Generalizability Theory and Its Application to Institutional Research

Paul W. Sturgis, Leslie Marchand, M. David Miller, Wei Xu, and Analia Castiglioni

About the Authors

Paul W. Sturgis and Leslie Marchand are statistical analysts at the University of Central Florida's College of Medicine. M. David Miller is professor of research and evaluation methodology, as well as director of collaborative assessment and program evaluation services in the School of Human Development and Organizational Studies in Education at the University of Florida. Wei Xu is psychometrician at the American Board of Anesthesiology. Analia Castiglioni is professor in the Department of Medical Education, and medical director of the Clinical Skills and Simulation Center, as well as assistant dean for clinical skills and simulation at the University of Central Florida College of Medicine.

Abstract

This article introduces generalizability theory (G-theory) to institutional research and assessment practitioners, and explains how it can be utilized to evaluate the reliability of assessment procedures in order to improve student learning outcomes. The fundamental concepts associated with G-theory are briefly discussed, followed by a discussion of the software needed to conduct a generalizability study (G-study) analysis. The article then presents a case study of a G-study analysis; this case study was conducted in order to evaluate the generalizability and dependability of an exam that third-year medical school students complete. The conclusion discusses several situations that institutional research and assessment practitioners are likely to encounter where G-theory can be used to evaluate and improve their assessment procedures in pursuit of improving student learning outcomes.

Keywords: generalizability theory, decision study, reliability, assessment

The AIR Professional File, Spring 2022

Article 156

https://doi.org/10.34315/apf1562022

Copyright © 2022, Association for Institutional Research

INTRODUCTION

Few institutional research and assessment professionals would argue that analyzing data in support of improving student learning outcomes is not central to their mission. In fact, this "student-focused paradigm for decision support" is explicitly recognized by the Association for Institutional Research's (AIR) statement of aspirational practice (Swing & Ross, 2016, p. 3). One of the ways that institutional research and assessment practitioners can improve student learning outcomes is by learning new skills and data analysis techniques, particularly skills that will allow them to more effectively analyze data, and to explain the results of that analysis to decision-makers.

The purpose of this article is to introduce generalizability theory (G-theory) to a new audience, and to explain how it can be used to improve assessment procedures in pursuit of improving student learning outcomes. This article will first briefly discuss the fundamental concepts associated with G-theory, and then discuss the software necessary to conduct a G-study. The article will then present the results of a G-study that was conducted in order to evaluate the generalizability and dependability of an exam that third-year medical school students complete. Finally, the article concludes with a discussion of how institutional research and assessment practitioners can utilize G-theory to evaluate and improve their assessment procedures in pursuit of improving student learning outcomes.

INTRODUCTION TO G-THEORY

G-theory is an extension of, and builds on, classical test theory (CTT). In CTT, the observed measurement is composed of true measurement and random error (Brennan, 2011; Sawaki, 2012; Teker et al., 2015; Willse, 2012). Stated more formally, in CTT "X = T + E, where X represents an observed score, T represents true score, and E represents error of measurement" (Willse, 2012, p. 150). As an example, a student's score on an exam (X) is equal to their true score (T) plus any errors associated with the exam (E). The error term (E) includes all sources of error, including such things as the day of the exam, the time of the exam, the consistency with which the rater(s) are evaluating the exam, and so on. The primary advantage of G-theory as compared to CTT is that multiple sources of error can be explicitly identified and estimated (Bloch & Norman, 2012; Sawaki, 2012; Teker et al., 2015). To return to our example above, this means that the unique amount of variance that various factors associated with the exam (e.g., the individual case and the number of raters evaluating the exam) can be estimated in a generalizability study (G-study), which of course cannot be done using the CTT framework. When comparing the two approaches, Mushquash and O'Connor (2006, p. 542) stated, "G theory is a more encompassing, informative, and useful alternative."

G-theory also builds on the familiar statistical concept of analysis of variance (ANOVA) (Sawaki, 2012; Teker et al., 2015). In fact, the variance components in a G-study are typically estimated by fitting a random-effects ANOVA model to the data (Sawaki, 2012, pp. 534–535).

