} w_v \, {}_v\mu_p \quad . \tag{15}$$

If the conditions of facets i and j are jointly sampled for all variables V, $\sigma^2(X)$ and $\sigma^2(\Delta)$ are determined not only by the magnitude of the variance components on each variable but also by the corresponding covariance components. For

$$\sigma^2(\Delta) = \sum_{v=1}^{n_V} \sum_{v'=1}^{n_V} w_v \, w_{v'} \, \sigma\,({}_v\Delta, \, {}_{v'}\Delta) \quad , \tag{16}$$

the covariance of the error Δ can be decomposed into the following components

$$\sigma({}_v\Delta, \, {}_{v'}\Delta) = \sigma({}_vI, \, {}_{v'}I) + \sigma({}_vJ, \, {}_{v'}J) + \sigma({}_vpI, \, {}_{v'}pI) + \sigma({}_vpJ, \, {}_{v'}pJ) + \sigma({}_vIJ, \, {}_{v'}IJ) + \sigma({}_vpIJ,e; \, {}_{v'}pIJ,e) \; . \tag{17}$$

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

In a multivariate generalizability study the variance components on each variable as well as the corresponding covariance components are estimated. The following sections will show how this information can be used to answer questions about construct validity, the optimal design of D studies, and the generalizability of D study outcomes.

The G Study

As mentioned earlier, the universe of admissible observations for a student's ability to paint in watercolors can be defined by completely crossing two facets: topics (i) and teachers (j).

Art educators (Denker, 1972; Kaiser, 1975) suggest three criteria for judging the paintings of elementary school children:

1. Are persons and things represented in an objective way?
2. Is the background appropriate?
3. Do relations between the objects become clear to the viewer?

Method

A teacher's ratings on these variables may be combined into a composite score X_{pij} for each painting. In order to estimate the variance components in Equation 2 for each variable and the corresponding covariance components, conditions were sampled from both facets i and j as well as from the person factor p. Sixty fourth-grade students were asked to make watercolor paintings on four topics arbitrarily chosen by the author. A sample of 25 art students was asked to judge the paintings. The sample consisted of students studying to be art teachers, who had answered an announcement in three different teacher's colleges in Germany.

After forty minutes of warming up in which the student teachers had to rate paintings that were not relevant to the G study, the 240 paintings were presented to them in random order. The raters did not know which paintings were done by the same student. For each of the 240 paintings, each student teacher made three ratings according to the three criteria. The students rated independently from one another. The ratings were made on a 10-point scale, with 10 indicating maximum ability. The raters were not informed of the purpose of the investigation. Time was not limited. Most participants took seven to eight hours to rate all 240 pictures. Raters were then paid for their participation.

Results

The rating data were analyzed by a three-way analysis of variance under the random model with one observation per cell. Using Cornfield & Tukey's (1956) equations for the expected mean squares, the values of the variance components were computed. The results are presented in the second column of Table 1. The third column shows the proportion of each variance component in relation to the estimated variance of scores in the universe of admissible observations. The fourth column contains confidence intervals for the estimates of the variance components, computed by the formulas of Satterthwaite (1941, 1946; see also Brennan, 1983, pp. 101).

The lower limits of the confidence intervals for $\sigma^2({}_v i)$ had negative values. Since the $\sigma^2({}_v i)$ must be considered essentially zero, no confidence intervals are reported for them. Looking at the other confidence intervals, it becomes evident that the estimates of the variance components may be relatively unstable. This finding is consistent with the results of monte carlo studies (cf. Smith, 1978) and larger G studies

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Estimates of the Variance Components

