Generalizability Theory Made Simple(r): An Introductory Primer to G-Studies

Sandra Monteiro, PhD (@monteiro_meded)
Gail M. Sullivan, MD, MPH (@DrMedEd_itor)
Teresa M. Chan, MD, MHPE (@TChanMD)

Imagine the sequence of events: You have designed a 6-station simulation objective structured clinical examination (OSCE) to assess the resuscitation skills of your trainees. At each station trainees are assessed on professionalism, communication, leadership, and technical skills relevant to the scenario. These individual scores are averaged to create a singular score for each station. Cognizant that gender of the trainee may play a role in how raters assess trainees, you wish to examine your OSCE for reliability and sources of variance, to ensure that gender effects are not driving competency decisions. How can you gather validity evidence to support your decisions?

Later, you read an article describing a new OSCE for assessing resident resuscitation skills. Eureka! This is exactly what you need for your program. However, the authors used Generalizability theory to examine validity for the OSCE. What exactly is G-theory and is this a credible approach?

If you can visualize yourself in the above situations and need an introduction or refresher in G-theory, read on. This article explains the basics and provides examples for further study.

Measuring Clinical Competence

In the age of competency-based medical education (CBME), we increasingly make decisions based on our assessment processes, which necessitate ensuring the reliability and validity of our assessments.[1–3] Can we rely on the score to discriminate between trainees based on competence? Can we trust the process? When we create new or modify older assessments to measure *relevant constructs* (ie, specific aspects of clinical competence), we must determine whether our assessment data maintain validity for decision-making. Any assessment can be considered a measurement tool; thus, we can apply measurement principles to examine validity.[3] As validity is not a stable characteristic of any measurement tool, it can be threatened by a multitude of *construct irrelevant* factors that potentially introduce measurement error.[3] Reliability, similarly, is not a stable characteristic of any measurement tool, and is sensitive to changes in context within competency-based assessments.

If you conduct a literature review regarding methods to reduce measurement error, you will find a dizzying multitude of study designs and analysis approaches. Of these approaches, readers are likely familiar with Cronbach's alpha to examine measurement reliability. A calculation of Cronbach's alpha can inform test score reliability, but not whether systematic rater bias has an influence on scores. That is, if trainee gender influenced a rater's assessment of performance, we could not discover this from Cronbach's alpha alone.[3]

A highly useful theory that informs reliability, validity, elements of study design, and data analysis is *Generalizability theory* (G-theory).[3–6] G-theory is a statistical framework for examining, determining, and designing the reliability of various observations or ratings.[3–6]

Using G-theory we can design *Generalizability studies* (G-studies) to better understand the composition of assessment scores (ie, what contributes to the actual score that you get at the end of an OSCE). We can then design *Decision studies* (D-studies) to help predict the reliability of the same data collected under different conditions.

In performance-based assessments we need to consider potential influences on assessment scores, such as rater bias, relative difficulty of items or stations, the rater's or examinee's attention or mood, the abilities of standardized patients, and the overall environment.[3,7] G-theory offers a way to quantify the variance contributed by these factors, which G-theory refers to as *facets*.[3,5,6] Each form of a given facet is called a *condition*.[3] In our example vignette, the trainees are a facet and their gender is a condition (and for this example, we assume only 2 genders). Let's use the above example to review important terminology and concepts, which are supplemented by definitions in TABLE 1.

DOI: http://dx.doi.org/10.4300/JGME-D-19-00464.1

Downloaded from http://meridian.allenpress.com/jgme/article-pdf/11/4/365/2372211/jgme-d-19-00464_1.pdf by guest on 13 July 2024

TABLE 1 Generalizability Theory Terminology[3]