Broadly speaking, conducting a G-theory analysis is a two-step process in which first a G-study is conducted, and then a dependability study (D-study) is conducted. As Croker et al. (1988) note, the purpose of the G-study is to "identify important sources of variation in a given set of observations collected under various measurement conditions" (p. 288). In simpler terms, this means that the primary purpose of the G-study is to estimate the variance components associated with the different facets of the study, which would normally be treated as an undifferentiated error term if one were to use the CTT framework. Croker et al. go on to note that the purpose of the D-study is to "obtain information that could then guide the researcher in deciding which measurement conditions should be controlled and how many levels of each condition should be included to obtain adequate generalizability" (p. 288). This means that, if we return to the example discussed above, the purpose of the D-study is to examine such things as how adding a rater that is grading some of the exams, or adding one or more cases, impacts the generalizability of the assessment.

A researcher who is considering conducting a G-study should be familiar with terms such as "facet," "universe score," and "dependability." A facet is defined as "a systematic source of variability that may affect the accuracy of the generalization one makes" (Sawaki, 2012, p. 535). To return to our above example, one of the facets that may be of interest could be the number of raters that we are using to evaluate the assessment. Other examples of facets include the individual exam items (in our example, the case); an exam given on different days/times could be a facet as well. Similarly, a universe score is defined as "the average score a candidate would have obtained across an infinite number of testing [*sic*] under measurement conditions that the investigator is willing to accept as exchangeable with one another" (pp. 534–535). This is, of course, very similar to the "true score" in CTT. Finally, dependability is defined as "the extent to which the generalization one makes about a given candidate's universe score based on an observed test score is accurate" (p. 534). As discussed above, the ultimate goal of a G-study is to determine the dependability of a measurement. In other words, the goal is to answer a research question such as this one: "If student A received a score of 90 percent on an exam, to what extent can we be confident that their 90 percent is an accurate reflection of their knowledge and abilities?"

Another strength of G-theory is that it incorporates the concept of relative and absolute decisions, which are related to the concept of norm-referenced and criterion-referenced testing. In norm-referenced testing, which is associated with the concept of relative decisions, the focus is on "the extent to which candidates are rank-ordered consistently across test tasks, test forms, occasions, and so on" (Sawaki, 2012, p. 534). Similarly, in criterion-referenced testing, which is associated with the concept of absolute decisions, the focus is on "the extent to which candidates are consistently classified into different categories (score or ability levels) across test forms, occasions, test tasks, and so on" (p. 534). The reliability index for relative decisions is typically referred to as the generalizability coefficient ($E\rho 2$). Likewise, the index of dependability (ϕ), which is often called the phi coefficient, is used to make absolute decisions (pp. 534–535).

SOFTWARE FOR CONDUCTING A G-STUDY

Despite the fact that G-theory has been discussed in the literature since the 1970s (Cronbach et al., 1972), for many years it was used infrequently because one needed specialty software in order to conduct a G-study. Readers that are interested in the history of software programs for conducting G-studies, or who are interested in conducting a G-study in a software program other than Statistical Package for the Social Sciences (SPSS) or Statistical Analysis Software (SAS) are encouraged to consult Bloch and Norman (2012), Huebner and Lucht (2019), Mushquash and O'Connor (2006), or Teker et al. (2015) for further information.

Regardless of the software package that will be used to conduct the analysis, the first step in conducting a G-study would be to ensure that your data file is in univariate format. If your data file is in multivariate format, then the VARTOCASES command in SPSS or the PROC TRANSPOSE procedure in SAS can be used to restructure your data file (IBM, 2011; SAS Institute, 2009). Table 1 illustrates the difference between univariate and multivariate data file formats.

Depending on the complexity of the design of the study, a G-study can be conducted in SPSS using the VARCOMP procedure, but the authors would recommend using SAS as discussed in the following section. For example, when using the VARCOMP procedure in SPSS, the highest order interaction term is confounded with residual error (Putka & McCloy, 2008), therefore the VARCOMP procedure obviously cannot be used to estimate the variance component associated with the highest order interaction term. The authors' experience is that, when using SPSS version 25, adding the highest order interaction term to the model using the VARCOMP procedure results in an error and all variance components receive an estimate of "0."

Readers that are interested in conducting a G-study in SPSS using the VARCOMP procedure are encouraged to consult the excellent discussion by Putka and McCloy (2008) for further details. An additional reference would be the SPSS syntax handbook available from within SPSS by selecting the "Help" menu, then selecting "Command Syntax Reference."

