Journal of Agricultural Education
Volume 38, Number 3, pp. 36-42
DOI: 10.5032/jae.1997.03036

ASSESSING RELIABILITY OF MEASUREMENTS WITH GENERALIZABILITY THEORY: AN APPLICATION TO INTER-RATER RELIABILITY

Dawn M. VanLeeuwen, Assistant Professor
New Mexico State University

Abstract

This article introduces the application of Generalizability Theory to assessing the reliability of measurements. Generalizability Theory can be used to assess reliability in the presence of multiple sources of error and also can be applied to assess reliability in the presence of different types of sources of error. In particular, the application of Generalizability Theory to measurements involving multiple raters is considered. An example illustrates its application to assessing reliability in the presence of inter-rater variability.

When reporting results from a study, the need to assess reliability of measurements is generally recognized and measures of reliability are reported. For example, in a recent issue of the Journal of Agricultural Education, articles were consistent in reporting reliability coefficients where appropriate. None of these papers, however, used Generalizability Theory (GT) to obtain the coefficients, even when multiple sources of error existed; all used Classical Reliability coefficients.

Under Classical Theory (CT), various reliability measures exist, but each can consider only a single source of error. Test-retest measures of reliability regard occasion as the source of error; parallel-forms measures of reliability regard the form as the source of error; and internal consistency reliability measures regard items as the source of error (Eason, 1989; Webb, Rowley & Shavelson, 1988).

GT provides a flexible alternative to CT that allows multiple sources of error to be estimated separately (Shavelson, Webb & Rowley, 1989). GT also allows the impact of a variety of different types of sources of error, such as items, occasions, forms, or raters, on the reliability of measurements to be examined within a unified framework. The crux of GT methodology is variance components estimation, usually in random linear models. While GT does provide coefficients that are analogous to CT's reliability coefficients, much more emphasis is placed on examining the magnitudes of the error from the different sources.

The GT literature emphasizes that reliability is a characteristic of the data., not of a given test or instrument (Eason, 1989; and Thompson, 1991, 1992). Each time an instrument is used, reliability of the data obtained should be reported. While this need was recognized in all articles of the example journal issue, there was some inconsistency in interpreting reliability coefficients and some seemed to attribute the reliability to the instrument rather than the data.

Purpose

This paper provides a brief introduction to the Generalizability Theory approach to assessing the reliability of measurements. It also provides an overview of the advantages of GT over CT. An example illustrates the application of GT to a situation involving multiple raters, so that inter-rater reliability is examined using GT. For the example, two reliability coefficients are obtained; the G coefficient for norm-referenced measurements and the phi coefficient for criterion-referenced measurements.

Advantages of Generalizability Theory

GT affords several advantages over CT. The following are among them:

1. GT considers multiple sources of error simultaneously and allows more accurate modelling of the measurement situation than methods modelling only a single source of error. CT considers only single sources of measurement error for relative decisions.

2. GT provides a unified approach to viewing various types of error. The same methodology can be applied whether the source of error is items, occasions, forms, or raters. Thus, GT can consider any of a number of different sources of error either in combination with one another or by themselves.

3. GT provides a unified approach for assessing the reliability of measurements taken for either relative decisions (norm-referenced measures) or absolute decisions (criterion-referenced measures). Relative decisions are based on an individual's ranking within a group rather than on an absolute score. Absolute decisions, on the other hand, are based on an absolute score *with no* comparative reference to the scores of others (Ary, Jacobs & Razavieh, 1996).

4. GT makes no assumptions concerning the overlap of sources of error but simultaneously estimates various sources of error, including interactions (Thompson 1992, 199 1). CT assumes that sources of error overlap and does not consider the possibility that they may interact to create additional measurement error.

5. CT assumes facet effects are zero. For example, if items are the source of error, CT assumes that all items are equally difficult. These assumptions are relaxed under GT. Removing these assumptions allows GT to consider reliability of both relative and absolute decisions. An additional benefit is the conceptual fit of a model that does not require, for example, items to be equally difficult.