Term	Definition	Comments
G-study	Generalizability studies provide a better understanding of the composition of assessment scores (ie, what contributes to the actual score that you get at the end of an OSCE).	Generalizability studies are used to estimate G-coefficients and describe the influence of various facets within the universe of scores.[3–7] For a review of formulas see references 3 and 12.
In G-Theory we can reinterpret several standard questions[7]:		
	Standard question	Reinterpreted in G-Theory
	What is the interrater reliability of this examination?	To what extent can we *generalize* these scores across raters?
	What is the test-retest reliability?	To what extent can we *generalize* these scores across occasions?
	What is the test-retest/interrater reliability?	To what extent can we *generalize* these scores across both occasions and raters?
D-study	Decision studies are used to ask optimization questions[7] (eg, to help predict the reliability of the same data collected under different conditions).	In our vignette, we computed the reliability of a 6-station OSCE and used the Spearman–Brown formula to estimate how reliable a 4-, 8-, or 10-station OSCE would be.[3] $k = \frac{r'(1-r)}{r(1-r')}$
Validity	Estimates whether an assessment tool finds meaningful, truthful results.	Validity is not a fixed property of the assessment tool, but varies with subjects, setting, purpose, and other factors.
Variance	Commonly referred to as measurement error, variance defines how multiple factors within a measurement context affect a measurement or score.	It is essential that the greatest source of measurement variance is actually due to "true" differences between individuals, rather than other, individually irrelevant factors.
Facets	Facets are variance components and can be considered *fixed* or *random*. Fixed facets are stable, for example, the same raters in every OSCE. Random facets are interchangeable, for example, randomly selected stations for each OSCE.	Ideally, facets are identified a priori. Knowledge of an assessment design informs the researcher whether facets should be considered fixed or random. When indicated, facets can be evaluated as having either fixed or random effects, for the purpose of D-studies. (See TABLE 4).
G-coefficients	Estimates the generalizability of a given aspect of the measurement (eg, interrater reliability).	Example of interrater G-coefficient: helps evaluate how well we might generalize a score from one rater (in one context/station) to another.
Relative G-coefficient	Estimates the generalizability of a given aspect of the measurement but only to the same context. Variance is defined only relative to the data collected in that context.	When in doubt, use absolute error as it is a more conservative estimate.
Absolute G-coefficient	Estimates the generalizability of a given aspect of the measurement that generalizes to other potential contexts. Variance is defined by considering the possible universe of scores yet to be collected.	When in doubt, use absolute error as it is a more conservative estimate.

Example: Application of G-Theory to Assessment

A standard OSCE design is ideal for a G-study because there are repeated measurements of the same construct (ie, clinical skills), much like we might use a measuring tape to measure some lumber several times before making a cut (or better yet, taking the average measurement from 5 people using different measuring tapes). Collecting repeated measures of the same construct has been shown to improve reliability. This is because random variance can cancel out in multiple measurements of the same construct. But there may remain systematic sources of measurement error. For

Downloaded from http://meridian.allenpress.com/jgme/article-pdf/11/4/365/2372211/jgme-d-19-00464_1.pdf by guest on 13 July 2024

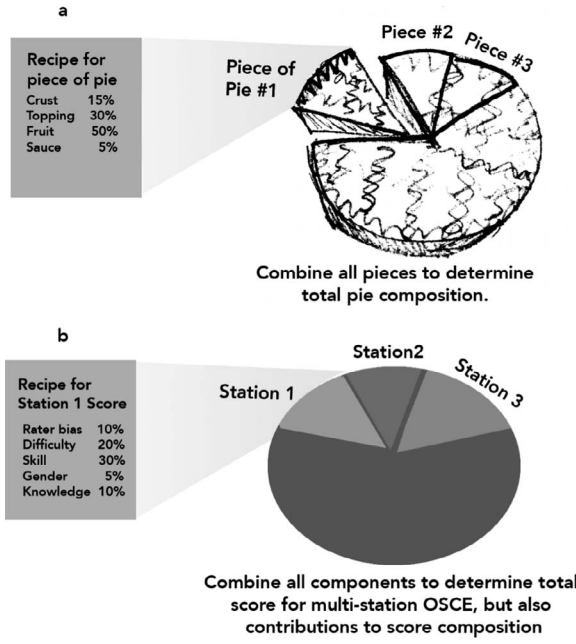

FIGURE
Analogy (a Pie Recipe) for Determining Score Composition for a Total OSCE

example, the rater cognition literature is filled with examples of different kinds of rater bias and its impact on assessment scores and decisions.[8]

