Macrothink Institute™

Journal of Studies in Education
ISSN 2162-6952
2012, Vol. 2, No. 1

Generalizability Theory: An Analysis of Variance Approach to Measurement Problems in Educational Assessment

Hussain Alkharusi

College of Education, Sultan Qaboos University

P.O.Box: 32 Al-Khod, P.C.: 123, Sultanate of Oman

Tel: 968-9622-2535 E-mail: hussein5@squ.edu.om

Received: January 3 Accepted: February 20 Published: February 22, 2012

doi:10.5296/jse.v2i1.1227 URL: http://dx.doi.org/10.5296/jse.v2i1.1227

Abstract

Of increasing interest to the educational assessment researchers is the reliability of scores obtained by various measurement procedures like tests, rating scales, surveys, and observation forms. Traditional methods of reliability based on Classical Test Theory (CTT) consider only one source of measurement errors. Generalizability Theory (GT) extends CTT by providing a flexible and practical framework for estimating the effects of multiple sources of measurement errors through an application of analysis of variance procedures. It helps educational assessment researchers to determine appropriate conditions in terms of items, occasions, raters conducive to obtaining an optimal level of score reliability. This paper highlights the utility and applicability of the GT analysis in the educational assessment research. The paper begins with a brief history of the GT tracing its growth and development and then discusses its defining features and advantages and disadvantages. The paper illustrates the application of the GT in the educational assessment research. The paper closes with a recommendation for educational assessment researchers to take advantage of the GT in the development of measurement procedures and further research agenda in this area.

Keywords: Generalizability theory, Educational measurement, Reliability

1. Introduction

The generalizability theory (GT) can be viewed as an extension of the classical test theory CTT through an application of certain analysis of variance (ANOVA) procedures to measurement issues. Just as the researcher, using ANOVA , attempts to identify and estimate the effects of potentially important factors, GT researcher attempts to identify and estimate the magnitude of the potentially important sources of measurements error. The connection between GT and CTT is paralleled by the connection between factorial and simple ANOVA. A researcher, using simple ANOVA, partitions variance into two components, usually named between-group variance and within-group variance. The between-group variance is thought of as a systematic variance associated with the factor that distinguishes groups from one another. The within-group variance is thought of as random and treated as error. In the same way, CTT partitions variance into true-score variance and error variance. The former is the thought of as a systematic variance associated with differences between objects of measurement. The latter is treated as random variance unrelated to the true-score variance (Shavelson & Webb, 1991).

By applying factorial ANOVA instead of simple ANOVA, the researcher acknowledges multiple factors contributing to the total variance in the observations, and hence partitions it into parts corresponding to each factor, to interactions among the factors, and to a random error. Similarly, GT acknowledges multiple influences on measurement variance. Whereas CTT, like simple ANOVA, partitions variance into only two sources, GT, like factorial ANOVA, partitions variance into many sources corresponding to a systematic variance among the objects of measurement, to multiple error sources, and to their interactions (Shavelson & Webb, 1991). This paper attempts to provide an overview of the fundamentals of the GT. It highlights its features, assumptions, advantages, and disadvantages. The paper begins with a brief history of the GT to trace its growth and development. Then, an example is provided to illustrate the application of the GT in the educational assessment research. The example focuses on the simple one-facet design, which is the most common measurement design used in the GT analysis. However, it should be noted that the generalizability analysis presented in this paper can be extended to almost any type of univariate and multivariate designs consisting of two or more factors, which may be random or fixed, and which may be crossed or nested. It should be acknowledged that the paper is an expository summary of the GT which is originated from Brennan (2001), Shaveslson and Webb (1991), and other measurement experts. The readers are encouraged to utilize the references cited in the paper for more details about the topic.

2. A brief Historical Overview of GT

Table 1 summarizes important and practical contributions in the history of GT (Brennan, 1997; Crocker & Algina, 1986; Shavelson & Webb, 1981). Although several researchers can be credited with paving the way for GT (e.g., Burt, 1936; Hoyt, 1941), it was formally introduced by Cronbach and his associates as an extension to the CTT (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965).