Generalizability Theory Basics

While GT provides a "generalizability coefficient" which is analogous to CT's reliability coefficient, it places much more emphasis on the magnitudes of the various sources of error. Additionally, GT considers two types of studies: generalizability (G) studies and decision (D) studies. G studies are designed to estimate as many sources of error as possible, while D studies obtain measurements for a particular purpose. In the example that follows, the G study is designed to allow estimation of variance components for person, item, rater, and all pairwise interactions as well as additional random error. But because D studies only need reliable measurements for decision-making, it may be reasonable to use a D study involving only a single rater. The measurement based on such a design may be reliable even though the D study design would not allow estimation of the variance component for rater or any of the variance components for interactions with rater. Information from G studies is used to design D studies in order to obtain a measure having the desired reliability level for decision-making purposes.

GT is concerned with an object of measurement (usually a person) and how accurately scores permit generalization to the person's behavior in a defined universe of situations. Known sources of error such as test items, testing occasions or raters are called facets, and levels of the facet are called conditions. In the example, rater and item are both facets. Each of four raters evaluated each response to each of the items. The four raters represent four conditions of the facet

rater and the five items represent five conditions of the facet item. Typically, facets are random effects because it is assumed that the testing situation includes either a random sample of conditions for each facet, or the conditions represented are an "exchangeable" subset of conditions. Thus, in the example, the four raters are either a random sample of qualified raters or an exchangeable subset of qualified raters. Often, facets and the object of measurement are factors in a factorial design. Since all these factors are random, GT essentially becomes a variance component estimation problem. Occasionally a facet is fixed. For example, if the four raters used in the study were the only qualified raters of interest, then raters would be a fixed effect. This is a special case discussed in Shavelson and Webb (1991).

In GT, the universe score is the average score for the object of measurement (person) over all combinations of conditions. This universe score is an idealized measurement that cannot possibly be obtained. Instead, a test score is obtained. This test score is an average of a random sampling of the conditions of each facet and is an attempt to estimate the universe score. When behavior is observed through time (i.e., occasion is a facet), an important assumption in GT is that individual differences remain constant over time.

The following section incorporates an example with discussion to illustrate the basics of GT, including the interpretation of the variance components, the computation of G and phi coefficients, and the fundamental notions and roles of G studies and D studies.

Example

The following hypothetical example illustrates the application of GT to assess the magnitudes of variability due to various sources of error and the reliability or generalizability of measurements. For purposes of example, the sample sizes are small (10 persons, 5 items, and 4 raters). It is important to remember that when estimating variance components, more stable estimates result from more replication at each level at which variability occurs. For example, including more people, more items, and more raters will improve estimators of variance components.

The example illustrates the application of GT to examining multiple sources of error including raters. Tools to examine inter-rater reliability are often restricted to pair-wise correlations between raters (Ary, Jacobs & Razavieh, 1996). GT can be applied to examine inter-rater reliability. While the example includes multiple items, if multiple raters were evaluating only a single item, the same concepts and tools would apply. The only difference would be that the analysis would be based on a simpler Analysis of Variance (ANOVA) because items would no longer be a factor in the analysis.

The example is presented as a G study and shows how information obtained in a G study can be used to plan a D study. When the purpose of a G study is to obtain estimates of as many variance components as possible, the G study is as fully crossed as possible. Crossed designs allow separate estimates of interaction and main effects components, while nested designs do not allow separate estimates of the magnitudes of some variance components. The example design is fully crossed because the four raters evaluated the same five items allowing separate estimates of the variance components for the item main effect, the rater main effect, and the item by rater interaction. If, instead, the four raters each evaluated different items, then items would have been nested within rater. While there would be a total of 20 different items in the study, it would not be possible to obtain separate estimates of the item main effect and the item by rater interaction variance components.

The hypothetical data appear in table 1. The ANOVA table and estimates of the variance

Table 1. Example Data

	Rater			
	1	2	3	4
	Item	Item	Item	Item
Person	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5	1 2 3 4 5
1	4 3 2 2 4	4 3 2 2 2	3 3 3 3 2	3 3 3 2 2
2	4 2 3 4 3	3 3 1 2 1	3 3 2 3 2	3 2 3 3 2
3	1 1 0 1 1	0 1 1 0 0	0 0 0 0 0	1 0 0 2 2
4	3 4 0 0 3	0 4 3 1 2	1 3 0 0 3	1 2 0 0 2
5	3 3 2 1 3	2 2 0 0 2	2 2 0 2 1	1 3 1 0 1
6	4 2 0 0 2	1 2 1 1 1	2 4 0 1 0	2 2 0 0 0
7	3 4 1 0 4	1 4 1 0 2	2 3 1 0 4	1 1 1 0 3
8	2 3 3 3 4	2 2 3 3 4	3 4 4 3 3	2 4 1 3 4
9	4 3 4 3 4	2 1 2 2 4	2 2 3 1 3	2 0 3 4 4
10	3 4 1 3 4	3 4 0 3 4	1 2 0 2 4	3 2 1 2 3

