Moving beyond Cronbach's Alpha and Inter-Rater Reliability: A Primer on Generalizability Theory for Pharmacy Education

Michael J. Peeters, PharmD, PhD
University of Toledo College of Pharmacy & Pharmaceutical Sciences

ABSTRACT

Background: *When available, empirical evidence should help guide decision-making. Following each administration of a learning assessment, data becomes available for analysis. For learning assessments, Kane's Framework for Validation can helpfully categorize evidence by inference (i.e., scoring, generalization, extrapolation, implications). Especially for test-scores used within a high-stakes setting, generalization evidence is critical. While reporting Cronbach's alpha, inter-rater reliability, and other reliability coefficients for a single measurement error are somewhat common in pharmacy education, dealing with multiple concurrent sources of measurement error within complex learning assessments is not. Performance-based assessments (e.g., OSCEs) that use raters, are inherently complex learning assessments.* ***Primer:*** *Generalizability Theory (G-Theory) can account for multiple sources of measurement error. G-Theory is a powerful tool that can provide a composite reliability (i.e., generalization evidence) for more complex learning assessments, including performance-based assessments. It can also help educators explore ways to make a learning assessment more rigorous if needed, as well as suggest ways to better allocate resources (e.g., staffing, space, fiscal). A brief review of G-Theory is discussed herein focused on pharmacy education.* ***Moving Forward:*** *G-Theory has been common and useful in medical education, though has been used rarely in pharmacy education. Given the similarities in assessment methods among health-professions, G-Theory should prove helpful in pharmacy education as well. Within this Journal and accompanying this Idea Paper, there are multiple reports that demonstrate use of G-Theory in pharmacy education.*

Keywords: pharmacy education, reliability, validation, generalizability theory

Why Is This Innovation Important?

Decades ago, a drive to base decisions in healthcare on empirical studies started in medicine (i.e., Evidence-Based Medicine[1]). After initial success, it moved further afield with evidence-based decision-making,[2] evidence-based leadership,[3] evidence-based management,[4] and evidence-based policy development[5]. The idea of basing decision-making on empirical evidence has become more widespread. Similarly, using evidence for decision-making has spread to education[6] and more specifically to health-professions education (e.g., Best Evidence in Medical Education,[7] Best Evidence in Pharmacy Education[8]). Thus, it may be assumed that learned educators and administrators will aim to base their decisions on empirical evidence whenever possible.

In reality this assumption may not always be the case, and so accreditors such as the Accreditation Council for Pharmacy Education have emphasized cultivating 'a culture of assessment'.[9,10] As Banta and Blaich highlight in *Closing the Assessment Loop*,[11] "the goal of assessment is not just to gather evidence, after all, but to make evidence-informed changes." Similarly, evidence should be a basis for developing and revising learning assessments whenever possible. This is especially so for high-stakes decision-making that involves use of test scores from a learning assessment (e.g., NAPLEX score for pharmacist licensing).[12]

Corresponding author: Michael J. Peeters, PharmD, PhD, FCCP, BCPS; University of Toledo College of Pharmacy & Pharmaceutical Sciences
Email: michael.peeters@utoledo.edu

The process of gathering and building evidence towards use and interpretation of test scores is termed validation. A framework for validation outlined in the influential text *Educational Measurement* focuses on four inferences for scores from a learning assessment—*scoring, generalization, extrapolation* and *implications*.[13] Kane further described this Framework in a 2013 series of articles.[14] There are medical education[15,16] and pharmacy education[17] primers of this important framework. While awareness of validation's importance appears lacking in medical education,[15] the small amount of scholarship is ahead of pharmacy education. Within this Idea Paper and accompanying examples, an aim of this series is to raise pharmacy educators' awareness of validation for learning assessments—to spurn educators to investigate and generate evidence for uses and interpretations of test-scores from their learning assessments (especially their high-stakes testing).