Table 1. A summary of important and practical contributions in the history of GT

Year	Contribution
1936	Burt applied ANOVA approach to measurement problems in the analysis of examination marks.
1941	Hoyt showed that ANOVA can be employed to compute reliability coefficients by treating students and items as sources of variations.
1951	- Finlayson's study of grades assigned to essays was probably the first treatment of reliability in terms of variance components. - Ebel published an article on the reliability of ratings in which he considered two types of error variance: one that included and another that excluded rater main effects. Ebel also considered single-facet crossed and nested designs.
1955 to 1959	The rater main effects in Ebel's (1951) article played the role of the item main effects in Lord's (1955, 1957, 1959) articles about conditional standard errors of measurement and reliability under the assumptions of the binomial error model. Ebel and Lord's works were eventually captured the distinction between relative and absolute error in GT.
1960 to 1965	Cronbach and his colleagues Gleser and Rajaratnam had pretty much completed their developmentnof univariate GT.
1966	Cronbach and Nada began the work on multivariate GT, in which each of the levels of one or more fixed facts is associated with a distinct universe score.
1972	Cronbach, Gleser, Nada, Rajaratnam published a book entitled, *The Dependability of Behavioral Measurements: Theory of Gneralizabilty for Scores and Profiles.* This monograph is still the most definitive treatment of the GT.
1976 to 1981	Cardinet, Tourneur, and Allal emphasized the role that facets other than students might play as objects of measurement, which was known as *principle of symmetry* of the GT.
1983	Crick and Brennan designed a computer program called *GENOVA* for conducting GT analysis.
1989	Feldt and Brennan devoted about one third of their chapter on reliability to GT.
1991	Shavelson and Webb published a relatively short monograph entitled, *Generalizabilty Theory: A primer.*

1992	Brennan provided a very brief introduction about GT in the *Educational Measurement: Issues and Practice* intended primarily for classroom use.

3. Defining Features of GT

GT is a measurement theory for estimating the dependability of measurements obtained by any kind of procedure like tests, rating scales, surveys, and observation forms. Dependability refers to the accuracy of generalizing from a student's observed score on a test or other measure to the average score that student would have received under the possible conditions such as all possible forms, all possible testing occasions, or all possible items. This average score is called a "universe score", which is analogous to the CTT's concept of "true-score". Thus, GT considers scores as dependable if they permit accurate inferences about the universe of admissible observations that they are meant to represent (Allal & Cardinet, 1997; Shavelson & Webb, 1991).

In GT, any observed score is considered to be a sample from a universe of admissible observations. This universe consists of all possible observations that would be acceptable as substitutes for the observation in question. For example, a score obtained by a particular student on a particular testing day is not the only acceptable indicator of his or her performance. A score obtained on a different day would also be acceptable, as would a score from a different form of the same test, or possibly a different set of items from the same form. Each of these characteristics of the testing situation is called a facet, and the levels of the facets are called conditions. The terms "facets" and "conditions" are analogous to "factors" and "levels" in the literature on experimental designs (Shavelson & Webb, 1991; Webb, Rowley, & Shavelson, 1988).

The universe is defined in terms of the facets of the observations that determine the conditions under which an acceptable score can be obtained. A one-facet universe is defined by one source of measurement error. For example, if the decision maker wants to generalize from the score on one set of test items to a much larger set of test items, items are a facet of the measurement and the item universe would be defined by all admissible items. If the decision maker wants to generalize from performance on one occasion to performance on a much larger set of occasions, occasions are a facet and the occasions' universe would be defined by all admissible occasions (Shavelson & Webb, 1991).

Traditional methods of reliability that are based on CTT consider only one source of error in a measurement at time. For example, test-retest reliability considers only the occasions of testing as the source of error. Parallel-forms reliability considers the forms of the test as the only source of error. Internal consistency reliability considers only the items as the source of error. As such, CTT provides very limited information. More specifically, information about one kind of reliability (e.g., test-retest) cannot be used to make inferences about other kinds of reliability (e.g., internal consistency). Furthermore, even if different kinds of reliability are presented, it is difficult to use the combined information to determine how many test forms, items, and occasions need to be used to obtain dependable measures (Webb, Rowley, & Shavelson, 1988).

However, GT extends CTT by providing a flexible and practical framework for estimating the effects of multiple sources of error. In particular, CTT states that an observed score for any student obtained through some measurement procedure can be decomposed into the true score and a single error. In contrast, GT recognizes that multiple sources for error such as error attributed to test items, testing occasions, and test forms may occur simultaneously in the measurement process. As such, the basic approach underlying GT is to decompose an observed score into a component for the universe score and one or more error components (Crocker & Algina, 1986; Shavelson & Webb, 1991).