In experimental designs the focus is on minimizing error of all types to find true differences between groups.[3,9] In G-studies the goal is to highlight sources of error (called variance) in order to determine if we can trust the critical measurement.[3] In our vignette, all 6 stations' scores are combined or averaged to determine the total score for the entire OSCE.[10] The use of multiple stations and an average OSCE score is a way to deal with measurement error introduced by raters.[8] However, each station score contributes variance toward the total score. Each station is more or less difficult or each rater is more or less lenient and has unique influences on the trainee's score; rather than ignore them, using G-theory we can quantify these contributions. The FIGURE demonstrates this concept. G-theory allows us to develop a recipe. Variance is like the distribution of ingredients in a single slice of pie; sometimes the variance works in an examinee's favor, and sometimes the variance may work against the trainee. Sometimes the variance acts in predictable ways (systematic error) and other times in unpredictable ways (random error).[3] It is essential to consider these aspects of variance in order to fully evaluate reliability, which ultimately sets an upper limit on the validity (accuracy) of the assessment.[3]

In our vignette, we could use G-theory to consider the variance among scores of professionalism, communication, leadership, and technical skills. However, the real power of G-theory is that we can also consider additional sources of variance, such as the gender of the trainee. We could, if we were interested, add in sources of variance like the time of day, different standardized patients, or number of OSCE stations. TABLE 2 shows how we can keep track of various facets in our vignette to evaluate them for different purposes.

To proceed with our G-study, we first identify *all likely sources of variance* or facets in the OSCE for resuscitation skills and determine if they are *fixed* or *random*. In order to ensure we have all the pieces to conduct a G-study, data are best organized as in TABLE 2. As a reminder, we are interested in (1) Determining if we collected reliable assessment data in the resuscitation OSCE, and (2) Whether there is any indication that gender—a factor that should be irrelevant to the evaluation of a clinical skill—contributes any variance to the overall assessment.

Generalizability Theory

> "*Any one measurement from an individual is viewed as a sample from a universe of possible measurements.*"[11]

In G-theory we first define the *universe of scores* and *facets* we wish to generalize from and to. In a G-study, the facets being considered are predetermined to be *fixed* or *random*. We then conduct several *G-studies* to calculate *G-coefficients*. Each calculated G-coefficient evaluates the reliability of a given aspect of the measurement tool, for example, interrater reliability. In D-studies we can evaluate the impact of changing a facet's label, such as from fixed to random. We can use these calculations to make predictions about performance in a similar assessment situation. For example, we can ask how the G-coefficient would be affected by reducing the number of OSCE stations. Or we can ask if multiple raters per station would increase the G-coefficient. Typically, as shown in TABLE 3, multiple stations can improve reliability, but multiple raters per station do not have a big impact.

When considering the reliability of a measurement tool, we can start with a basic formula to describe how different sources of variance or error relate.[3,6]

$$\text{Reliability} = \frac{\text{Trainee}_{\text{variance}}}{\text{Trainee}_{\text{variance}} + \text{Error}_{\text{variance}}}$$

Using the pie analogy, consider that $\text{Error}_{\text{variance}}$ (the recipe) is composed of multiple components

Downloaded from http://meridian.allenpress.com/jgme/article-pdf/11/4/365/2372211/jgme-d-19-00464_1.pdf by guest on 13 July 2024

TABLE 2
Definition of Facets in the Example Vignette

		Station 1 and Rater 1 Trauma Resuscitation Station				Station 2 and Rater 2 Breaking Bad News Station	
		Competencies					
Genders	**Trainee IDs**	Professionalism	Communication	Leadership	Technical Skill	Professionalism	Communication
Male	1	x	x	x	x	x	x
Male	2	x	x	x	x	x	x
Male	3	x	x	x	x	x	x
Male	4	x	x	x	x	x	x
Male	5	x	x	x	x	x	x
Female	6	x	x	x	x	x	x
Female	7	x	x	x	x	x	x
Female	8	x	x	x	x	x	x
Female	9	x	x	x	x	x	x
Female	10	x	x	x	x	x	x
Female	11	x	x	x	x	x	x

Note: all facets of interest are highlighted in **bold text**.
An "x" demarcates a space where a unique score would be entered based on the parameters set forth within this objective structured clinical examination.