Table 1. Multivariate vs. Univariate Format

Multivariate Format			Univariate Format		
Student_ID	**Rater 1 Score**	**Rater 2 Score**	**Student_ID**	**Rater_ID**	**Score**
1	**90**	**95**	**1**	**1**	**90**
2	**80**	**85**	**1**	**2**	**95**
3	**70**	**75**	**2**	**1**	**80**
			2	**2**	**85**

Note: Adapted from Putka & McCloy (2008, p. 1).

CASE STUDY: OBJECTIVE STRUCTURED CLINICAL EXAM ANALYSIS

The purpose of this study was to evaluate the generalizability and dependability of an objective structured clinical exam (OSCE) that third-year medical students at a state university complete. An OSCE involves medical students rotating through a series of timed stations where they perform certain clinical tasks. Each station represents a separate medical case; the required tasks for each case may range from taking a patient history, to performing a physical exam, interpreting diagnostic studies or lab results, counseling a patient, and so forth. OSCEs often include the use of standardized patients (SPs), who are individuals who have been trained to portray patients with the particular signs or symptoms of a medical condition in a consistent manner. In some instances, due to the length of time it takes for all medical students to rotate through all OSCE cases, multiple SPs might be trained for the same case. Student performance at each station is scored using a checklist that is specific to the content of the relevant case. The trained SPs are usually the ones who also serve as raters and who complete the checklist for each student that they interact with or observe. For the purposes of this article, the use of the term "case" implies one station of an OSCE that includes the SP and the medical condition they are portraying. The primary purpose of this project was to determine how much of the variance on the exam was attributable to the student, to the case, and/or to the rater. An additional research question involved determining the overall generalizability of the assessment.

The design of the OSCE used six stations or cases, five raters per case (34 raters in total, meaning that not all raters rated each case), and 117 students. Based on these data, a G-study was conducted using the PROC HPMIXED procedure in SAS.[1]

The following variance components were estimated in this study:

- Student (p)
- Case (c)
- Rater (r(c))
- Student * Case (p * c)
- Student * (Rater: Case), and residual (p * (r:c))

The results of the analysis are summarized in table 2. The table illustrates how the variance components and so forth change as the number of cases increase from six to eleven. As the results demonstrate, the largest variance components were those associated with student * case (p * c), and with student * case nested within rater and the residual (p * (r:c)). It is of course not surprising that a large amount of the variance is attributable to (p * (r:c)), since that includes the residual, which accounts for all unmeasured error. However, the fact that 33.1 percent of the variance is attributable to (p * c) is a promising finding. The variance associated with this component indicates that students are learning different skills across the different cases.

However, more variance is attributable to the rater than is attributable to either the student or the case. This indicates that more of the variation in performance on the OSCE is attributable to the subjective evaluation of the raters than is ideal.

1. See appendix 1 for the SAS syntax used in this analysis.

Table 2. Generalizability and Dependability Study

Third-Year Student OSCE									
Effect	**G-study Variance Component**	**% of Variance**		**Rater = 5, Case = 6**	**Rater = 5, Case = 7**	**Rater = 5, Case = 8**	**Rater = 5, Case = 9**	**Rater = 5, Case = 10**	**Rater = 5, Case = 11**
Student	5.72	11.13		5.72	5.72	5.72	5.72	5.72	5.72
Case	4.27	8.31		0.71	0.61	0.53	0.47	0.43	0.39
Rater (Case)	7.35	14.30		0.25	0.21	0.18	0.16	0.15	0.13
Student * Case	17.03	33.13		2.84	2.43	2.13	1.89	1.7	1.55
Student * Rater (Case) and residual	17.03	33.13		0.57	0.49	0.43	0.38	0.34	0.31
Total	51.40	100.00							
			Relative Error Variance	3.41	2.92	2.55	2.27	2.04	1.86
			Absolute Error Variance	4.36	3.74	3.27	2.91	2.62	2.38
			G Coefficient	0.63	0.66	0.69	0.72	0.74	0.75
			Dependability Index	0.57	0.6	0.64	0.66	0.69	0.71

