SHIKEN: JALT Testing & Evaluation SIG Newsletter. *9* (1) Spring 2005 (p. 12 - 16)

Statistics Corner
Questions and answers about language testing statistics:

Generalizability and Decision Studies

James Dean Brown (University of Hawai'i at Manoa)

QUESTION: Recently, I read a paper in the *2004 JALT Pan-SIG Proceedings* by Lars Molloy and Mika Shimura about situational sensitivity in medium-scale interlanguage pragmatics research. They analyzed how three situationally-sensitive measures (word count, speech act features, and a coded set of actions) vary among Japanese university-age students when initiating twelve scripted complaints in English. They used G- and D-studies to analyze the data. My question is threefold: Basically what are "G-studies" and "D-studies"? How do they differ? When should they be used?

ANSWER: My guess is that most readers of *Shiken* probably have no idea what generalizability theory is, so I will first provide some background material on that topic. I will then proceed to answer your three questions more directly, and finally I will have some concluding thoughts.

Generalizability Theory

The reliability of measures in social sciences research has typically been estimated using a classical theory approach (replete with statistical estimates of reliability based on test-rest, parallel forms, split-half, K-R20, K-R21, Cronbach alpha, etc.). One useful extension of classical theory reliability is called generalizability theory (G theory)1. G theory was pioneered by Cronbach, Rajaratnam, and Gleser (1963). In G theory, reliability becomes an issue of the degree to which we can generalize from one observation to a universe of observations (Cronbach, Gleser, Nanda, & Rajaratnam, 1972, p. 15). Hence, G theory takes the view that the observed score is a *universe score*, and permits generalizing from a specific sample to the universe of interest through the use of clearly defined procedures (Shavelson & Webb, 1981, pp. 133-137). These procedures involve the use of a conceptual framework to estimate sources of measurement error in the measurement context. Analysis of variance (ANOVA) procedures are used to segregate and estimate the variance components associated with whatever facets of measurement the tester wants to examine. The mean squares values found in the ANOVA procedure are used to estimate the components of variance in a potentially multifaceted way, which provides a more comprehensive and detailed explanation of test variance than was ever possible in classical theory reliability (Shavelson & Webb, 1981, p. 133). The variance components are then used to study how various test designs affect the *generalizability coefficient* (also known as the *G coefficient* , which is analogous to a *reliability coefficient*). Test design decisions can then be based on more accurate estimation of the effects of various sources of testing error. Shavelson and Webb (1991) describe G-theory techniques in theoretical terms, and Brennan (1983) describes ways to do G-studies in somewhat more practical terms. The single most comprehensive book on G theory currently in print is Brennan (2001). G theory has been applied in many fields other than language measurement research. Naturally, I cannot include all of those references here.

What Are G-Studies and D-Studies?

The various procedures described in the previous section can be broken up into two distinct stages: the generalizability study (G study) and the decision study (D study).

In the *G-study stage* , the ANOVA techniques are used to estimate components of variance for whatever testing facets are of interest in the study. This involves running the ANOVA, using the resulting mean squares to calculate the estimated variance components (one for each facet and one for each of the possible interactions among facets), and interpreting those variance components. Note that the testing facets of interest to a particular researcher may involve one, two, or more of the following sorts of variables: numbers of items, subtests, task types, passages, testing occasions, raters, etc. All of this is explained in great detail in the books cited above.

1 G theory should not be confused with the *G factor*, which is the *general intelligence factor* that some researchers believe in.

The Molloy and Shimura (2005) study (mentioned in your question at the top of this article) provides an appropriate and innovative use of G theory to explore the relative effects of situations (i.e., prompts in interlanguage pragmatics research) on the results for individuals (the participants being studied) in three one-facet (situations) G studies, one for each of three sets of measures (number of words, number of speech acts, and number of actions). Their interpretations of the three sets of G study variance components are correct and useful as far as they go:

> Roughly, in this study, the variance components estimates for individuals can be interpreted as an estimate of how much the individuals in the study varied in terms of the three measures; that is, roughly this shows simple differences between individuals. The situation variance component estimates how much prompts affect the scores. Roughly, this shows how much situations affect scores. The interaction variance components estimates [sic] the extent to which the relative ranking of individuals changes according to prompts. Basically, this shows the extent to which the scores depended on differing reactions to the 12 complaint-initiation prompts. (p. 19).