(individual ingredients). The variance may be slightly altered by various error-inducing facets (eg, the measuring cup precision, water purity, altitude of the bakery, etc). These facets would each introduce some element of error in the final composition of the resulting pie.

In our vignette, assuming that the resuscitation OSCE is conducted on one occasion, the facets would be the trainees, gender of the trainees, and stations (which include raters). The facet of trainee is nested in the facet of gender. Raters (nested in stations) are facets of generalization, as we hope to generalize from one observation or score at one station, or recorded by one rater, to another score from a different rater. In this example, trainees are the facet of differentiation as we wish to differentiate between individual trainees based on their skill level as measured by the OSCE. These facets describe the known universe of scores in this study. If the OSCE stations or scenarios are never going to change in future administrations, we can consider the facet of station to be fixed. Similarly, if you foresee the same clinical faculty acting as raters for this OSCE every time, the facet of rater may also be considered fixed. Note that in high-stakes OSCEs both of these facets are random as stations change for test security reasons. In program

TABLE 3
Example of Decision-Study Results[3]

Changing Number of Stations			Changing Number of Raters		
$N_{station}$	N_{rater}	Reliability	$N_{station}$	N_{rater}	Reliability
2	2	0.62	2	2	0.62
4	2	0.74	2	4	0.66
6	2	0.79	2	6	0.68
8	2	0.82	2	8	0.69

Note: This demonstrates that changing the number of stations will increase reliability, or reproducibility, and changing the number of raters does not.

TABLE 4
Study Designs for Various Goals[3]

Forms of Reliability	Facets	
	Gender	Station/Rater
	Trainee (Nested in Above)	Competencies (Nested in Above)
Interrater	D	G/R
Test-retest	D	G/F
Overall test	D	G/R
Hawks versus doves	G/R	D

Abbreviations: D, differentiation; G, generalization; R, random; F, fixed.

Downloaded from http://meridian.allenpress.com/jgme/article-pdf/11/4/365/2372211/jgme-d-19-00464_1.pdf by guest on 13 July 2024

BOX Articles Using Generalizability Study Design to Examine Test Properties

- Lang VJ, Berman NB, Bronander K, Harrell H, Hingle S, Holthouser A, et al. Validity evidence for a brief online key features examination in the internal medicine clerkship. *Acad Med*. 2019;94(2):259–266. doi:10.1097/ACM.0000000000002506.
- Monteiro S, Sibbald D, Coetzee K. i-Assess: evaluating the impact of electronic data capture for OSCE. *Perspect Med Educ*. 20181;7(2):110–119. doi:10.1007/s40037-018-0410-4.
- Lord JA, Zuege DJ, Mackay P, Roze des Ordons A, Jocelyn L. Picking the right tool for the job: a reliability study of 4 assessment tools for central venous catheter insertion. *J Grad Med Educ*. 2019;11(4):422–429.

assessments, faculty also may change as they are typically volunteers and not dedicated assessment staff. Whether a facet is *fixed* or *random* changes what variance components are included in the calculation of the G-coefficient (see TABLE 4).[3–5]

For any administration of the resuscitation OSCE, the average OSCE score acts as the universe score against which all individual scores are evaluated. G-studies have the same starting point for determining standard deviation, mean square error, and variance as an analysis of variance (ANOVA).[3–5] We start with *factors* in a standard ANOVA, but then continue with the variance components, or *facets* in a G-study. The difference from an ANOVA analysis is that in G-theory we are not as concerned about establishing a significant difference between groups, but rather to determine how error variance is distributed among the various facets. The goal is to extend classical reliability coefficients to describe how much variance is due to the object of measurement, in this case the trainees. Ideally, the greatest source of variance is the trainees themselves, which would indicate individual differences in ability. Large amounts of variance attributed to raters or other facets such as gender are undesirable, as these factors should not influence decisions about clinical competence.