Variance Component	Estimate	Percentage of $\sigma^2({}_vX_{pij})$	95% Confidence Interval
$\sigma^2({}_1p)$	0.4156	11.0	$0.2762 \le \sigma^2({}_1p) \le 0.6958$
$\sigma^2({}_1i)$	0.0424	1.1	—
$\sigma^2({}_1j)$	1.3568	36.0	$0.8331 \le \sigma^2({}_1j) \le 2.5865$
$\sigma^2({}_1pi)$	0.3805	10.1	$0.3048 \le \sigma^2({}_1pi) \le 0.4878$
$\sigma^2({}_1pj)$	0.1351	3.6	$0.1043 \le \sigma^2({}_1pj) \le 0.1820$
$\sigma^2({}_1ij)$	0.0854	2.3	$0.0590 \le \sigma^2({}_1ij) \le 0.1348$
$\sigma^2({}_1pij,e)$	1.3501	35.9	$1.2947 \le \sigma^2({}_1pij,e) \le 1.4095$
$\sigma^2({}_1X_{pij})$	3.7659	100.0	
$\sigma^2({}_2p)$	0.3153	8.6	$0.2024 \le \sigma^2({}_2p) \le 0.5578$
$\sigma^2({}_2i)$	0.0053	0.1	—
$\sigma^2({}_2j)$	1.1601	31.7	$0.6943 \le \sigma^2({}_2j) \le 2.3248$
$\sigma^2({}_2pi)$	0.3782	10.3	$0.3010 \le \sigma^2({}_2pi) \le 0.4894$
$\sigma^2({}_2pj)$	0.1909	5.2	$0.1535 \le \sigma^2({}_2pj) \le 0.2438$
$\sigma^2({}_2ij)$	0.1236	3.4	$0.0865 \le \sigma^2({}_2ij) \le 0.1910$
$\sigma^2({}_2pij,e)$	1.4850	40.6	$1.4234 \le \sigma^2({}_2pij,e) \le 1.5496$
$\sigma^2({}_2X_{pij})$	3.6584	100.0	
$\sigma^2({}_3p)$	0.3897	9.0	$0.2732 \le \sigma^2({}_3p) \le 0.6005$
$\sigma^2({}_3i)$	0.0252	0.6	—
$\sigma^2({}_3j)$	1.5703	36.4	$0.9477 \le \sigma^2({}_3j) \le 3.0935$
$\sigma^2({}_3pi)$	0.4534	10.5	$0.3627 \le \sigma^2({}_3pi) \le 0.5831$
$\sigma^2({}_3pj)$	0.1218	2.8	$0.0880 \le \sigma^2({}_3pj) \le 0.1799$
$\sigma^2({}_3ij)$	0.0862	2.0	$0.0584 \le \sigma^2({}_3ij) \le 0.1400$
$\sigma^2({}_3pij,e)$	1.6652	38.6	$1.5953 \le \sigma^2({}_3pij,e) \le 1.7385$
$\sigma^2({}_3X_{pij})$	4.3118	100.0	

(cf. Nußbaum, 1980), which indicate that stable estimates of variance components can only be obtained by sampling large numbers of conditions in each facet. For this reason the values in Table 1 must be interpreted with care.

The results show nearly the same pattern for each of the three variables. The proportion of the residual variance in relation to the total variance is quite high. Since there is only one observation per

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

cell, the variance component for the triple interaction cannot be estimated separately from the error variance $\sigma^2({}_v e)$. The variance between the students $\sigma^2({}_v p)$ is approximately 10% of the observed variance. The values for $\sigma^2({}_v pi)$ are of the same magnitude, indicating that a student's achievement varied substantially over different topics. The low values for $\sigma^2({}_v i)$, which are essentially zero, show that the different topics were of the same difficulty. About one-third of the total variance is due to the fact that the raters differed from each other in the strength of their ratings. As indicated by the low values for $\sigma^2({}_v pj)$, a rater did not change his or her standard from one student to another. His or her standard was also constant over different topics, as indicated by the low values for $\sigma^2({}_v ij)$.

Table 2 shows the results for the covariance components. Since no adequate procedure is known to the author, no confidence intervals were computed for the covariance components. In view of the relatively small sample size, the estimates of the covariance components are also assumed to be quite unstable. In the fourth column of Table 2, correlation coefficients are reported in order to facilitate the interpretation of covariance components. Each correlation indicates how strongly the respective effects are correlated. The correlation between person effects on variable 1 and variable 2, for example, is defined by

$$r_{{}_1p\,{}_2p} = \frac{\sigma({}_1p, {}_2p)}{[\sigma^2({}_1p)\sigma^2({}_2p)]^{1/2}} \quad . \tag{18}$$

Since the variance components $\sigma^2({}_v i)$ as well as the corresponding covariance components $\sigma({}_v i,\ {}_{v'} i)$ must be considered zero, no correlation coefficient is defined for them.

