



ELSEVIER

## CONTRIBUTORS' PROOFREADING INSTRUCTIONS FOR – ENCYCLOPEDIA OF SOCIAL MEASUREMENT

### PROOFREADING

The text content and layout of your article is not in final form when you receive proofs. Read proofs for accuracy and clarity, as well as for typographical errors, but **DO NOT REWRITE**.

Check titles and headings for spelling and capitalization. Ensure that the right typeface and size have been used to indicate the proper level of each heading. Review numbered items such as figures, tables, and lists for proper order. Proofread the captions and credit lines for illustrations and tables. Ensure that any permissioned material has the required credit line. Artwork appears in its final size and placement, but may not match the image quality of the final printed publication and is for content checking only.

Any questions from the copy editor will be listed following the article proof. Please address these questions as necessary. While it is appreciated that some articles may require updating/revising, please try to keep alterations to a minimum. Excessive alterations may be charged to the contributors.

Please keep a copy of any corrections you make.

### RETURNING PROOFS

Proofs should be returned to Aimee Bishop at Elsevier by **May 26** by one of the following methods:

1. If corrections are minor, they should be sent in an e-mail to **SOME\_proofs@elsevier.com**. This e-mail should state the encyclopedia title and article title and ID number. Each correction listed should clearly identify the page number, paragraph number, line number within that paragraph, and the correction itself.

2. If corrections are more substantial, the amended hard copy should be mailed by express courier or faxed (only if marked very clearly) to:

Aimee Bishop  
Production Project Manager  
Elsevier  
The Boulevard  
Langford Lane  
Kidlington  
Oxford OX 5 1GB  
UK

Fax: **+44 (0) 1865 843974**

3. If the proofs are OK as is, please inform Aimee Bishop via either email (**SOME\_proofs@elsevier.com**) or fax (**+44 (0) 1865 843974**); be sure to list the article ID number and title.

Please note that your corrections should be returned as quickly as possible. Scheduling is crucial at this stage of production, and a delay in the return of the proofs could mean missing the press date, resulting in delayed publication.

### CHECKLIST

- |   |                          |
|---|--------------------------|
| Author name(s) and affiliation(s) checked?  | <input type="checkbox"/> |
| Figures and tables checked?   | <input type="checkbox"/> |
| Further Reading section checked and completed?  | <input type="checkbox"/> |
| Outstanding permissions letters returned to Elsevier?   | <input type="checkbox"/> |
| Author's mailing address checked and updated?   | <input type="checkbox"/> |
| [Note: this information will be used only to update our records;<br>this section will not be printed in the final article.] |                          |
| Manuscript queries addressed/answered?  | <input type="checkbox"/> |

If you have any questions regarding your proofs, contact Aimee Bishop at **SOME\_proofs@elsevier.com**.



a0005

# Generalizability Theory

*Richard J. Shavelson*

*Stanford University, Stanford, California, USA*

*Noreen M. Webb*

*University of California, Los Angeles, Los Angeles, California, USA*

## Glossary

- g0005 condition** The levels of a facet (e.g., task 1, task 2, . . . , task  $k$ ).
- g0010 decision (D) study** A study that uses information from a G study to design a measurement procedure that minimizes error for a particular purpose.
- g0015 facet** A characteristic of a measurement procedure such as a task, occasion, or observer that is defined as a potential source of measurement error.
- g0020 generalizability (G) study** A study specifically designed to provide estimates of the variability of as many possible facets of measurement as economically and logistically feasible considering the various uses a test might be put to.
- g0025 universe of admissible observations** All possible observations that a test user would consider acceptable substitutes for the observation in hand.
- g0030 universe of generalization** The conditions of a facet to which a decision maker wants to generalize.
- g0035 universe score** The expected value of a person's observed scores over all observations in the universe of generalization (analogous to a person's true score in classical test theory); denoted  $\mu_p$ .
- g0040 variance component** The variance of an effect in a G study.