This explanation could have been a bit more explicit and clear in terms of prose explanations of the relative magnitude of the variance components, the percentages of variance accounted for in each study, the meaning of the signal-to-noise ratio, and so forth. But on the whole, as I pointed out above, the authors do provide correct and useful interpretations of the three sets of G-study results, as far as they go.

In the *D-study stage* that follows the G-study, the values determined for the estimated variance components are used to further calculate estimates of the effects of various measurement designs on the *dependability* (analogous to *reliability*) of the scores. Decisions about alternative test designs and score uses can be made rationally based on estimates of the error involved in whatever choices are involved. Then, the *dependability coefficients* (analogous to *reliability coefficients*) for these various possible alternatives can be calculated, examined, and compared.

Once again turning to the Molloy and Shimura (2005) example. The researchers present their D study results in the last two columns of their Tables 2, 4, and 6. Unfortunately, they do nothing with those results, not even explaining what they indicate, thereby missing important opportunities to explore and illustrate the effects of numbers of situations on the dependability of their measures. Consider the results (for number of words) reported in their Tables 1 and 2. The authors show D-study variance component estimates for individuals, situations, and the individuals-by-situations interaction of 25.7197, .85747, nd 3.19213, respectively (apparently for the current design with 12 situations). Based on these results, I was able to calculate the values shown in my Table A in about 45 minutes using my spreadsheet program.

Table A *Estimated Generalizability Coefficients for Relative and Absolute Decisions for All Three Measures*

	Number of Words		Number of Speech Acts		Number of Actions	
Situations	Relative Decisions	Absolute Decisions	Relative Decisions	Absolute Decisions	Relative Decisions	Absolute Decisions
1	0.4017	0.3461	0.3671	0.3491	0.3188	0.2961
2	0.5732	0.5142	0.5371	0.5175	0.4834	0.4570
3	0.6682	0.6136	0.6351	0.6167	0.5840	0.5580
4	0.7287	0.6792	0.6988	0.6820	0.6518	0.6273
5	0.7705	0.7258	0.7436	0.7283	0.7006	0.6778
6	0.8011	0.7605	0.7768	0.7629	0.7374	0.7163
7	0.8246	0.7875	0.8024	0.7896	0.7661	0.7465
8	0.8431	0.8089	0.8227	0.8110	0.7892	0.7710
9	0.8580	0.8265	0.8393	0.8284	0.8081	0.7911
10	0.8704	0.8411	0.8530	0.8428	0.8239	0.8080
11	0.8808	0.8534	0.8645	0.8550	0.8373	0.8223
12	***0.8896***	***0.8640***	***0.8744***	***0.8655***	***0.8488***	***0.8347***
13	0.8972	0.8731	0.8829	0.8745	0.8588	0.8454
14	0.9038	0.8811	0.8904	0.8825	0.8676	0.8549
15	0.9097	0.8881	0.8969	0.8894	0.8753	0.8632
20	0.9307	0.9137	0.9207	0.9147	0.9035	0.8938
30	0.9527	0.9408	0.9457	0.9415	0.9335	0.9266
40	0.9641	0.9549	0.9587	0.9555	0.9493	0.9439
50	0.9711	0.9636	0.9667	0.9640	0.9590	0.9546
60	0.9758	0.9695	0.9721	0.9699	0.9656	0.9619
70	0.9792	0.9737	0.9760	0.9741	0.9704	0.9672
80	0.9817	0.9769	0.9789	0.9772	0.9740	0.9711
90	0.9837	0.9794	0.9812	0.9797	0.9768	0.9743
100	0.9853	0.9815	0.9831	0.9817	0.9791	0.9768
150	0.9902	0.9876	0.9886	0.9877	0.9860	0.9844
200	0.9926	0.9906	0.9915	0.9908	0.9894	0.9883