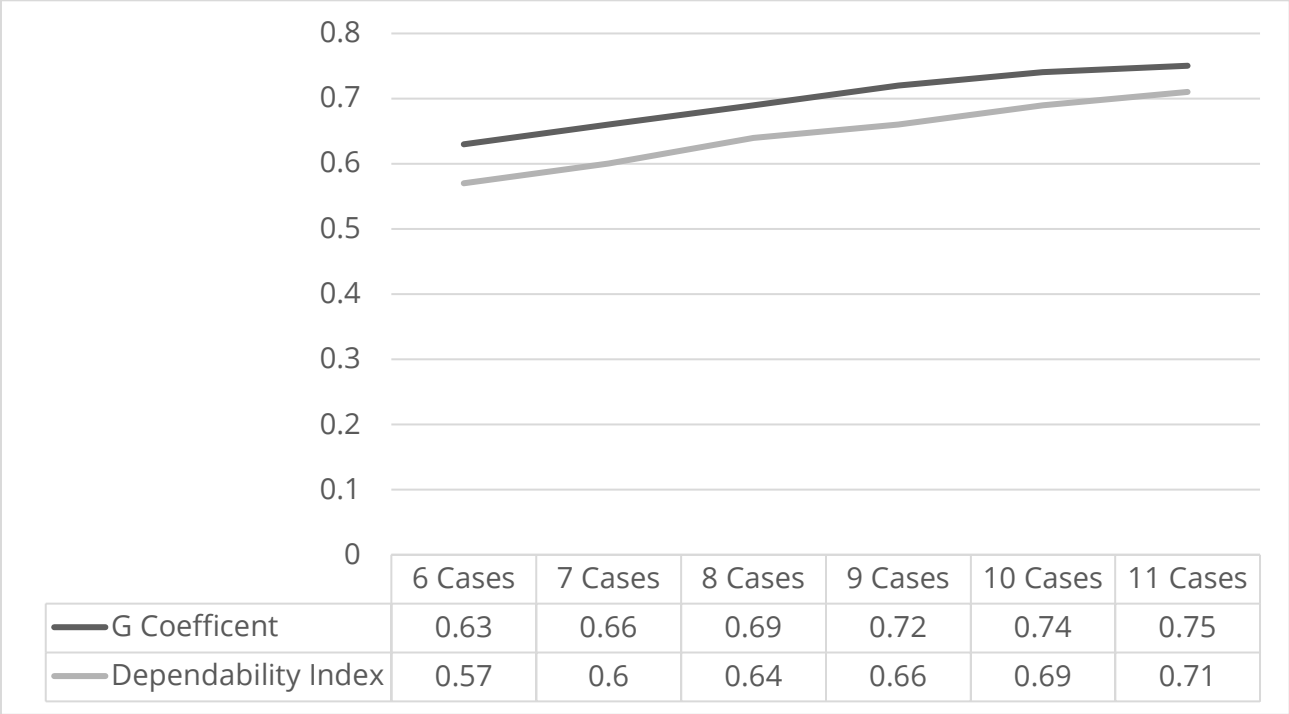
Table 2 also presents the results of the D-study.[2] As the table illustrates, as the OSCE is currently operationalized (five raters and six cases), the generalizability is .63 for relative interpretations[3] and .57 for absolute interpretations[4]. This indicates that the OSCE as operationalized is suitable for making low-stakes decisions, such as estimating student mastery of material in order to assign student grades. The generalizability of the OSCE could be increased to the .7 threshold needed for making high-stakes decisions, such as licensure or certification exams, for this type of assessment (Downing, 2004) by adding three to five additional cases. See figure 1 for additional details on the results of the D-study.

2. The D-study variance components were calculated by dividing the G-study variance component estimates by the number of cases in the study.

3. Relative error variance was calculated by summing all of the D-study variance components that include interactions with the student. The G coefficient was calculated by dividing the student variance component by the sum of the student variance component and the relative error variance.

4. Absolute error variance was calculated by summing all of the D-study variance components. The dependability index was calculated by dividing the student variance component by the sum of the student variance component and the absolute error variance. All of these calculations can easily be done in an Excel spreadsheet.