**p0005** Generalizability (G) theory, a statistical theory for evaluating the dependability (reliability) of behavioral measurements, grew from the recognition that the undifferentiated error in classical test theory provided too gross a characterization of the multiple sources of measurement error. Whereas in classical test theory, measurement error is random variation and the multiple error sources are undifferentiated, G theory considers both systematic and unsystematic sources of error variation

and disentangles them simultaneously. Moreover, in contrast to the classical parallel-test assumptions of equal means, variances, and covariances, G theory assumes only randomly parallel tests sampled from the same universe. These developments expanded the conceptions of error variability and reliability that can be applied to different kinds of decisions using behavioral measurements. In G theory a behavioral measurement (e.g., achievement test score) is conceived of as a sample from a universe of admissible observations, which consists of all possible observations that decision makers consider to be acceptable substitutes for the observation in hand. Each characteristic of the measurement situation (e.g., test form, item, or occasion) is called a facet and a universe of admissible observations is defined by all possible combinations of the levels of the facets. To estimate different sources of measurement error, G theory extends earlier analysis-of-variance approaches to reliability and focuses heavily on variance component estimation and interpretation to isolate different sources of variation in measurements and to describe the accuracy of generalizations made from the observed to the universe scores of individuals. In contrast to experimental studies, analysis of variance is not used to formally test hypotheses.

## Generalizability Studies

s0005

In order to evaluate the dependability of behavioral measurements, a generalizability (G) study is designed to isolate particular sources of measurement error. The facets that the decision maker might want to generalize over (e.g., items or occasions) must be included.

s0010 **Universe of Generalization**

p0015 The universe of generalization is defined as the set of conditions to which a decision maker wants to generalize. A person's universe score (denoted  $\mu_p$ ) is defined as the expected value of his or her observed scores over all observations in the universe of generalization (analogous to a person's true score in classical test theory).

s0015 **Decomposition of Observed Score**

p0020 With data collected in a G study, an observed measurement can be decomposed into a component or effect for the universe score and one or more error components. Consider a random-effects two-facet crossed  $p \times i \times o$  (person by item by occasion) design. The object of measurement, here people, is not a source of error and, therefore, is not a facet. In the  $p \times i \times o$  design with generalization over all admissible test items and occasions taken from an indefinitely large universe, the observed score for a particular person ( $p$ ) on a particular item ( $i$ ) and occasion ( $o$ ) is:

$$\begin{aligned}
 X_{pio} = \mu & && \text{grand mean} \\
 + \mu_p - \mu & && \text{person effect} \\
 + \mu_i - \mu & && \text{item effect} \\
 + \mu_o - \mu & && \text{occasion effect} \\
 + \mu_{pi} - \mu_p - \mu_i + \mu & && \text{person} \times \text{item effect} \\
 + \mu_{po} - \mu_p - \mu_o + \mu & && \text{person} \times \text{occasion effect} \\
 + \mu_{io} - \mu_i - \mu_o + \mu & && \text{item} \times \text{occasion effect} \\
 + X_{pio} - \mu_p - \mu_i - \mu_o & && \\
 + \mu_{pi} + \mu_{po} + \mu_{io} - \mu & && \text{residual} \quad (1)
 \end{aligned}$$

where  $\mu = E_o E_i E_p X_{pio}$  and  $\mu_p = E_o E_i X_{pio}$ , and  $E$  means expectation. The other terms in (1) are defined analogously. Assuming a random-effects model, the distribution of each effect, except for the grand mean, has a mean of zero and a variance  $\sigma^2$  (called the variance component). The variance of the person effect,  $\sigma_p^2 = E_p(\mu_p - \mu)^2$ , called universe-score variance, is analogous to the true-score variance of classical test theory. The variance components for the other effects are defined similarly. The residual variance component,  $\sigma_{pio,e}^2$ , indicates that the person  $\times$  item  $\times$  occasion interaction is confounded with residual error because there is one observation per cell. The collection of observed scores,  $X_{pio}$ , has a variance,  $\sigma^2(X_{pio}) = E_o E_i E_p \times (X_{pio} - \mu)^2$ , which equals the sum of the variance components:

$$\sigma^2(X_{pio}) = \sigma_p^2 + \sigma_i^2 + \sigma_o^2 + \sigma_{pi}^2 + \sigma_{po}^2 + \sigma_{io}^2 + \sigma_{pio,e}^2 \quad (2)$$

