Generalizability Research on Alternate Assessments in Reading

The purpose of this chapter is to provide states with a brief explanation of generalizability Theory, including the theory-specific nomenclature. By understanding the terminology and procedures for conducting a generalizability study, it is possible to better understand specific sources of error. In the reliability chapter, measurement error was described as coming from the student or various aspects of the test (the way it is constructed, administered, or scored). This chapter extends that presentation by specifically parsing error from the test into different sources (called facets or conditions of measurement).

The critical topics of this report include an overview of generalizability theory, including both G-studies (addressing variance source and amount) and D-studies (addressing selection of items for a test, including the number and type of items). As noted above, central to this treatment of generalizability is the idea that error can be parsed in terms of its source and analyzed separately (into facets or conditions) rather than considering it as single dimension left over from the true score. Finally, we address the rationale for and findings from a research study on generalizability.

Overview of Generalizability Theory

"Generalizability theory provides a conceptual and statistical framework for identifying major sources of measurement error and for estimating measurement precision for various measurement procedures" (Gao and Brennan, 2001, p. 192). According to *Standards* * (AERA, APA, NCME, 1999), measurement developers and users benefit from this type of information when properly estimated and presented. Similar standards are endorsed by the International Test Commission (whose membership includes most national psychological associations internationally, and the International Test Commission, 2001). Clearly, knowing the extent to which measurements are dependable is widely accepted as good practice, relevant to anyone involved in testing and assessment.

Generalizability theory (G Theory) is not new, though its wide use is relatively current. Introduced originally by Cronbach and colleagues (1963, 1972), the procedures address some important limitations of the true score test model and classical reliability theory. All measurement is assumed to contain some kind of error; while classical test theory considers an observed score to include true score and undifferentiated error, generalizability theory decomposes the error term into components that are related to various facets. “Generalizability theory liberalizes classical [test] theory by employing ANOVA [analysis of variance] methods that allow an investigator to untangle multiple sources of error that contribute to the undifferentiated E [error] in classical theory” (Brennan, 2001, p. 3). Furthermore, “Generalizability analyses are useful for not only understanding the relative importance of various sources of error but also for designing efficient measurement procedures” (Brennan, 2001, p. 4).

A sampling perspective provides intuitive appreciation for G theory procedures. Most error is introduced when observations are poorly sampled. This is especially true with assessments based on constructed responses, e.g., performance assessments, in contrast to selected response formats. The systematic influences on responses and scores comprise the sampling space for any measurement procedure. Items, raters, formats, and occasion are examples of these types of influences that can produce error, thereby resulting in unreliable measurement. A generally accepted approach to diminishing this type of error is to increase the sample size using the Spearman-Brown prophecy formula (Feldt and Brennan, 1989). Perhaps more items, another rater, or maybe another format will reduce the error. Knowing where to increase or decrease the sampling is the purpose of a well-designed generalizability study.

Generalizability and decision studies. Parsing error into specific components and estimating necessary sampling configurations is accomplished with G theory by conducting a sequence of two types of studies: (a) a generalizability study (G-study) that addresses *universes of admissible observations* to variance associated with sources of error and (b) a decision study (D-study) that considers *universes of generalizations* for user-specified hypothetical measurement sampling configurations (i.e., numbers of items, raters, etc.). A D-study estimates examinee universe scores (true scores) along with various reliability and dependability indices. "Generalizability analyses are useful for not only understanding the relative importance of various sources of error but also for designing efficient measurement procedures" (Brennan, 2001, p. 4). Together, a G-study and D-study aid measurement developers and users with very specific information about measurement error and optimal measurement designs.

Measurement facets. The universe of admissible observations typically is discussed in terms of the measurement 'facet': "A facet is simply a set of similar conditions of measurement" (Brennan, 2001, p. 6), e.g., items, raters. Note the term is not applied to populations (persons), who serve as the primary objects of measurement. Basically, various facets are identified for data collection using a carefully designed study. Then, ANOVA procedures are used to understand how much variance is associated with each facet (by decomposing the total observed score into variance from separate facets as well as variance from the interaction of facets). In G-studies, the focus is on understanding (estimating) the amount of variance associated with a universe of admissible observations (operationalized through facets). After the G-study is completed, and the variance components have been estimated, then a decision study is conducted. D-studies "emphasize the estimation, use, and interpretation of variance components for decision-making with well-specified measurement procedures" (Brennan, 2001, p. 9); basically the focus is on making generalizations (over replications) based on the results of the G-study.