Of note, various sources of evidence for validation can be gathered quantitatively and/or qualitatively; the approach should fit the circumstance. Both approaches can have strengths, when used in a timely and appropriate manner. That said, the remainder of this article is focused on one quantitative approach, and not because it is always better than a qualitative approach, but because it has the advantages of: 1) casting a large, wide net on many participants, 2) summarizing findings from many participants into a few indices (e.g., mean plus/minus standard deviation for test-scores), and 3) quantifying the rigor (quality) of scores from a learning assessment. Additionally, while assessments can include instruments and processes beyond learning (e.g., Doctor of Pharmacy selection/admissions), this article is focused on

DOI: https://doi.org/10.24926/iip.v12i1.2131

assessments of Doctor of Pharmacy (PharmD) students' learning.

What Has Been Learned So Far?

Overview of Validation. Reviews of Kane's Framework for Validation have appeared in multiple health-professions' education.[15-17] The first column of Table 1 provides a brief overview of this framework. This involves transitions from *scoring* to *generalization* to *extrapolation* to *implications* and so transitions from the narrowest to the broadest inferences. An analogy is with using an internet-based map (e.g., Google Maps, MapQuest) to find a location. From this map, you can zoom in and you can zoom out to answer various questions. *Generalization* would first look at "where is this location within this city?" *Scoring* would be to zoom in, such as obtaining directions on "how do I get there?" Alternatively, *extrapolation* would be zooming out from the city and looking at a larger picture, such as a location beyond your city and even into adjacent counties, answering "How far away is this location from me?" *Implications* comes with further zooming out, such as with a location in another state, and considering "Is it worth the travel or can I just get something comparable closer to me?

Table 1. Overview of Kane's Framework for Validation

Kane' Framework for Validation	Focuses on	Example in an OSCE
Scoring	translating an observed performance into an observed score	Scoring for one individual OSCE station
Generalization	Generating and examining the total-scores from an entire exam	The total-score for an entire OSCE (over multiple stations)
Extrapolation	examining the total-score in relation to other real-world performances	An OSCE total-score's relationship to performance on APPEs and/or pharmacist licensing
Implications	exploring consequences of the test, including standard-setting	The passing score for an OSCE, and identification of who will need to remediate

OSCE=objective structured clinical examination. APPE=Advanced Pharmacy Practice Experience

Similar to the analogy description above, it often makes sense to begin with generating evidence for the *generalization* inference. For the generalization inference, a student's performance on an entire test as a whole is evaluated (i.e., a student's final or total-score on a test; a test-score we would like to generalize to other test-takers). Evidence for this inference mainly comes from two sources. First, during test development, initial blueprinting of content or tasks should be indicative of desired course or programmatic outcomes. Second, after administering a learning assessment to students, reliability should be evaluated. That is, the test should be fair and statistically-discriminate among test-takers. This evaluation of reliability can be done after *any* test administration (including multiple-choice, long-answer or performance-based assessments),[18] and should be done after *every* test administration.[19] (Although reliability can be analyzed for many types of questions in learning assessments, in pharmacy education reliability is most often only described as internal consistency using Cronbach's alpha or KR-20 coefficient.[20])

Next, we can zoom in to an individual item's scoring. At the smallest (most zoomed-in) level, evidence for the *scoring* inference looks at translating an observed performance into an observed score. It is focused on whether the scoring criteria were appropriate, as well as whether criteria were applied accurately and consistently by a grader or raters (e.g., grading a long-answer question on an exam, machine-scoring for a multiple-choice exam question, rubric use by multiple raters). Because some methods for scoring a test are straightforward and well-documented (e.g., multiple-choice questions, true-false, and possibly a traditional objective structured clinical examination [OSCE]), if adequate generalization can be demonstrated, then scoring can often be assumed to also be adequate.[17] If generalization is questionable, scoring should next be evaluated.[17] Thus, based on the adequacy of generalization, limited assessment resources may be put to better use with exploring the extrapolation and/or implications inferences.