The three variables are moderately correlated. This means that the three ratings a rater gave on a single picture tended to take the same position on the 10-point scale. The pattern of the results is quite similar for all combinations of the variables. The high values for $\sigma({}_v j,\ {}_{v'} j)$ show that nearly half of the covariance between the three variables is due to the fact that the three ratings come from the same rater. Hence, differences between the judges concerning the strength of their ratings are constant over all three variables. Approximately one-fourth of the total variance is due to the covariance between the residual terms. Since $\sigma({}_v e,\ {}_{v'} e)$ and $\sigma({}_v pij,\ {}_{v'} pij)$ are confounded, no unique interpretation is possible.

The universe score covariance $\sigma({}_v p,\ {}_{v'} p)$ as well as the covariance components for $\sigma({}_v pi,\ {}_{v'} pi)$ account for 12% of the total covariance. The respective correlation coefficients are very high, indicating that the abilities addressed by the different criteria are very similar to each other. This means that the ability to paint in watercolors must be considered a very general one that cannot be divided into several independent aspects, at least not into the ones that have been discussed. The remaining covariance components are very small and of little importance.

D Studies with Generalization Over Topics and Raters

D studies are made for the purpose of estimating universe scores. A teacher, who wants to assess a student's ability to paint in watercolors, may take the student's composite score X_{pIJ} as an estimate for his or her universe score μ_p:

$$\hat{\mu}_p = X_{pIJ} \quad . \tag{19}$$

In doing so, he or she generalizes from one topic and one rater over all possible topics and raters. The error variance $\sigma^2(\Delta)$ is the weighted sum of several variance and covariance components, as shown in Equations 4, 16, and 17. The larger the number of conditions sampled for each facet in the D study, the smaller will be $\sigma^2(\Delta)$, because values for the variance and covariance components that enter Equation 16 decrease with increasing numbers of conditions (n_i and n_j).

For each multivariate D study, the weights for the different variables must be specified. Denker (1972) considers criterion 1 to be most important when paintings of elementary school children are to be judged, followed by criterion 2 and 3. Under the restriction $\Sigma w_v = 1$; $w_v \geqslant 0$ for all v, the values w_1

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
Estimates of the Covariance Components

Covariance Component	Estimate	Percentage of $\sigma({}_vX_{pij}, {}_{v'}X_{pij})$	Correlation Coefficient
$\sigma({}_1p, {}_2p)$	0.2932	12.0	0.81
$\sigma({}_1i, {}_2i)$	-0.0314	-1.3	—
$\sigma({}_1j, {}_2j)$	1.1070	45.4	0.88
$\sigma({}_1pi, {}_2pi)$	0.2710	11.1	0.71
$\sigma({}_1pj, {}_2pj)$	0.1039	4.3	0.65
$\sigma({}_1ij, {}_2ij)$	0.0328	1.3	0.32
$\sigma({}_1pij{,}e, {}_2pij{,}e)$	0.6638	27.2	0.47
$\sigma({}_1X_{pij}, {}_2X_{pij})$	2.4402	100.0	0.65
$\sigma({}_1p, {}_3p)$	0.3908	12.9	0.97
$\sigma({}_1i, {}_3i)$	0.0196	0.6	—
$\sigma({}_1j, {}_3j)$	1.3196	43.6	0.90
$\sigma({}_1pi, {}_3pi)$	0.3518	11.6	0.85
$\sigma({}_1pj, {}_3pj)$	0.1033	3.4	0.81
$\sigma({}_1ij, {}_3ij)$	0.0663	2.2	0.77
$\sigma({}_1pij{,}e, {}_3pij{,}e)$	0.7722	25.5	0.52
$\sigma({}_1X_{pij}, {}_3X_{pij})$	3.0236	100.0	0.75
$\sigma({}_2p, {}_3p)$	0.3295	11.7	0.94
$\sigma({}_2i, {}_3i)$	-0.0217	-0.7	—
$\sigma({}_2j, {}_3j)$	1.2549	44.4	0.93
$\sigma({}_2pi, {}_3pi)$	0.3264	11.6	0.79
$\sigma({}_2pj, {}_3pj)$	0.1412	5.0	0.93
$\sigma({}_2ij, {}_3ij)$	0.0357	1.3	0.35
$\sigma({}_2pij{,}e, {}_3pij{,}e)$	0.7556	26.8	0.48
$\sigma({}_2X_{pij}, {}_3X_{pij})$	2.8217	100.0	0.71