Two types of measurement decisions and error estimates. Corresponding to classification and ranking decisions, both absolute and relative errors are considered, respectively, when interpreting generalizability results. Absolute error refers to the difference between a person's observed and universe (true) score, while relative error refers to the difference between a person's observed deviation and universe deviation score (and are usually less error prone than absolute error).

Reliability Indices

In generalizability theory, two coefficients are used:

- $E\rho^2$ (generalizability coefficient): universe score variance/(universe score variance + relative error variance); it is the analogue of a reliability coefficient. $E\rho^2$ is valuable when appraising how reliably a measurement system will consistently rank individuals with respect to their universe (true) scores.

- ϕ (index of dependability): universe score variance/(universe score variance + absolute error variance); it is appropriate when scores are given 'absolute' interpretations and is valuable when considering classification problems and the need to consistently make absolute types of decisions about an individual.

Facets can be random (unrestricted) or fixed (restricted). "The estimated generalizability coefficient is larger when facets are considered fixed because a universe of generalization with a fixed facet is narrower than a universe of generalization [both] facets random. That is generalization to narrow universes are less error prone than generalizations to broader universes" (Brennan, 2001, p. 15).

Rationale for Research on Generalizability

The reason that this topic is important relates to the way that measurement error is considered. In classical test theory, a person's observed score is comprised of true score and measurement error. If we remove the error, then the observed score would be the same as the true score. The problem arises in what we call measurement error, which is generally assumed to be random variation that is unrelated to the construct being measured. But what if the variation is not random and yet is still unrelated to the construct being measured? Generalizability allows us to study different sources of error that are part of the measurement process but unrelated to the construct being measured. Following are three examples of facets that are traditionally studied as part of generalizability research.

1. Tasks – All measurement systems must present a stimulus to which students respond. In large-scale tests, this stimulus is usually a multiple-choice item or a performance task. Generalizability theory frequently has been used to study the influence that tasks have on performance. Although multiple-choice test items rarely vary much, performance tasks often contain considerable variation.

2. Occasions – If measures are given more than once, it would be possible to determine the influence that time (or occasion) has on performance. The administration conditions (including the teacher and context) may be an important source of measurement error.

3. Raters – When performance tasks are administered, raters need to be given a rubric or scoring guide, trained how to conduct the rating, and then systematically make judgments of the performance. It is very possible that the raters themselves may influence the scores obtained by the raters (e.g. some may be harsh while others are lenient).

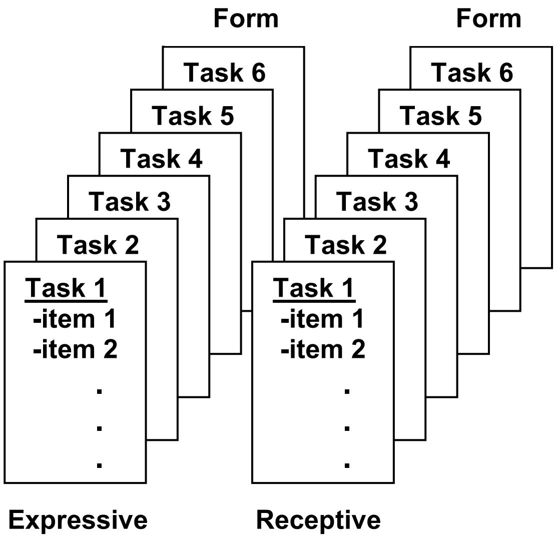
Other facets (the term used in generalizability theory to reference the measurement dimension of influence) may be studied, but tasks, occasions, and raters are typically studied with G-theory.

Summary of a Research Study on Generalizability of Reading Measurements

In this study, we used generalizability theory to study two formats of a reading test: (a) an expressive format in which the student had to read the words and (b) a receptive format in which the test administrator read the words and only asked the student to respond as appropriate. Because each student took both formats, we have to use different items in each format (and counter balance the order of administration). Nevertheless, we considered form to be a random variable (not fixed). Following is a brief summary of this study.