4. Assumptions of GT

The assumptions underlying GT are basically the same assumptions of CTT. First, the data examined in a generalizabilty analysis should be interval or ordinal in nature. Second, GT assumes that a student's observed score is comprised of his or her universe score and/or more sources of error. Third, the errors are assumed to be independent of the universe score and uncorrelated. In other words, all of the effects in the measurement model are independent. Fourth, GT assumes that the samples used to estimate the error variances and selected of students, items, or occasions and comprise random samples from their respective populations. However, these facts can sometimes be treated as fixed. In particular, the concept of "randomness" states that even though conditions of a facet have not been sampled randomly, the facet may be considered to be random if conditions not observed in the study can be exchanged with the observed conditions. For example, if the researcher is willing to exchange the 30 items on a test for another sample of 30 items, the facet might be treated reasonably as random. The fifth assumption is that the standard errors are the same at all score levels. In other words, the same standard error of measurement is often applied to all objects of measurement regardless of the underlying universe score (Shavelson & Webb, 1991; Strube, 2002).

5. An illustration of Generalizability Analysis

To illustrate the application of GT, consider the following example: A test consisting of a random sample of n_i = 5 items, from a universe of items, was administered to a random sample of n_s = 5 students, from a population of students. Table 2 presents hypothetical data for this example. This design is called a one-facet design because the items facet is the only facet of potential measurement error being investigated. The generalizability question is that how dependable are scores made under different conditions of test items to draw inferences about the universe consisting of all conditions?

Table 2. Data from a hypothetical one-facet design

Student	Items				
	1	2	3	4	5
1	7	7	8	7	8
2	5	6	8	6	6
3	5	5	6	5	5
4	9	5	7	3	4
5	8	6	8	4	3

The observed score (X_{si}) for one student *(s)* on one item *(i)* can be expressed in terms of the following linear model (Brennan, 2001; Shavelson & Webb, 1991):

$$X_{si} = \mu + \mu_s + \mu_i + \mu_{si} + e, \qquad \text{model (1)}$$

Where

μ = is the overall mean in the population of students and universe of items.

μ_s = is the score effect attributable to student *s*.

μ_i = is the score effect attributable to item *i*.

μ_{si} = is the score effect attributable to the interaction of student and item.

e = is the random error for student's *s* score.

Because there is only one score for each student-item combination in Table 2, the terms μ_{si} and e are confounded. In other words, after accounting for student effect and item effect, we do not know if differences between item scores reflect the student-item interaction, the random variability, or both. Consequently, this confounding is represented with the notation *(si,e)* and is often referred to as the residual effect. Thus, model (1) is implemented as a two-way non-factorial analysis in which the interaction is not estimated. As such, model (1) can be written as follows (Brennan, 2001; Shavelson & Webb, 1991):

$$X_{si} = \mu + \mu_s + \mu_i + \mu_{si,e}, \qquad \text{model (2)}$$

The universe score for student *s* (μ_s) is defined as the expected value of a student's observed

score across the universe of items. Similarly, the population mean for item i (μ_i) is defined as the expected value over students. The basic assumptions underlying model (2) are that all effects are sampled independently, and the expected value of each effect over the population of students and the universe of items is equal to zero. Given these assumptions, the model is considered a random-effects students-crossed-with-items ($s \times i$) one-facet design. In fact, the choice of the appropriate ANOVA model is determined by the model of sampling, either fixed or random, of the levels of each facet (Brennan, 2001; Shavelson & Webb, 1991).

Once the data have been collected, the standard procedures of ANOVA are applied to determine the mean squares and to estimate the variance components corresponding to all sources of variation in the design. For this simple design in which students are crossed with items, variance components can be estimated by the random effects model for the three sources of variations: students (σ_s^2), items (σ_i^2), and residual ($\sigma_{si,e}^2$). Table 3 shows the standard ANOVA table for the students-by-items design along with corresponding computational formulas (Brennan, 2001).