components appear in Table 2. Note that a variance component cannot be negative, but estimates of variance components can be negative. If a negative estimate that is small in magnitude occurs, one approach is to simply truncate the negative estimate at zero. Such negative estimates may occur due to random error when the variance component is either small or zero. If a negative estimate has a relatively large magnitude, a problem may exist in the model's specification. For the example, some of the components are small, but none are negative.

Interpreting the variance components is at the heart of GT. For the example data, all variance components associated with rater are relatively small. Components associated with person, error, and the person by item interaction are relatively large, while the component associated with item is moderate.

All variance components affect the reliability of absolute decisions, but only those variance components reflecting an interaction with persons (including the error component) affect the reliability of relative decisions. The reason is relative decisions are really only concerned with relative placings or rankings of individuals. Item main effects (i.e., overall differences in the difficulty of items) have no effect on the rankings; that is, when item main effects exist but no interactions with item exist, the range of absolute scores will differ from item to item but rankings based on each item will not differ. Even rater by item interactions will not affect the relative standing of persons on the test scores. Only error and interactions involving persons, such as the person by item interaction or the person by rater interaction, will cause changes in the rank orderings of individuals.

The reliability of a measurement is the proportion of the total variability (i.e., person variability + random error variability) in the measurement that is due to differences among individuals (person variability). As discussed in the last paragraph, which variance components contribute to the random error variability depend on whether the decision is norm- or criterion-referenced. Thus, GT provides two different reliability coefficients. The G coeffkient is the GT

Table 2. ANOVA Table

Source of Variance	df	sum of Squares	Mean Squares	Estimated Variance Component	% of Variance
Person	9	112.22000	12.46889	0.4992	26.62%
Item	4	42.67000	10.66750	0.2017	10.76%
Rater	3	15.64000	5.21333	0.0872	4.65%
Person*Item	36	87.03000	2.41750	0.4542	24.22%
Person*Rater	27	18.06000	0.66889	0.0136	0.73%
Item*Rater	12	9.41000	0.78417	0.0183	0.98%
Error	108	64.89000	0.60083	0.6008	32.04%

analog to the CT coefficients, which are for norm-referenced or for relative decisions. Thus, for relative decisions, random error variability is the sum of the variance components for error and all components for interaction terms involving persons. The phi coefficient is a reliability coefficient for criterion-referenced or absolute decisions. For absolute decisions, random error variability is the sum of all but the person variance components. Since most measurements are sums or averages across several items or raters, the random error variability must take into account the number of different items or raters used in obtaining the measurement. When calculating the random error variability, the use of multiple items or raters is taken into account by dividing each variance component by the number of levels of each facet that the component reflects.

To illustrate the computation of reliability coefficients, consider the reliability coefficients corresponding to the G study design. Relative decisions are affected by three sources of error, (1) person by item interaction, (2) person by rater interaction, and (3) error. Because there are 5 items, the person by item interaction component is divided by 5; because there are 4 raters, the person by rater interaction component is divided by 4; and because there are 5 items and 4 raters the error component is divided by 20 which is equivalent to dividing first by 5 and then dividing by 4. This yields the following random error variability for relative decisions 0.09084 + 0.0034 + 0.03004 = 0.12428.

Thus, the G coefficient for measurements is

$$\frac{0.4992}{0.4992+0.12428} \approx 0.80.$$

To obtain the phi coefficient, continue the adjustments made above to the interaction and error variance components. In addition, divide the item component by 5, the rater component by 4, and the item by rater interaction component by 20. The random error variability for absolute decisions then becomes:

$$0.04034 + 0.0218 + 0.09084 + 0.0034 + 0.000915 + 0.03004 = 0.187335.$$

And the phi coefficient is

$$\frac{0.4992}{0.4992+0.187335} \approx 0.73.$$

From these computations, greater reliability clearly can be achieved by increasing the numbers of items or raters. Also, when looking at the raw components, it is best if the component for persons

