Generalizability Theory: Overview

Noreen M. Webb and Richard J. Shavelson

Volume 2, pp. 717–719

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Generalizability Theory: Overview

Generalizability (G) theory is a statistical theory for evaluating the dependability ('reliability') of behavioral measurements [2]; see also [1], [3], and [4]. G theory pinpoints the sources of measurement error, disentangles them, and estimates each one. In G theory, a behavioral measurement (e.g., a 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 that a decision maker would be indifferent to (e.g., test form, item, occasion, rater) is a potential source of error and is called a *facet of a measurement*. The universe of admissible observations, then, is defined by all possible combinations of the levels (called *conditions*) of the facets. In order to evaluate the dependability of behavioral measurements, a *generalizability* (G) *study* is designed to isolate and estimate as many facets of measurement error as is reasonably and economically feasible.

Consider a two-facet crossed person x item x occasion G study design where items and occasions have been randomly selected. The *object of measurement*, here persons, is not a source of error and, therefore, is not a facet. In this design with generalization over all admissible test items and occasions taken from an indefinitely large universe, an observed score for a particular person on a particular item and occasion is decomposed into an effect for the grand mean, plus effects for the person, the item, the occasion, each two-way interaction (*see* **Interaction Effects**), and a residual (three-way interaction plus unsystematic error). The distribution of each component or 'effect', except for the grand mean, has a mean of zero and a variance σ^2 (called the **variance component**). The variance component for the person effect is called the *universe-score variance.* The variance components for the other effects are considered error variation. Each variance component can be estimated from a traditional **analysis of variance** (or other methods such as **maximum likelihood**). The relative magnitudes of the estimated variance components provide information about 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.

The *decision* (D) *study* deals with the practical application of a measurement procedure. A D study uses variance component information from a G study to design a measurement procedure that minimizes error for a particular purpose. In planning a D study, the decision maker defines the universe that he or she wishes to generalize to, called the *universe of generalization*, which may contain some or all of the facets and their levels in the universe of admissible observations. In the D study, decisions usually will be based on the mean over multiple observations (e.g., test items) rather than on a single observation (a single item).

G theory recognizes that the decision maker might want to make two types of decisions based on a behavioral measurement: relative ('norm-referenced') and absolute ('criterion- or domain-referenced'). A *relative decision* focuses on the rank order of persons; an *absolute decision* focuses on the level of performance, regardless of rank. Error variance is defined differently for each kind of decision. To reduce error variance, the number of conditions of the facets may be increased in a manner analogous to the Spearman–Brown prophecy formula in classical test theory and the standard error of the mean in sampling theory. G theory distinguishes between two reliability-like summary coefficients: a Generalizability (G) Coefficient for relative decisions and an Index of Dependability (Phi) for absolute decisions.

Generalizability theory allows the decision maker to use different designs in G and D studies. Although G studies should use crossed designs whenever possible to estimate all possible variance components in the universe of admissible observations, D studies may use nested designs for convenience or to increase estimated generalizability.

G theory is essentially a random effects theory. Typically, a random facet is created by randomly sampling levels of a facet. A fixed facet arises when the decision maker: (a) purposely selects certain conditions and is not interested in generalizing beyond them, (b) finds it unreasonable to generalize beyond the levels observed, or (c) when the entire universe of levels is small and all levels are included in the measurement design (*see* **Fixed and Random Effects**). 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. Alternatives include conducting a separate G study within each condition of the fixed facet, or a multivariate analysis with the levels of the fixed facet comprising a vector of dependent variables.

As an example, consider a G study in which persons responded to 10 randomly selected science items on each of 2 randomly sampled occasions. Table 1 gives the estimated variance components from the G study. The large $\hat{\sigma}^2_{\mathrm{p}}$ (1.376, 32% of the total variation) shows that, averaging over items and occasions, persons in the sample differed systematically in their science achievement. The other estimated variance components constitute error variation; they concern the item facet more than the occasion facet. The nonnegligible $\hat{\sigma}^2_{\mathrm{i}}$ (5% of total variation) shows that items varied somewhat in difficulty level. The large $\hat{\sigma}^2_{\mathrm{pi}}$ (20%) reflects different relative standings of persons across items. The small $\hat{\sigma}^2_{\mathrm{o}}$ (1%) indicates that performance was stable across occasions, averaging over persons and items. The nonnegligible $\hat{\sigma}^2_{\mathrm{po}}$ (6%) shows that the relative standing of persons differed somewhat across occasions. The zero $\hat{\sigma}^2_{\mathrm{io}}$ indicates that the rank ordering of item difficulty was similar across occasions. Finally, the large $\hat{\sigma}^2_{\mathrm{pio,e}}$ (36%) reflects the varying relative standing of persons across occasions and items and/or other sources of error not systematically incorporated into the G study.

Table 1 Estimated variance components in a generalizability study of science achievement (p × i × o design)

Source	Variance component	Estimate	Total variability (%)
Person (p)	σ^2_{p}	1.376	32
Item (i)	σ^2_{i}	0.215	05
Occasion (o)	σ^2_{o}	0.043	01
p × i	σ^2_{pi}	0.860	20
p × o	σ^2_{po}	0.258	06
i × o	σ^2_{io}	0.001	00
p × i × o,e	$\sigma^2_{\mathrm{pio,e}}$	1.548	36

Because more of the error variability in science achievement scores came from items than from occasions, changing the number of items will have a larger effect on the estimated variance components and generalizability coefficients than will changing the number of occasions. For example, the estimated G and Phi coefficients for 4 items and 2 occasions are 0.72 and 0.69, respectively; the coefficients for 2 items and 4 occasions are 0.67 and 0.63, respectively. Choosing the number of conditions of each facet in a D study, as well as the design (nested vs. crossed, fixed vs. random facet), involves logistical and cost considerations as well as issues of dependability.

References

[1] Brennan, R.L. (2001). *Generalizability Theory*, Springer-Verlag, New York.

[2] Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*, Wiley, New York.

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

[4] Shavelson, R.J. & Webb, N.M. (1991). *Generalizability Theory: A Primer*, Sage Publications, Newbury Park.

(*See also* **Generalizability Theory: Basics**; **Generalizability Theory: Estimation**)

Noreen M. Webb and Richard J. Shavelson