Table A shows the dependability coefficients for various numbers of situations for relative decisions (i.e., norm-referenced decisions) and absolute decisions (i.e., criterion-referenced decisions) for each of the three measures. The table clearly illustrates the effect of increasing the number of situations on the dependability for each measure. Notice that, for 12 situations (in boldfaced italics), the results in Table A match those reported by Molloy and Shimura, which were .88959, .86379, .87493, .86550, .84883, and .83469 for relative and absolute decisions for number of words, number of speech acts, and number of actions, respectively. Notice also, however, that this table shows estimates for other possible numbers of situations, thereby allowing me to explore and illustrate the effect of numbers of situations on the dependability of each of the measures. For instance, the table shows that designing a measure for number of words (columns 2 & 3) with only one situation results in low dependability (.4017 and .3461 for relative and absolute decisions, respectively). However, adding only one more situation increases the dependability considerably to .5732 and .5142, respectively. Notice also, for number of words, that considerable gains in dependability are garnered by adding additional situations all the way up to about 5 or 6 situations, but that thereafter the "bang for the buck" gained by adding situations decreases considerably. If a particular group of researchers is (a) going to be using this measure for relative decisions, (b) has limited time for gathering data, and (c) feels that dependability of .80 is sufficient, they may decide, on the basis of these D-study results, to design their measure with six situations, with which they can reasonably expect to produce dependability of about .8011 if their participants are similar to those studied in Molloy and Shimura.

Columns 4 through 7 provide equivalently useful D-study information for the other two measures. Notice that each of these sets of results is interesting in its own right, but also that comparing the three measures can provide useful insights into the relative number of situations needed to produce such-and-such dependability with each of the three measures. For example, for relative decisions, only six situations are needed for the number of words measure to produce dependability of .80 (see column 2), while seven situations are required for the number of speech acts measure to reach the same level (see column 4), and nine situations are needed for the number of actions measure to reach .80 (see column 6).

In short, using the sort of D-study information shown in Table A, the designer of a measure can decide, in terms of the type of decision (relative or absolute) and any practical considerations, how many situations will be needed to produce the necessary level of dependability for a given measure in a given research project; the researcher can also compare the relative efficiency of several measures. Clearly, such information can be very helpful in making design decisions about research instruments (even if, or especially if, there are two, three, or even more facets in the G study).

How Do G-Studies and D-Studies Differ?

It should be clear from the previous section that the *G study stage* involves using ANOVA procedures to determine and interpret the variance components. Then, the *D study stage* takes over to use those variance components to estimate the effects of various design conditions (numbers of situations, items, subtests, testing occasions, etc.) on dependability (for relative decisions, absolute decisions, or both).

When Should G-Studies and D-Studies Be Used?

Clearly, the G study should be used first; then and only then, a follow-up D study should be applied. To do either without the other makes little sense. Hence, there are no "merits or demerits" involved in using either method of analysis. They should both be used together, and they should be applied sequentially.

Conclusion

Molloy and Shimura (2005) made an excellent contribution to pragmatics research in their article. They provided an appropriate and innovative use of G theory analysis to explore the relative effects of situations for individuals in a one-facet G study for each of three different measures. With the additional D-study results in Table A, researchers (with research participants similar to those used by Molloy and Shimura) will now be able to plan how many situations they will need to include in their studies to achieve whatever dependability they feel is adequate and necessary given their practical constraints.

G theory can be applied to many sorts of measurement problems in language testing, measurement, and research. Anyone interested in pursuing these sorts of research may want to visit www.education.uiowa.edu/casma/computer_programs.htm. At the time of this writing, GENOVA software (in both PC or Mac versions) and manuals are free and downloadable at that website.