= .5, w_2 = .33, and w_3 = .17, were specified. The proportions between these values are somewhat arbitrarily chosen and should only serve for demonstration.

In a real D study there are at least three ways to specify the weights. One is to define them by theory (e.g., a theory of human abilities). Another is to define them by practice (e.g., optimization of a classification procedure). A third is used by Webb & Shavelson (1981), who determine the weights with

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

respect to maximum generalizability of the composite score. These points are discussed in greater detail in Shavelson & Webb (1981) and Webb & Shavelson (1981).

Figure 1 shows the expected values for the standard error of measurement $\sigma(\Delta)$ under different combinations of n_i and n_j. The values for $\sigma(\Delta)$ are quite large. The best way to reduce $\sigma(\Delta)$ is to increase the number of raters in a D study. But even if several conditions of each facet are sampled in the D study, the generalization from X_{pij} to μ_p is burdened with a considerable amount of error. For three raters and four topics a 95% interval for the universe score, formed by

$$X_{pIJ} - z_{(1-\alpha/2)}\,\sigma(\Delta) \leq \mu_p \leq X_{pIJ} + z_{(1-\alpha/2)}\,\sigma(\Delta) \quad , \tag{20}$$

ranges over 3.1 points of the 10-point scale.

In a similar manner, confidence intervals can be specified for the deviation scores $(\mu_p - \mu)$. They would be considerably smaller because they would involve $\sigma(\delta)$ instead of $\sigma(\Delta)$.

A generalizability coefficient $\varepsilon\rho^2$ may be determined for D studies with different numbers of topics and raters similar to the reliability coefficient in classical test theory. However, to take $\varepsilon\rho^2$ as the estimated generalizability coefficient for a particular D study may turn out to be risky, since the conditions of facet i and j are not equivalent in terms of classical test theory. This will at least be the case for D studies with small numbers of conditions (see Cronbach et al., 1972, pp. 100–101).

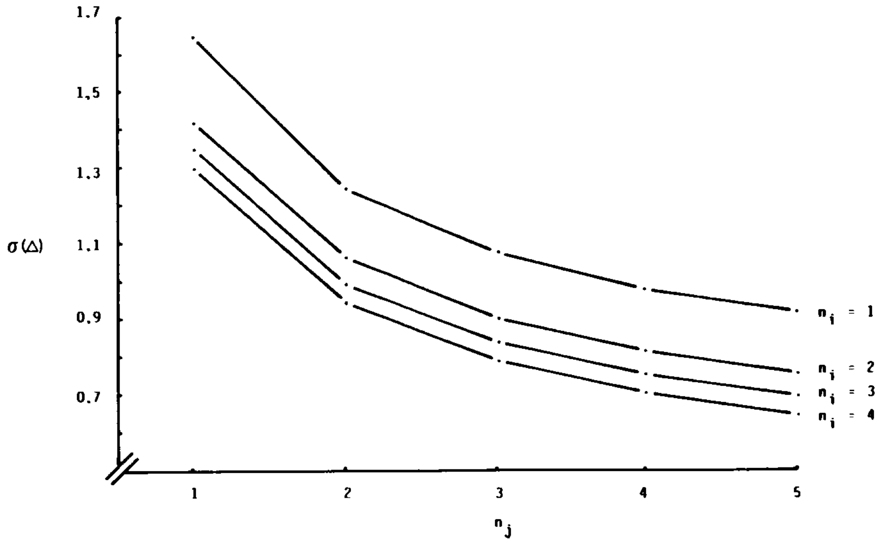
D Studies with Generalization Over Topics Only