An estimate of each variance component can be p0025 obtained from a traditional analysis of variance (or other methods such as maximum likelihood). The relative magnitudes of the estimated variance components provide information about potential sources of error influencing a behavioral measurement. Statistical tests are not used in G theory; instead, standard errors for variance component estimates provide information about sampling variability of estimated variance components.

**Decision Studies**

s0020

G theory distinguishes a decision (D) study from a G p0030 study. The G study is associated with the development of a measurement procedure and the D study uses information from a G study to design a measurement that minimizes error for a particular purpose. In planning a D study, the decision maker defines the universe that he or wishes to generalize to, called the universe of generalization, which may contain some or all of the facets and conditions in the universe of admissible observations. In the D study, decisions usually are based on the mean over multiple observations rather than on a single observation. The mean score over a sample of  $n'_i$  items and  $n'_o$  occasions, for example, is denoted as  $X_{pIO}$ , in contrast to a score on a single item and occasion,  $X_{pio}$ . A two-facet, crossed D-study design in which decisions are to be made on the basis of  $X_{pIO}$  is, then, denoted as  $p \times I \times O$ .

**Types of Decisions and Measurement Error**

s0025

G theory recognizes that the decision maker might want p0035 to make two types of decisions based on a behavioral measurement: relative and absolute.

**Measurement Error for Relative Decisions**

s0030

A relative decision concerns the rank ordering of indivi- p0040 duals (e.g., norm-referenced interpretations of test scores). For relative decisions, the error in a random-effects  $p \times I \times O$  design is defined as:

$$\delta_{pIO} = (X_{pIO} - \mu_{IO}) - (\mu_p - \mu) \quad (3)$$

where  $\mu_p = E_o E_i X_{pIO}$  and  $\mu_{IO} = E_p X_{pIO}$ . The variance of the errors for relative decisions is:

$$\begin{aligned}
 \sigma_\delta^2 &= E_p E_i E_o \delta_{pIO}^2 = \sigma_{pI}^2 + \sigma_{pO}^2 + \sigma_{pIO,e}^2 \\
 &= \frac{\sigma_{pi}^2}{n'_i} + \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{pio,e}^2}{n'_i n'_o} \quad (4)
 \end{aligned}$$

In order to reduce  $\sigma_\delta^2$ ,  $n'_i$  and  $n'_o$  may be increased (analogous to the Spearman-Brown prophecy formula in classical test theory and the standard error of the mean in sampling theory).

s0035 **Measurement Error for Absolute Decisions**

p0045 An absolute decision focuses on the absolute level of an individual's performance independent of others' performance (cf. domain-referenced interpretations). For absolute decisions, the error in a random-effects  $p \times I \times O$  design is defined as:

$$\Delta_{pIO} = X_{pIO} - \mu_p \quad (5)$$

and the variance of the errors is:

$$\begin{aligned} \sigma_{\Delta}^2 &= E_p E_I E_O \Delta_{pIO}^2 \\ &= \sigma_I^2 + \sigma_O^2 + \sigma_{pI}^2 + \sigma_{pO}^2 + \sigma_{IO}^2 + \sigma_{pIO,e}^2 \\ &= \frac{\sigma_i^2}{n_i'} + \frac{\sigma_o^2}{n_o'} + \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{io}^2}{n_i' n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'} \end{aligned} \quad (6)$$

s0040 **Coefficients**

p0050 Although G theory stresses the importance of variance components and measurement error, it provides summary coefficients that are analogous to the reliability coefficient in classical test theory (i.e., true-score variance divided by observed-score variance; an intraclass correlation). The theory distinguishes between a generalizability coefficient for relative decisions and an index of dependability for absolute decisions.