Those researchers intrigued by the possibilities inherent in G theory should definitely try to get copies of Shavelson and Webb (1991) and Brennan (1983, 2001), but they may also find a number of language testing studies interesting. Bolus, Hinofotis, and Bailey (1982) were the first to suggest the usefulness of G theory to language testing. Brown (1984) was the first to actually use G theory in language testing (to study the effects of numbers of items and passages on the dependability of an engineering English reading comprehension test). Brown and Bailey (1984) examined the effects of numbers of raters and scoring categories on the dependability of essay ratings. Brown (1988, 1989, 1990a, 1991) applied G theory to studying the effects of numbers of raters and topic types on L1 writing placement test scores. Stansfield and Kenyon (1992) used G theory to investigate the effects of numbers of tests and raters on the dependability of oral interview scores. Brown (1990b, 1993) used G theory to examine score dependability in criterion-referenced tests. Kunnan (1992) also used G-theory to investigate the dependability of a criterion-referenced test. Bachman, Lynch, and Mason (1995) used G-theory to study the effects of test tasks and rater judgments on speaking test dependability. Brown and Ross (1996) investigated the relative contributions of numbers of item types, sections, and tests to the dependability of TOEFL scores. Brown (1999) examined the relative contributions of numbers of items, subtests, languages, and their various interactions to TOEFL score dependability. And finally, Brown and Hudson (2002) explain how to apply G theory to criterion-referenced tests in both domain score approaches and squared-error loss approaches.

References

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12* (2), 239-257.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32,* 245-258.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: American College Testing Program.

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

Brown, J. D. (1984). A norm-referenced engineering reading test. In A. K. Pugh, & J. M. Ulijn (Eds.) *Reading for professional purposes: studies and practices in native and foreign languages* (pp. 213-222). London: Heinemann Educational Books.

Brown, J. D. (1988). 1987 Manoa Writing Placement Examination. *Manoa Writing Board Technical Report #1*. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1989). 1988 Manoa Writing Placement Examination. *Manoa Writing Board Technical Report #2*. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1990a). 1989 Manoa Writing Placement Examination. *Manoa Writing Board Technical Report #5*. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1990b). Short-cut estimates of criterion-referenced test consistency. *Language Testing, 7* (1), 77-97.

Brown, J. D. (1991). 1990 Manoa Writing Placement Examination. *Manoa Writing Board Technical Report #14*. Honolulu, HI: Manoa Writing Program, University of Hawaii at Manoa.

Brown, J. D. (1993). A comprehensive criterion-referenced language testing project. In D. Douglas and C. Chapelle (Eds.) *A New Decade of Language Testing Research* (pp. 163-184). Washington, DC: TESOL.

Brown, J. D. (1999). Relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16* (2), 216-237.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34,* 21-42.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Brown, J. D., & Ross, J. A. (1996). Decision dependability of item types, sections, tests, and the overall TOEFL test battery. In M. Milanovic & N. Saville (Eds.), *Performance Testing , Cognition and Assessment* (pp. 231-265). Cambridge: Cambridge University.

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

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: Wiley.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analysis. *Language Testing, 9* (1), 30-49.

Molloy, H., and Shimura, M. (2005). An examination of situational sensitivity in medium-scale interlanguage pragmatics research. In T Newfields, Y. Ishida, M. Chapman, & M. Fujioka (Eds.) *Proceedings of the May. 22-23, 2004 JALT Pan-SIG Conference* Tokyo: JALT Pan SIG Committee.(p. 16 -32). Available online at www.jalt.org/pansig/2004/HTML/ShimMoll.htm. [accessed 8 Feb. 2005].

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

Stansfield, C. W., & Kenyon, D. M. (1992). Research of the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20,* 347-364.

HTML: http://jalt.org/test/bro_21.htm / **PDF**: http://jalt.org/test/PDF/Brown21.pdf