Table 3. ANOVA formula for the students-by-items design

Source of variation	Sum of squares	Degrees of freedom	Mean squares	Variance component
Students (s)	$SS_s = n_i \sum_s (\bar{X}_s - \bar{X})^2$	$df_s = n_s - 1$	$MS_s = \frac{SS_s}{df_s}$	$\sigma_s^2 = \frac{MS_s - MS_{si,e}}{n_i}$
Items (i)	$SS_i = n_s \sum_i (\bar{X}_i - \bar{X})^2$	$df_i = n_i - 1$	$MS_i = \frac{SS_i}{df_i}$	$\sigma_i^2 = \frac{MS_i - MS_{si,e}}{n_s}$
Residuals (si,e)	$SS_{si,e} = \sum_s \sum_i (X_{si} - \bar{X}_s - \bar{X}_i - \bar{X})^2$	$df_{si,e} = (n_s - 1)(n_i - 1)$	$MS_{si,e} = \frac{SS_{si,e}}{df_{si,e}}$	$\sigma_{si,e}^2 = MS_{si,e}$

Table 4 present the ANOVA results from the hypothetical one-facet study. The variance components reveal information about how different sources of variability affect the response to a single item. To interpret the magnitude of the estimated variance components, we can take the sum of the variance components, called the total variance, and create percentages of this sum that each estimated variance component accounts for (Shavelson & Webb, 1991). In this case, the variance component for students accounts for 18% of the total variance, suggesting that averaging over all the items, the students in the sample differ in the construct being measured. In this example, students constitute the object of measurement, not error, and as such this variability is desirable. It reflects systematic individual differences in the construct being measured. This variance components is known as the universe score variance. It presents the variance of scores averaged over all conditions of test items, defined by the universe of admissible observations (Strube, 2002).

The variance component for items accounts for 37% of the total variance, suggesting that some items were more difficult than others, averaging over all the students. The residual term accounts for 45% of the total variance. This suggests that there are important sources of variance not accounted for by differences between students, differences in item difficultly, or both. It should be emphasized that in GT, the magnitudes of the estimated variance components are of central importance rather than their statistical significance (Brennan, 2001). Once the variance components have been estimated, the principles of GT are used to determine the allocation of the components for the estimation of two types of error variance: relative and absolute error variances.

Table 4. ANOVA estimates of variance components for the one-facet crossed design example

Source of variation	Sum of squares	Degrees of freedom	Mean squares	Variance component	% of total variance
Students (*s*)	14.16	4	3.54	.34	12
Items (*i*)	21.36	4	5.34	.70	24
Residuals (*si*,*e*)	29.44	16	1.84	1.84	64

6. Relative Error Variance

GT distinguishes between two types of error variance that corresponds to relative decisions and absolute decisions. Relative decisions are decisions about individual differences between students. Absolute decisions are decisions about the absolute level of performance (Shavelson & Webb, 1991; Strube, 2002).

The relative error variance (σ_{δ}^{2}) is of primary concern when researchers are interested in decisions that involve the rank ordering of individuals. In this case, the error sources are limited to the interactions of the individuals with the facet(s) formed by random sampling of the conditions of measurements. This is because interactions involving the object of measurement reflect changes in relative standing across facet levels (Brennan, 2001; Shavelson & Webb, 1991; Strube, 2002). For the one-facet design presented in Table 2, the estimate of the relative error variance can be found by averaging the residual variance over the number of items used in the measurement (Brennan, 2001). Using the estimates obtained in Table 4, the relative error variance estimate is .37. The square root of this index, which is .61, is considered the relative standard error of measurement (Shavelson & Webb, 1991; Strube, 2002).

7. Absolute Error Variance

For some situations, particularly in the areas of criterion-referenced assessment, one may make decisions about whether a student can perform at a prespecified level. In these instances,

the absolute error variance (σ_{Δ}^{2}) is of concern (Brennan, 2001; Shavelson & Webb, 1991). It reflects both information about the rank ordering of students and any differences in the average scores (Shavelson & Webb, 1991). All sources other than the object of measurement are a source of error for absolute decisions (Strube, 2002). As such, in the one facet ($s \times i$) design example, the absolute error variance includes the variance components due to both the item effect and the residual effect averaged over the number of items used in the measurement. Using the estimates obtained in Table 4, the estimate of the absolute error variance is .51. The square root of this index, which is .71, represents the absolute standard error of measurement (Shavelson & Webb, 1991; Strube, 2002).

8. Generalizability Coefficient

The dependability of a measurement procedure is assessed by a generalizability coefficient, an index that is analogous to the CTT's reliability coefficient. It ranges from 0 to 1 with higher values reflecting more dependable measurement procedures. Values approaching 1 indicate that the scores of interest can be differentiated with a high degree of accuracy despite the random fluctuations of the measurement conditions (Alla & Cardinet, 1997; Shavelson & Webb, 1991; Strube, 2002).