s0045 **Generalizability Coefficient**

p0055 The generalizability coefficient is analogous to classical test theory's reliability coefficient (the ratio of the universe-score variance to the expected observed-score variance; an intraclass correlation). For relative decisions and a  $p \times I \times O$  random-effects design, the generalizability coefficient is:

$$E\rho^2(X_{pIO}, \mu_p) = E\rho^2 = \frac{E_p(\mu_p - \mu)^2}{E_O E_I E_p (X_{pIO} - \mu_{IO})^2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} \quad (7)$$

s0050 **Dependability Index**

p0060 For absolute decisions with a  $p \times I \times O$  random-effects design, the index of dependability is:

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{\Delta}^2} \quad (8)$$

The right-hand side of Eqs. (7) and (8) are generic expressions that apply to any design and universe. For domain-referenced decisions involving a fixed cutting score  $\lambda$  (often called criterion-referenced measurements), and assuming that  $\lambda$  is a constant that is specified *a priori*, the error of measurement is:

$$\Delta_{pIO} = (X_{pIO} - \lambda) - (\mu_p - \lambda) = X_{pIO} - \mu_p \quad (9)$$

and the index of dependability is:

$$\Phi(\lambda) = \frac{E_p(\mu_p - \lambda)^2}{E_O E_I E_p (X_{pIO} - \lambda)^2} = \frac{\sigma_p^2 + (\mu - \lambda)^2}{\sigma_p^2 + (\mu - \lambda)^2 + \sigma_{\Delta}^2} \quad (10)$$

An unbiased estimator of  $(\mu - \lambda)^2$  is  $(\bar{X} - \lambda)^2 - \sigma^2(\bar{X})$ , MQ2 where  $\bar{X}$  is the observed grand mean over sampled objects of measurement and sampled conditions of measurement in a D-study design.

s0055

## Generalizability- and Decision-Study Designs

G theory allows the decision maker to use different p0065 designs in the G and D studies. Although G studies should use crossed designs whenever possible to avoid confounding of effects, D studies may use nested designs for convenience or for increasing sample size, which typically reduces estimated error variance and, hence, increases estimated generalizability. For example, compare  $\sigma_{\delta}^2$  in a crossed  $p \times I \times O$  design and a partially nested  $p \times (I : O)$  design, where facet  $i$  is nested in facet  $o$ , and  $n'$  denotes the number of conditions of a facet under a decision maker's control:

$$\begin{aligned} \sigma_{\delta}^2 \text{ in a } p \times I \times O \text{ design} &= pI + \sigma_{pO}^2 + \sigma_{pIO}^2 \\ &= \frac{\sigma_{pi}^2}{n_i'} + \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pio,e}^2}{n_i' n_o'} \end{aligned} \quad (11)$$

$$\begin{aligned} \sigma_{\delta}^2 \text{ in a } p \times (I : O) \text{ design} &= \sigma_{pO}^2 + \sigma_{pI:O}^2 \\ &= \frac{\sigma_{po}^2}{n_o'} + \frac{\sigma_{pi,pio,e}^2}{n_i' n_o'} \end{aligned} \quad (12)$$

In Eqs. (11) and (12),  $\sigma_{pi}^2$ ,  $\sigma_{po}^2$ , and  $\sigma_{pio,e}^2$  are directly available from a G study with design  $p \times xi \times o$  and  $\sigma_{pi,pio,e}^2$  is the sum of  $\sigma_{pi}^2$  and  $\sigma_{pio,e}^2$ . Moreover, given cost, logistics, and other considerations,  $n'$  can be manipulated to minimize error variance, trading off, in this example, items and occasions. Due to the difference in the designs,  $\sigma_{\delta}^2$  is smaller in Eq. (12) than in (11).