Ideally, G-studies would assess all sources of variance, or error. The limitation for any evaluation of an assessment is the inability to estimate contributions from unknown sources of variance. However, by focusing a light on as many *known important variables*, it is possible to begin to understand what may be missing.

Next time you read an article that includes a G-study, remember that this strategy will help determine whether the largest source of variance was the subjects being tested—which we would expect in an assessment to determine different trainee competence—or due to other factors, such as the person rating the trainee, time of day, number of test situations, or other factors. For reference, we list 3 articles that use G-theory to examine measurement error (BOX). When creating your own assessment programs, consider using G-theory to understand the role of sources of variance, not only to enhance the reliability of your own measurements, but also to add benefit when disseminating your work to others. We look forward to your questions and comments about G-theory and reliability studies.

References

1. Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Acad Med*. 2014;89(5):721–727. doi:10.1097/ACM.0000000000000221.
2. Iobst WF, Sherbino J, Cate OT, Richardson DL, Dath D, Swing SR, et al. Competency-based medical education in postgraduate medical education. *Med Teach*. 2010;32(8):651–656. doi:10.3109/0142159X.2010.500709.
3. Streiner DL, Norman G, Cairney J. *Health measurement scales: a practical guide to their development and use*. Oxford: Oxford University Press; 2015.
4. Brennan R. (Mis)Conceptions about generalizability theory. *Educ Measure Iss Pract*. 2000;19(1):510.
5. Brennan R. Performance assessments from the perspective of generalizability theory. *App Psychol Measure*. 2000;24(4):339–353.
6. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess*. 2003;80(1):99–103. doi:10.1207/S15327752JPA8001_18.
7. Keszei AP, Novak M, Streiner DL. Introduction to health measurement scales. *J Psychosom Res*. 2010;68(4):319–323. doi:10.1016/j.jpsychores.2010.01.006.
8. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. *Med Educ*. 2016;50(5):511–522. doi:10.1111/medu.12973.
9. Cronbach L. Beyond the two disciplines of scientific psychology. *Amer Psycholo*. 1975;30(2).
10. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ*. 1979;13(1):39–54. doi:10.1111/medu.12801.
11. Llabre MM, Ironson GH, Spitzer SB, Gellman MD, Weidler DJ, Schneiderman N. How many blood pressure measurements are enough? An application of generalizability theory to the study of blood pressure reliability. *Psychophysiology*. 1988;25(1):97–106.
12. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE

Downloaded from http://meridian.allenpress.com/jgme/article-pdf/11/4/365/2372211/jgme-d-19-00464_1.pdf by guest on 13 July 2024

Guide No. 68. *Med Teach*. 2012;34(11):960–992. doi:10.3109/0142159X.2012.703791.

Sandra Monteiro, PhD, is Assistant Director of Research, Centre for Simulation Based Learning, and Scientist, McMaster Faculty of Health Sciences Education Research, Innovation and Theory (MERIT) Program, McMaster University, Hamilton, Ontario, Canada; **Gail M. Sullivan, MD, MPH,** is Editor-in-Chief, *Journal of Graduate Medical Education* (*JGME*), and Associate Director for Education, Center on Aging, and Professor of Medicine, University of Connecticut Health Center; and **Teresa M. Chan, MD, MHPE,** is Associate Professor of Emergency Medicine and Adjunct Scientist, MERIT Program, McMaster University, Hamilton, Ontario, Canada, and Associate Editor, *JGME*.

Corresponding author: Teresa M. Chan, MD, MHPE, Hamilton General Hospital, McMaster Clinic, Room 255, 237 Barton Street E, Hamilton, ON L8L 2X2 Canada, 905.521.2100 ext 76207, teresa.chan@medportal.ca

Downloaded from http://meridian.allenpress.com/jgme/article-pdf/11/4/365/2372211/jgme-d-19-00464_1.pdf by guest on 13 July 2024