A teacher may have good reason not to generalize over judges. He or she may, for example, regard his or her own statement as the only valid one because other teachers have no knowledge of the conditions under which the pictures were painted. In order to indicate that no generalization over facet j will be made, an observed composite score may be written as X_{pij*}. The observed variance of each variable V may be partitioned in the following way

$$\sigma^2({}_vX_{pij*}) = \sigma^2({}_vp|j*) + \sigma^2({}_vi|j*) + \sigma^2({}_vpi,e|j*) \quad . \tag{21}$$

Figure 1

$\sigma(\Delta)$ for Different Numbers of Conditions i and j



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

The variance components in Equation 21 are estimated by

$$\sigma^2({}_vp|j^*) = \sigma^2({}_vp) + \sigma^2({}_vpj) \quad , \tag{22}$$

$$\sigma^2({}_vi|j^*) = \sigma^2({}_vi) + \sigma^2({}_vij) \quad , \tag{23}$$

$$\sigma^2({}_vpi,e|j^*) = \sigma^2({}_vpi) + \sigma^2({}_vpij,e) \quad , \tag{24}$$

and the corresponding covariance components by

$$\sigma({}_vp|j^*, {}_{v'}p|j^*) = \sigma({}_vp, {}_{v'}p) + \sigma({}_vpj, {}_{v'}pj) \quad , \tag{25}$$

$$\sigma({}_vi|j^*, {}_{v'}i|j^*) = \sigma({}_vi, {}_{v'}i) + \sigma({}_vij, {}_{v'}ij) \quad , \tag{26}$$

$$\sigma({}_vpi,e|j^*, {}_{v'}pi,e|j^*) = \sigma({}_vpi, {}_{v'}pi) + \sigma({}_vpij,e; {}_{v'}pij,e) \quad . \tag{27}$$

Figure 2 shows the expected values of σ(Δ) for D studies with different numbers of topics. These values are considerably smaller than those in Figure 1.

Whether the advantage of higher accuracy is worth the price that must be paid by abandoning the possibility of generalizing over judges, may be decided in each single case. Problems like this are merely an extension of the well-known bandwidth-fidelity dilemma discussed by Cronbach & Gleser (1965).

Summary and Conclusions

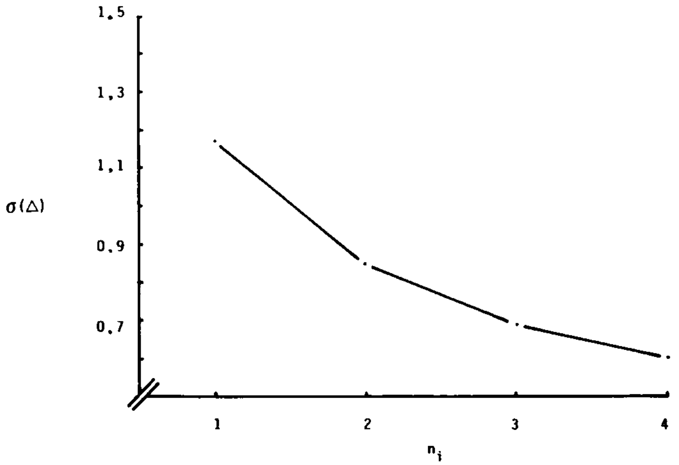
It has been shown that multivariate generalizability theory can be relevant to the convergent validity and the reliability of educational measures, for which the equivalence assumptions of classical test theory do not hold.

The results of this G study indicated that—with respect to different criteria on which the teacher ratings were based—the ability of students to paint with watercolors can be regarded as a relatively homogeneous attribute. However, the results can also be interpreted as evidence against the discriminant validity of the three criteria involved or simply be indicative of a "halo effect" in the judgments of the raters.