s0060

## Random and Fixed Facets

Generalizability theory is essentially a random effects the- p0070 ory. Typically a random facet is created by randomly sampling conditions of a measurement procedure (e.g., tasks from a job in observations of job performance). When the conditions of a facet have not been sampled randomly from the universe of admissible observations but the intended universe of generalization is infinitely large, the concept of exchangeability may be invoked to consider the facet as random.

p0075 A fixed facet (cf. fixed factor in analysis of variance) arises when (1) the decision maker purposely selects certain conditions and is not interested in generalizing beyond them, (2) it is unreasonable to generalize beyond conditions, or (3) the entire universe of conditions is small and all conditions are included in the measurement design. G theory typically treats fixed facets by averaging over the conditions of the fixed facet and examining the generalizability of the average over the random facets. When it does not make conceptual sense to average over the conditions of a fixed facet, a separate G study may be conducted within each condition of the fixed facet or a full multivariate analysis may be performed.

p0080 G theory recognizes that the universe of admissible observations encompassed by a G study may be broader than the universe to which a decision maker wishes to generalize in a D study, the universe of generalization. The decision maker may reduce the levels of a facet (creating a fixed facet), select (and thereby control) one level of a facet, or ignore a facet. A facet is fixed in a D study when  $n' = N'$ , where  $n'$  is the number of conditions for a facet in the D study and  $N'$  is the total number of conditions for a facet in the universe of generalization. From a random-effects G study with design  $p \times i \times o$  in which the universe of admissible observations is defined by facets  $i$  and  $o$  of infinite size, fixing facet  $i$  in the D study and averaging over the  $n_i$  conditions of facet  $i$  in the G study ( $n_i = n'_i$ ) yields the following estimated universe-score variance:

$$\sigma_{\tau}^2 = \sigma_p^2 + \sigma_{pIO}^2 = \sigma_p^2 + \frac{\sigma_{pi}^2}{n'_i} \quad (13)$$

where  $\sigma_{\tau}^2$  denotes estimated universe-score variance in generic terms.  $\sigma_{\tau}^2$  in Eq. (13) is an unbiased estimator of universe-score variance for the mixed model only when the same levels of facet  $i$  are used in the G and D

studies. Estimates of relative and absolute error variance, respectively, are:

$$\sigma_{\delta}^2 = \sigma_{pO}^2 + \sigma_{pIO}^2 = \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{pio,e}^2}{n'_i n'_o} \quad (14)$$

$$\begin{aligned} \sigma_{\Delta}^2 &= \sigma_o^2 + \sigma_{pO}^2 + \sigma_{IO}^2 + \sigma_{pIO}^2 \\ &= \frac{\sigma_o^2}{n'_o} + \frac{\sigma_{po}^2}{n'_o} + \frac{\sigma_{io}^2}{n'_i n'_o} + \frac{\sigma_{pio,e}^2}{n'_i n'_o} \end{aligned} \quad (15)$$

## Numerical Example

s0065

As an example, consider the following 1998 G study, by p0085 Webb, Nemer, Chizhik, and Sugrue, of science achievement test scores. In this study, 33 eighth-grade students completed a six-item test on knowledge of concepts in electricity on two occasions, 3 weeks apart. The test required students to assemble electric circuits so that the bulb in one circuit was brighter than the bulb in another circuit and to answer questions about the circuits. Students' scores on each item ranged from 0 to 1, based on the accuracy of their judgment and the quality of their explanation about which circuit, for example, had higher voltage. The design was considered fully random.