The generalizability coefficients are available for both relative error and absolute error. For the case of relative comparisons of observed scores, the corresponding estimate of the generalizability coefficients for the ($s \times i$) design is defined by the following formula (Barennan, 2001; Shavelson & Webb, 1991):

$$\mathrm{E}_{s\delta}^{2} = \frac{\sigma_{s}^{2}}{\sigma_{s}^{2} + \sigma_{\delta}^{2}},$$

If decisions were based on the absolute values of the observed scores, the corresponding estimate of the generalizability coefficient would be as follows (Brennan, 2001; Shavelson & Webb, 1991):

$$\Phi = \frac{\sigma_{s}^{2}}{\sigma_{s}^{2} + \sigma_{\Delta}^{2}},$$

Using the values of the one-facet design presented in Table 4, the estimates of the generalizability coefficients are $\mathrm{E}_{s\delta}^{2} = .48$ and $\Phi = .40$.

As indicated earlier, GT provides a framework for examining the dependability of measurement procedures. Performing a generalizability analysis to pinpoint the sources of measurement error allows the researcher to determine how many conditions of each facet are needed (e.g., number of items, number of occasions) to obtain an optimal level of generalizability (Marcoulides & Goldstein, 1990). The items, in the design presented in Table 2, represent a source of measurement error. Thus, increasing the number of items in the

measurement procedure increase the generalizability coefficients because an increase in the number of items would lead to a decrease in the estimates of both relative and absolute errors.

9. Generalizability Studies and Decision Studies

There are two stages in the application of GT. The first is the generalizability (G) study carried out by the developer of the measurement procedure. It is designed to provide information about the sources of variability (i.e., facets of the universe) that influence the generalizability of observations. On the basis of this information, various modifications of the initial G study design can be analyzed in the second stage called a D study (Shavelson & Webb, 1981, 1991).

The purpose of D study is to determine the most appropriate measurement procedure for a particular situation in which the information obtained in the G study will be used. For example, if the results of a G study show that some sources of error are small such as the error attributable to items, then the researcher may elect a measurement procedure that reduces the number of levels of the facet (i.e., number of items) or elect to change the actual data collection design, say, for example, from a crossed ($s \times i$) design in which every student was administered the same sample of test items to design in which the items are nested within students (i.e., each student is administered a different sample of items). Alternatively, if the results of a G study show that some sources of error in the design are large, the researcher may need to increase the levels of that facet in order to obtain an acceptable level of generalizability. In general, the D study addressed the question: What should be done differently if you are going to rely on this measurement procedure for making future decisions? In the case where no changes should be made, the G study acts as the D study (e.g., employs the same sample of items used in the initial G study) (Brennan, 2001; Shavelson & Webb, 1991).

10. Advantages and Disadvantages of GT

The foregoing discussion indicates that GT has four main advantages (Shavelson, & Webb, 1991; Thompson, 2003):

1. It can be used to asses multiple sources of error in a specific measurement situation. Compared to CTT, GT provides more reasonable estimates of dependability in circumstances where multiple sources of error are present. A researcher with a measure consisting of several items done by different raters to measure one construct in respondents is seriously inflating reliability when only looking at the separate reliability estimates.

2. It informs the researcher about the magnitude of the types of errors, so that decisions concerning whether error magnitudes are within acceptable ranges can be applied to future studies. Hence, a desired level of generalizability can be obtained in those studies. If errors are considered larger than desired, researchers can use GT to more precisely evaluate methods for reducing the error in future studies. Once a researcher knows which sources of errors, he or she can identify those error sources which are mutable, plan ways to reduce them, and design the most optimal future measurement situations.

3. GT distinguishes between relative and absolute decisions. Relative decisions are those used to compare individuals to each other. Absolute decisions are those based on an individual's absolute level of performance.

4. The possibility of treating sources of error as fixed or random allows the researcher the flexibility to consider measurement errors that will generalize to either a universe of facets or to only a fixed number of facets.

Despite the power of GT, it has a number of disadvantages or limitations (Shavelson, & Webb, 1991; Strube, 2002; Webb, Rowley, & Shavelson, 1988). These include:

1. It has not been readily accessible to researchers because of its technical development and presentation.

2. It may result in coefficients that are tethered to the particular sample used in conducting the study. The ability to generalize the findings to another population is limited, particularly when levels of facets of the larger population are not incorporated in the study's sample.