Methods. All materials (six different tasks in reading with several items in each task) were developed with administration and scoring directions embedded in a teacher edition with the student materials separately bundled (containing no directions but only the stimulus words, sentences, passages, and questions, all of which were presented on flash cards). See Figure 1.

Figure 1. Form, Task, and Item Sampling Structure



These six tasks each contained a number of items as listed in Table 2 below.

Table 2. Number of Items per Task by Format

	Administration Format	
Task	Expressive	Receptive
1. Identify Signs and Symbols	10	10
2. Letter Naming	10	10
3. Word Reading	10	10
4. Sentence Reading	5	5
5. Passage Reading	6	6
6. Comprehension Passage	6	6

After administration of all the tasks, the scoring protocols (administrator copy) was placed in a folder and five judges rated the overall reading performance using a 1-5 rating scale. See Table 3 for a description of the anchors of each value.

Table 3. Rating Scale for Rating Overall Student Performance Across All Six Tasks

0	1	2	3	4	5
Student demonstrates no behavior.	Student demonstrates virtually no comprehension and few symbol, letter, or word skills.	Student demonstrates limited comprehension with some symbol, letter, and word reading skills.	Student demonstrates emerging comprehension with some sentence reading and few passage reading skills.	Student demonstrates basic comprehension with accurate word and sentence reading skills, and some passage reading skills.	Student demonstrates full comprehension, including accurate symbol, letter, word, sentence, and passage reading.

Results. We found that students performed better with the receptive format. With expressive formats, performance across the six reading tasks was .44 on form A and .31 on Form B. In contrast, with the receptive format, the average performance across the six tasks was .59 and .56 on Forms A and B, respectively. This difference is about 15%-25%. Basically, the receptive format resulted in higher performance. We also found, however, considerable variation in the difficulty of the tasks. While letter naming was easy, reading words and passages was difficult.

An important finding was that items interacted with persons and that this effect was more pronounced with the receptive than the expressive format. As we note in the full report, “when comparing expressive versus receptive formats, the variance accounted by persons interacting with items was greater with receptive tasks than expressive tasks, except for naming letters. This means that the format of the task had a significant influence on the variance for p x i (persons interacting with items). This finding has an important implication in that, though receptive tasks may result in higher average performance, this performance has more variance that cannot be explained until specific reference is made to the person and the item. Typically, this variance accounted for between 11% and 16% of the performance” (p. 20).

In the D-Study, we found that, for most tasks, a relatively few number of items would, nevertheless, result in high generalizability coefficients. “Increasing the number of items beyond 5 rarely resulted in (practically) significant improvements in the generalizability coefficient. Increasing the number of items beyond 5 rarely resulted in (practically) significant improvements in the generalizability coefficient. Finally, when considering the number of raters needed to achieve high generalizability coefficients, either 2 or 3 would be sufficient, beyond which only small increases would be found. For expressive tasks, these coefficients were typically higher for decoding tasks (naming letters, reading words, sentences, and passages); however, when the task focused on meaning (signs/symbols and comprehension), more items would be needed. The same was true for receptive format, though all tasks required more items to achieve sufficient generalizability coefficients” (p. 21).

These results are important, particularly given the nature of the students' disabilities with an important lesson learned: standardization comes at a cost. For students who can only express themselves receptively (blink, nod, eye gaze, etc.), it is possible to develop alternatives to the traditional constructed response in which students must express themselves. However, more items are likely to be necessary in obtaining accurate levels of performance.

Books on Generalizability Theory

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

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*. New York: John Wiley.

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

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

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.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, *16*, 137-163.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (pp. 105-146). New York: Macmillan Publishing Company.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*, 191-203.

International Test Commission (2001). International guidelines for test use. *International Journal of Testing*, *1*, 93-114.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72-101.

Tindal, G., Yovanoff, P., & Geller, J. (2006). *Generalizability theory applied to reading assessments for students with significant cognitive disabilities*. Unpublished study reported at http://daata.org.