Table I gives the estimated variance components from p0090 the G study.  $\sigma_p^2$  (0.03862) is fairly large compared to the other components (27% of the total variation). This shows that, averaging over items and occasions, students in the sample differed in their science knowledge. Because people constitute the object of measurement, not error, this variability represents systematic individual differences in achievement. The other large estimated variance components concern the item facet more than the occasion facet. The nonnegligible  $\sigma_i^2$  (5% of the total variation) shows

t0005 **Table I** Generalizability Study and Alternative Decision Studies for the Measurement of Science Achievement

Source of variation	$\sigma^2$	G study		Alternative D studies				
		$n'_i = 1$ $n'_o = 1$	$n'_i =$ $n'_i =$	6 1	6 2	8 3	12 1	12 2
Person	$\sigma_p^2$	0.03862	$\sigma_p^2$	0.03862	0.03862	0.03862	0.03862	0.03862
Item	$\sigma_i^2$	0.00689	$\sigma_i^2$	0.00115	0.00115	0.00086	0.00057	0.00057
Occasion	$\sigma_o^2$	0.00136	$\sigma_o^2$	0.00136	0.00068	0.00045	0.00136	0.00068
$pi$	$\sigma_{pi}^2$	0.03257	$\sigma_{pi}^2$	0.00543	0.00543	0.00407	0.00271	0.00271
$po$	$\sigma_{po}^2$	0.00924	$\sigma_{po}^2$	0.00924	0.00462	0.00308	0.00924	0.00462
$io$	$\sigma_{io}^2$	0 <sup>a</sup>	$\sigma_{io}^2$	0	0	0	0	0
$pio,e$	$\sigma_{pio,e}^2$	0.05657	$\sigma_{pio,e}^2$	0.00943	0.00471	0.00236	0.00471	0.00236
	$\sigma_{\delta}^2$	0.09838		0.02410	0.01476	0.00951	0.01667	0.00969
	$\sigma_{\Delta}^2$	0.10663		0.02661	0.01659	0.01082	0.01860	0.01095
	$\rho^2$	0.28		0.62	0.72	0.80	0.70	0.80
	$\Phi$	0.27		0.59	0.70	0.78	0.67	0.78

MQ1

<sup>a</sup> Negative estimated variance component (−0.00093) set to zero.

that items varied somewhat in difficulty level. The large  $\sigma_{pi}^2$  (22%) reflects different relative standings of people across items. The small  $\sigma_o^2$  (1% of the total variation) indicates that performance was stable across occasions, averaging over students and items. The nonnegligible  $\sigma_{po}^2$  (6%) shows that the relative standing of students differed somewhat across occasions. The zero  $\sigma_{io}^2$  indicates that the rank ordering of item difficulty was the same across occasions. Finally, the large  $\sigma_{pio,e}^2$  (39%) reflects the varying relative standing of people across occasions and items and/or other sources of error not systematically incorporated into the G study.

p0095 Table I also presents the estimated variance components, error variances, and generalizability coefficients for several decision studies varying in the number of items and occasions. Because more of the variability in achievement scores came from items than from occasions, changing the number of items has a larger effect on the estimated variance components and coefficients than does changing the number of occasions. The optimal number of items and occasions is not clear; for a fixed number of observations per student, different combinations of numbers of items and occasions give rise to similar levels of estimated generalizability. Choosing the optimal number of conditions of each facet in the D study involves logistical and cost considerations as well as issues of generalizability (reliability). Because administering more items on fewer occasions is usually less expensive than administering fewer items on more occasions, a decision maker will probably choose a 12-item test administered twice over an eight-item test administered three times. No feasible test length will produce a comparable level of generalizability for a single administration, however—

p0100 The optimal D study design need not be fully crossed. In this example, administering different items on each occasion (*i:o*) yields slightly higher estimated generalizability than does the fully crossed design; for example, for 12 items and two occasions,  $\rho^2 = 0.82$  and  $\phi = 0.80$ . The larger values of  $\rho^2$  and  $\phi$  for the partially nested design than for the fully crossed design are solely attributable to the difference between Eqs. (11) and (12).