On the other hand, a student's achievement varies considerably over different topics. Moreover, the raters differ very much in the strength of their ratings. The generalizability of D study outcomes could

Figure 2

σ (Δ) for Different Numbers of Conditions *i*



Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227. May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

be shown to be quite low even if several conditions of each facet are sampled in the D study. If generalization over judges is abandoned, the expected error of measurement is much smaller.

The distinction between G study and D study enables the investigator to plan his or her D study individually, and to estimate the generalizability of the D study outcomes on the basis of G study outcomes. This is of great practical importance. However, the large confidence intervals for the estimates of the variance components to be found in this study as well as in the area of achievement testing (Nußbaum, 1980) indicate that caution is needed when the results of a G study are used to determine the generalizability of D study outcomes. Additional empirical research is necessary to clarify this point.

References

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

Brennan, R. L., & Kane, M. T. (1977a). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277–289.

Brennan, R. L., & Kane, M. T. (1977b). Signal/noise rations for domain-referenced tests. *Psychometrika, 42,* 609–625; Errata, 1978, *43,* 289.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *Annuals of Mathematical Statistics, 27,* 907–949.

Cronbach, L. J. (1976). On the design of educational measures. In D. N. M. De Gruijter & L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement* (pp. 199–208). New York: Wiley.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

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: John Wiley & Sons, Inc.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16,* 137–163.

Denker, J. (1972). *Kunstunterricht in der Grundschule*. Oldenburg: Isensee.

Endler, N. S., & Hunt, J. M. (1966). Sources of behavioral variance as measured by the S-R Inventory of Anxiousness. *Psychological Bulletin, 65,* 336–346.

Endler, N. S., & Hunt, J. M. (1969). Generalizability of contributions from sources of variance in the S-R Inventories of Anxiousness. *Journal of Personality, 37,* 1–14.

Erlich, O., & Borich, G. (1979). Occurrence and generalizability of scores on a classroom interaction instrument. *Journal of Educational Measurement, 16,* 11–18.

Erlich, O., & Shavelson, R. J. (1978). The search for correlations between measurements of teacher behavior and student achievement: Measurement problem, conceptualization problem, or both? *Journal of Educational Measurement, 15,* 77–89.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver and Boyd.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15,* 1–13.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A ''universe-defined'' system of arithmetic achievement tests. *Journal of Educational Measurement, 5,* 275–290.

Kaiser, G. (1975). *Kunstunterricht in der Eingangstufe*. Ravensburg: O. Maier.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47,* 267–292.

Kane, M. T., & Brennan, R. L. (1980). Agreement coefficients as indices of dependability for domain-referenced tests. *Applied Psychological Measurement, 4,* 105–126.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13,* 171–183.

Levy, P. (1974). Generalizability studies in clinical settings. *British Journal of Social and Clinical Psychology, 13,* 161–172.

Nußbaum, A. (1980). *Konstruktion, Planung und Analyse lehrzielorientierter Tests auf der Grundlage der Generalisierbarkeitstheorie*. Unpublished doctoral dissertation, RWTH Aachen.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika, 6,* 309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Shavelson, R., & Atwood-Russo, N. (1977). Generalizability of measures of teacher effectiveness. *Educational Research, 19,* 171–183.

Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of measures of teaching behavior. *Review of Educational Research, 46,* 553–611.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133–166.

Smith, P. L. (1978). Sampling errors of variance components in small sample multifacet generalizability studies. *Journal of Educational Statistics, 3,* 319–346.

Smith, P. L. (1979). The generalizability of student ratings of courses: Asking the right question. *Journal of Educational Measurement, 16,* 77–87.

Webb, N. M., & Shavelson, R. J. (1981). Multivariate generalizability of general educational development ratings. *Journal of Educational Measurement, 18,* 13–22.

Author's Address

Send requests for reprints or further information to Albert Nußbaum, Institut für Erziehungswissenschaft, Lehrstuhl Pädagogik III der RWTH Aachen, Eilfschornsteinstr. 7, D-5100 Aachen, Federal Republic of Germany.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/