Figure 1. Generalizability Based on Number of Cases



	6 Cases	7 Cases	8 Cases	9 Cases	10 Cases	11 Cases
G Coefficent	0.63	0.66	0.69	0.72	0.74	0.75
Dependability Index	0.57	0.6	0.64	0.66	0.69	0.71

DISCUSSION AND ADDITIONAL APPLICATIONS

Although G-theory is a niche type of statistical analysis, it has many applications that those who work in institutional research and assessment are likely to encounter. The analysis that is discussed above was designed to determine how much of the variance in an exam was due to the student, the case, the rater, and so on. One of the primary findings was that, although more of the variance than is ideal is due to the raters, the majority of the variance was attributable to factors other than the raters (such as student * case and student * rater(case) and the residual), which suggests that the raters were evaluating the students' performance objectively and reliably. Those that work in institutional research and assessment, particularly those that are associated with health- and medicine-related programs, are often called on to answer these types of research questions, and hopefully the analysis presented above is useful to those researchers and can be used as a blueprint for conducting similar research projects. Those that are interested in additional ways that G-theory concepts can improve assessment procedures in the medical school curriculum are encouraged to consult Bloch and Norman's (2012) excellent discussion on the topic.

Those that work in institutional research and assessment are likely to encounter many research projects where a G-study is useful. For example, many large universities have substantial sections of writing-intensive courses where students are

responding to more than one essay prompt, and where the grading is done by multiple teaching assistants. In this type of situation, G-theory can be used to determine how much of the variance in the students' scores on the essays is due to the teaching assistants, which would help to empirically determine if the teaching assistants are grading the essays in a reliable fashion. Additionally, the amount of variance that is due to the different essay prompts can be determined, which would assist faculty in making decisions about the relative difficulty of the essay prompts.

Another situation where G-theory could be useful to improve student learning is when multiple faculty members are evaluating student portfolios. Similar to the above discussion, G-theory could be used to determine the amount of variance that is due to the faculty members grading the portfolios, which would help to determine if the faculty members are grading the portfolios in a reliable fashion.

The above discussion of the possible uses for a G-theory analysis in institutional research and assessment is certainly not exhaustive. G-theory is undoubtedly a useful analytic procedure, and it can help answer many research questions related to student learning outcomes that institutional research and assessment practitioners are called on to examine.

REFERENCES

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher, 34*(11), 960–992. http://doi.org/10.3109/0142159X.2012.703791

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. https://www.tandfonline.com/doi/abs/10.1080/08957347.2011.532417

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

Croker, L., Llabre, M., & Miller, M. David. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement, 25*(4), 287–299. https://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1988.tb00309.x

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education, 38*:1006–1012. https://pubmed.ncbi.nlm.nih.gov/15327684/

Huebner, A., & Lucht, M. (2019). Generalizability theory in R. *Practical Assessment, Research, and Evaluation, 24*(5). https://doi.org/10.7275/5065-gc10

IBM. (2011). IBM SPSS statistics 20 documentation (pdf). https://www.ibm.com/support/pages/ibm-spss-statistics-20-documentation?msclkid=13863f66b4f211ecbf65cc95f02d060c

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*(3), 542–547. https://link.springer.com/article/10.3758/BF03192810

Putka, D. J., & McCloy, R. A. (2008). Estimating variance components in SPSS and SAS: An annotated reference guide. Human Resources Research Organization.

Salkind, N. J., Ed. (2010). *Encyclopedia of research design*. SAGE.

SAS Institute. (2009). *SAS/STAT ® 9.2: User's guide*, 2nd edition. https://support.sas.com/documentation/cdl/en/statug/63033/PDF/default/statug.pdf

Sawaki, Y. (2012). Generalizability theory. In Salkind, *Encyclopedia of research design*, pp. 534–538. https://dx.doi.org/10.4135/9781412961288

Swing, R. L., & Ross, L. E. (2016). *Statement of aspirational practice for institutional research*. Association for Institutional Research. http://www.airweb.org/aspirationalstatement

Teker, G. T., Guler, N., & Uyanik, G. (2015). Comparing the effectiveness of SPSS and EduG using different designs for generalizability theory. *Educational Sciences: Theory & Practice*, 15(3), 635–645. http://doi.org/10.12738/estp.2015.3.2278

Willse, J. T. (2012). Classical test theory. In Salkind, *Encyclopedia of research design*, pp. 534–538. https://dx.doi.org/10.4135/9781412961288

APPENDIX 1. SAS SYNTAX

```
Data rating;

infile “C:\Users\paul\Desktop\G_study.csv”

delimiter=”,” dsd;

Input ID $ Case $ Rater $ Score;

run;

ods rtf file= “C:\Users\paul\Desktop\G_study.rft”;

PROC HPMIXED method=REML;

CLASS ID Case Rater;

MODEL Score = ;

Random ID Case Rater(Case) ID*Case;

run;

ods rtf close;
```