## s0070 Multivariate Generalizability

p0105 For behavioral measurements involving multiple scores describing individuals' aptitudes or skills, multivariate generalizability can be used to (1) estimate the reliability of difference scores, observable correlations, or universe-score and error correlations for various D study designs and sample sizes; (2) estimate the reliability of a profile of scores using multiple regression of universe scores on the observed scores in the profile; or (3) produce a composite

of scores with maximum generalizability. For all these purposes, multivariate G theory decomposes both variances and covariances into components. In a two-facet, crossed  $p \times i \times o$  design with two dependent variables, the observed scores for the two variables for person  $p$  observed under conditions  $i$  and  $o$  can be denoted as  ${}_1X_{pio}$  and  ${}_2X_{pio}$ , respectively. The variances of observed scores,  $\sigma^2({}_1X_{pio})$  and  $\sigma^2({}_2X_{pio})$ , are decomposed as in Eq. (2).

The covariance,  $\sigma({}_1X_{pio}, {}_2X_{pio})$ , is decomposed in p0110 analogous fashion:

$$\begin{aligned} \sigma({}_1X_{pio}, {}_2X_{pio}) &= \sigma({}_1p, {}_2p) + \sigma({}_1i, {}_2i) + \sigma({}_1o, {}_2o) \\ &+ \sigma({}_1pi, {}_2pi) + \sigma({}_1po, {}_2po) \\ &+ \sigma({}_1io, {}_2io) + \sigma({}_1pio, e, {}_2pio, e) \end{aligned} \quad (16)$$

In Eq. (16) the term  $\sigma({}_1p, {}_2p)$  is the covariance between universe scores on variables 1 and 2, say, ratings on two aspects of writing: organization and coherence. The remaining terms in Eq. (16) are error covariance components. The term  $\sigma({}_1i, {}_2i)$ , for example, is the covariance between scores on the two variables due to the conditions of observation for facet  $i$ .

An important aspect of the development of multivariate G theory is the distinction between linked and p0115 unlinked conditions. The expected values of error covariance components are zero when conditions for observing different variables are unlinked, that is, selected independently (e.g., the items used to obtain scores on one variable in a profile, writing organization, are selected independently of the items used to obtain scores on another variable, writing coherence). The expected values of error covariance components are nonzero when conditions are linked or jointly sampled (e.g., scores on two variables in a profile, organization and coherence, come from the same items).

In 1976, Joe and Woodward presented a G coefficient p0120 for a multivariate composite that maximizes the ratio of universe-score variation to the universe score plus error variation. Alternatives to using canonical weights that maximize the reliability of a composite are to determine variable weights on the basis of expert judgment or to use weights derived from a confirmatory factor analysis.

## Issues in the Estimation of Variance Components

s0075

Given the emphasis on estimated variance components in p0125 G theory, any fallibility of their estimates is a concern. One issue is the sampling variability of estimated variance components; a second is how to estimate variance components, especially in unbalanced designs.

s0080 **Sampling Variability of Variance Component Estimates**

p0130 Assuming that mean squares are independent and score effects have a multivariate normal distribution, the sampling variance of an estimated variance component ( $\sigma^2$ ) is:

$$\sigma^2(\sigma^2) = \frac{2}{c^2} \sum_q \frac{E(MS_q)^2}{df_q} \quad (17)$$

where  $c$  is the constant associated with the estimated variance component;  $E(MS_q)$  is the expected value of the mean square,  $MS_q$ ; and  $df_q$  is the degrees of freedom associated with the  $MS_q$ . In the  $p \times i \times o$  design, for example,  $\sigma_p^2$  is estimated by  $(MS_p - MS_{pi} - MS_{po} + MS_{pio,e}) / (n_i * n_o)$ . Using Eq. (17) to estimate the variance of  $\sigma_p^2$ ,  $c$  refers to  $n_i * n_o$ , and  $MS_q$  refers to  $MS_p$ ,  $MS_{pi}$ ,  $MS_{po}$ , and  $MS_{pio,e}$ . The more mean squares that are involved in estimating variance components, the larger the estimated variances are likely to be (e.g., compare standard errors  $\sigma(\sigma_p^2) = 0.01360$  and  $\sigma(\sigma_{pio,e}^2) = 0.00632$  for the results in Table I). Furthermore, the variances of estimated variance components will be larger with smaller numbers of observations per person (reflected in smaller  $df_q$ ). Although exact confidence intervals for variance components are generally unavailable (due to the inability to derive exact distributions for variance component estimates), approximate confidence intervals are available, as are resampling techniques such as bootstrapping.