3. Its application requires substantial effort in the design, data collection, and analysis. Estimation of error sources requires that the design of the generalizability study includes all relevant facets and that enough data are collected on each facet to accurately estimates its error variance.

4. By taking multiple sources of error into account, the generalizability coefficients tend to be lower than reliability coefficients from CTT.

11. Conclusion

GT, as a measurement theory, provides a framework for examining the dependability of almost any type of measurement procedure in almost any type of design. It extends CTT in several important ways. First, it recognizes multiple sources of measurement error, estimates each source separately, and provides a mechanism for optimizing the reliability. Second, although GT provides a reliability coefficient, called a "generalizability coefficient", the theory focuses on variance components that index the magnitude of each source of error. Third, GT distinguishes between relative decisions, where interest focuses on the dependability of the differences among individuals, and absolute decisions, where scores are themselves interpretable without reference to others' performance. Fourth, GT distinguishes between generalizability (G) studies and decision (D) studies. G studies estimate the magnitude of as many potential sources of measurement error as possible. D studies use information from a G study to design a measurement that minimizes error for particular purpose. Although GT provides the most flexible and practical approach for examining the dependability of the measurement procedures, it has not been readily accessible to researchers because of its technical development. It should be emphasized that GT is a continuous work in progress, where there are some important theoretical and statistical topics that clearly need to be addressed more fully, and that there are potential areas of application where the theory has been largely not used.

References

Allal, L., & Cardinet, J. (1997). Generalizability theory. In J.P. Keeres (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd, pp. 737- 741). Cambridge, United Kindom: Cambridge University.

Brennan, R.L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*, 27-34. http://dx.doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brennan, R.L. (1997). A perspective on the history of generalizability theory. *Educational Measurement*: *Issues and Practice, 16*, 14-20. http://dx.doi.org/10.1111/j.1745-3992.1997.tb00604.x

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer.

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E.C. Rhodes (Eds.), *the marks of examiners* (pp. 245- 314). London: Macmillan.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Application to educational measurement. *Journal of Educational Measurement, 13*, 119-135. http://dx.doi.org/10.1111/j.1745-3984.1976.tb00003.x

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extensions of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, *18*, 183-204. http://dx.doi.org/10.1111/j.1745-3984.1981.tb00852.x

Crick, J.E., & Brennan, R.L. (1983). Manual for GENOVA: a generalized analysis of variance system (American College Testing Technical Bukketin No. 43). Iowa City, IA: American College Testing.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Belmont, CA: Wadsworth Group/ Thomson Learning.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability scores and profiles*. New York: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberization of reliability theory. *British Journal of statistical Psychology, 16*, 137-163. http://dx.doi.org/10.1111/j.2044-8317.1963.tb00206.x

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407- 424. http://dx.doi.org/10.1007/BF02288803

Feldt, L. S., & Brennan, R.L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 127-144). New York: Macmillan.

Finlayson, D. S. (1951). The reliability of marking essays. *British Journal of educational Psychology, 35*, 143-162.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30*, 395-418.

http://dx.doi.org/10.1007/BF02289531

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160. http://dx.doi.org/10.1007/BF02289270

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325- 336.

Lord, F. M. (1957). Do test of the same length have the same standard errors of measurement? *Educational and Psychological Measurement, 17*, 510-521. http://dx.doi.org/10.1177/001316445501500401

Lord, F. M. (1959). Test of the same length do have the same standard errors of measurement? *Educational and Psychological Measurement, 19*, 233-239. http://dx.doi.org/10.1177/001316445901900208

Marcoulides, G.A., & Goldstein, Z., (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement, 50*, 782- 789. http://dx.doi.org/10.1177/0013164490504004

Shavelson, R.J., & Webb, N.M. (1981). Generalizability theory: 1973 – 1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133- 166. http://dx.doi.org/10.1111/j.2044-8317.1981.tb00625.x

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

Strube, M. J. (2002). Reliability and generalizability theory. In L.G. grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Thompson, B. (2003). A brief introduction to generalizability theory. In B. Thompson (Eds.), *Score reliability: Contemporary thinking on reliability issues* (pp. 43- 58). Thousand Oaks, CA: Sage.

Webb, N.M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Development, 21, 81- 90.