Table 3. Planning a D Study

Source of Variation	Raters 1 Items 1	1 10	1 15	3 5	3 10
Person	0.4992	0.49920	0.499200	0.499200	0.499200
Item	0.2017	0.02017	0.013447	0.040340	0.020170
Rater	0.0872	0.08720	0.087200	0.029067	0.029067
Person*Item	0.4542	0.04542	0.030280	0.090840	0.045420
Person*Rater	0.0136	0.01360	0.013600	0.004533	0.004533
Item*Rater	0.0183	0.00183	0.001220	0.001220	0.000610
Error	0.6008	0.06008	0.040053	0.040053	0.020027
Generalizability Coefficients					
G Coefficient	0.32	0.81	0.86	0.79	0.88
Phi Coefficient	0.27	0.69	0.73	0.71	0.81

is one of the larger components. Constructing a measurement that reflects individual differences more than random error is the goal. Difficulty occurs when the variance component for persons is too small, suggesting that responses differ little between individuals.

Not only does GT provide estimates of measurement reliability that can replace the traditional CT reliability estimates in research, but GT also considers the reliability of measurements for particular decision-making purposes. For example, students are often given placement tests. Since it is desired to place them appropriately, the need for the test to be a reliable measure is obvious. Another familiar example of a D study is the written driver's test that must be taken to obtain a learner's permit. These criterion-referenced tests are used as decision-making tools. The decision is whether or not an individual has enough familiarity with the rules of the road to begin to drive on the road. Estimates of variance components from a G study can be used to plan and design D studies that will produce measurements having the desired reliability. To find the design that provides the desired reliability, G study estimates of variance components are used to compute generalizability coefficients for measurements for different designs. Table 3 computes reliability coefficients for several designs and illustrates how information generated in a G study is used to plan a D study.

In the example, rater (and interactions with rater) contribute little to the variability, so it may not be worth basing the D study measurements on multiple raters. That is, if, in practical terms, administering more items scored by a single rater is cheaper than having fewer items but multiple raters, it may not be worth using multiple raters. Certainly for relative decisions where the G coefficient for the design with one rater and 15 items is 0.86 while the design with three raters and 10 items has a G coefficient of 0.88 there would be little advantage to using a measurement requiring multiple raters.

Conclusion

GT provides a powerful alternative to CT. GT can be applied to obtain reliability coefficients in any situation where the usual CT coefficients

might be used. In situations where appropriate CT methodology exists there is no real disadvantage to using CT. However CT methodology is inadequate for many situations encountered in practice. In particular, CT cannot adequately model situations where multiple sources of error are present. GT can model and estimate multiple sources of error as well as the interactions among the sources of error and should be used whenever multiple sources of error are present. GT also has a clear advantage in some single source of error situations. For example, suppose a single item were evaluated by two raters. Rater is the source of error and inter-rater reliability must be assessed. CT can deal adequately with this situation by basing a reliability coefficient on the correlation between the two raters. But CT does not have a coefficient if there are more than two raters. GT logically extends the methods of CT and can easily handle the estimation of inter-rater reliability whether there are two or more raters. Assessment of the reliability of criterion-referenced tests represents a third situation where GT is recommended over CT. In addition, GT considers both G and D studies. Not only can GT be used to obtain estimates of the reliability in research but GT also provides theory and methods to assist in the construction of reliable measurements for on-going decision-making processes.

References

Ary, D., Jacobs, L., & Razavieh, A. (1996). Introduction to Research in Education (5th ed.). U.S.: Harcourt Brace & Co.

Eason, S. (1989). Why generalizability theory yields better results than classical test theory. Paper presented at the Annual Meeting of the Mid-South Educational Research Association (Little Rock, AR, November 8-10, 1989). (ERIC Document Reproduction Service No. ED 3 14 434)

Shavelson, R., Webb, N., & Rowley, G. (1989, June). Generalizability theory. American Psychologist, 44(6), 922-932.

Shavelson, R., & Webb, N.(1991). Generalizability theory: A primer. Newbury Park, CA: SAGE.

Thompson, B . (1992, January/February). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70(3), 434-438.

Thompson, B. (1991, winter). Review of the book Generalizability theory: A mimer, Educational and Psychological Measurement, 51(4), 1069-1075.

Webb, N., Rowley, G., & Shavelson, R. (1988, July). Using generalizability theory in counseling and development. Measurement and Evaluation in Counseling and Develoument, 2 l(2), 81-90.