s0085 **Estimates of Variance Components**

p0135 Although analysis-of-variance methods for estimating variance components is straightforward when applied to balanced data and has the advantages of requiring few distributional assumptions and producing unbiased estimators, problems arise with unbalanced data. They include many different decompositions of the total sums of squares without an obvious basis for choosing among them (which leads to a variety of ways in which mean squares can be adjusted for other effects in the model), biased estimation in mixed models (not a problem in G theory because G theory averages over fixed facets in a mixed model and estimates only variances of random effects, or mixed models can be handled via multivariate G theory), and algebraically and computationally complex rules for deriving expected values of mean squares.

p0140 In 1987, Searle reviewed several alternative methods of estimating variance components that do not have the drawbacks of analysis-of-variance methods. Maximum likelihood (ML) and restricted maximum likelihood (REML) methods of estimation produce estimators that are normally distributed and have known sampling

variances at least under large-sample conditions. Minimum norm quadratic unbiased estimation (MINQUE) and minimum variance quadratic unbiased estimation (MIVQUE), unlike ML and REML, do not assume normality and do not involve iterative estimation, thus reducing computational complexity. However, MINQUE and MIVQUE can produce different estimators from the same data set, and estimates may be negative and are usually biased. In 2001, Brennan described two resampling techniques, bootstrap and jackknife, that can be used to estimate variance components and standard errors. Drawing on Wiley's 2001 dissertation, bootstrap now appears to be potentially applicable to estimating variance components and their standard errors and confidence intervals when the assumption of normality is suspect.

Another concern with variance component estimation is when a negative estimate arises because of sampling errors or because of model misspecification. Possible solutions when negative estimates are small in relative magnitude are to (1) substitute zero for the negative estimate and carry through the zero in other expected mean square equations from the analysis of variance, which produces biased estimates; (2) set negative estimates to zero but use the negative estimates in expected mean square equations for other components; (3) use a Bayesian approach that sets a lower bound of zero on the estimated variance component; and (4) use ML or REML methods, which preclude negative estimates.

### Further Reading

- Brennan, R. L. (2001). "Generalizability Theory." Springer-Verlag, New York.
- Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). "The Dependability of Behavioral Measurements." John Wiley, New York.
- Feldt, L. S., and Brennan, R. L. (1989). Reliability. In "Educational Measurement" (R. L. Linn, ed.), 3rd Ed., pp. 105–146. American Council on Education/Macmillan, Washington, D.C.
- Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educ. Psychol. Meas.* **54**, 3–7.
- Searle, S. R. (1987). "Linear Models for Unbalanced Data." John Wiley, New York.
- Shavelson, R. J., and Webb, N. M. (1981). Generalizability theory: 1973–1980. *Br. J. Math. Statist. Psychol.* **34**, 133–166.
- Shavelson, R. J., and Webb, N. M. (1991). "Generalizability Theory: A Primer." Sage, Newbury Park, CA.
- Webb, N. M., Nemer, K., Chizhik, A., and Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *Am. Educ. Res. J.* **35**, 607–651.

Webb, N. M., Shavelson, R. J., and Maddahian, E. (1983). Multivariate generalizability theory. *In* "Generalizability Theory: Inferences and Practical Applications" (L. J. Fyans, ed.), pp. 67–81. Jossey-Bass, San Francisco, CA.

Wiley, E. (2000). "Bootstrap Strategies for Variance Component Estimation: Theoretical and Empirical Results." Unpublished doctoral diss., Stanford University, Stanford, CA.

ELSEVIER FIRST PROOFS

**Manuscript queries and/or remarks - please respond to these points on returning your proofs**

- (MQ1) What is this symbol?  $\sigma$ ?
- (MQ2) Please identify symbol. Is this  $\sigma$ ?

ELSEVIER FIRST PROOFS