Evidence for the *extrapolation* inference zooms out from generalization and extends from a test-score to the level of real-world performance. This shifts beyond reliability and should

DOI: https://doi.org/10.24926/iip.v12i1.2131

focus on statistical associations (e.g., correlations, regressions) with other measures of success. Extrapolation evidence can be of higher or lower quality. Looking at the measure of success—is it consequential? And, is it being measured inside or outside the same classroom as the learning assessment? Higher quality extrapolation evidence involves measures of success in the real-world *outside* of the classroom. For instance, how does this OSCE of clinical skills for PharmD students correlate with their performance on Advanced Pharmacy Practice Experiences (APPEs) or with these students' successful pharmacist licensure? Meanwhile, evidence of more limited quality describes a relationship between scores from two learning assessments *inside* that same classroom. For example, between scores from an OSCE and a series of written SOAP (Subjective/Objective/Assessment/Plan) notes in the same course. Thus, improvement in quality of extrapolation evidence attempts to examine the degree that a consequential learning assessment is associated with other measures of success that identify real-world behaviors or consequences; better evidence for extrapolation goes beyond the classroom.

Often taking much more time to unfold and then to gather, evidence for *implications* has been zoomed out the furthest and can be more difficult to generate (though is also most important[15,16,21]). This evidence regards the rigor of the decision-rule (e.g., cut-scores for a learning assessment) to inform a decision or action, such as academic progression or graduating from a program, as well as consequences following from the decision-rule for scores of a learning assessment. This level of evidence is not needed for every assessment of student's learning but should be investigated if scores from a learning assessment are being used to make high-stakes decisions, such as progression in a PharmD program or graduation from that program.

In an example, the entirety of Kane's Framework for Validation was described for a high-stakes situation—admission to medical school. In this summary report, Pieris[22] describes validation evidence for all four inferences with the modified personal interview. Notably, all inferences need not be included in a single investigation; in fact, Pieris references a prior investigation[23] for some evidence within his report. Admission interviews are similarly high-stakes situations in pharmacy education, and so one suggestion would be for documentation of validation for interviews at each college/school of pharmacy.

Some Validation Tools. There are a number of tools to analyze and produce evidence for the *scoring* inference, including item-analysis for multiple-choice items, content analysis of scoring rubrics created by experts, as well as intra-rater and inter-rater reliability to evaluate rater consistency. The Rasch Measurement Model could also be used for scoring evidence (and had been used in this Journal Issue's study of PharmD student case presentations[24]). (While the rest of this article focuses on *generalization* evidence and does not discuss *extrapolation* or *implications* evidence further, evidence for those inferences are also important—especially for using scores from learning assessments in a high-stakes situation.) Evidence for both the *scoring* and (especially) the *generalization* inferences can be evaluated from data obtained the first time a learning assessment is administered.

A tool that can help with generating evidence for the *generalization* inference is Generalizability Theory (G-Theory). G-Theory is a powerful tool in educational assessment and was first described half a century ago.[25,26] Notably, it has had substantial utility in medical education,[27-30] as well as among some investigators in general education.[31-33] (Of note, G-Theory was the foundation for generalization evidence in the modified personal interview example above.[22,23]) Noting the array of contextual and implementation differences among over one-hundred medical schools, Crossley and colleagues concluded that "Generali[z]ability [T]heory is particularly useful in medical education because of the variety and complexity of assessments used and the large number of factors (examinees, assessors, types of assessment, cases and items within cases, contexts, etc.) that impact on scores."[28] Among health-professions, there are similarities with conceptualization of educational assessments. However, use of G-Theory has been minimal in pharmacy education. Internationally, there have only been three investigations that have reported use of G-Theory in the pharmacy education literature,[34-36] and no studies that investigated student performances during a PharmD program. Because pharmacy education shares variety and complexity of assessments similar to medical education, G-Theory should also be particularly useful in pharmacy education regardless of the specific degree program (BScPharm or PharmD).

While more complete reviews of G-Theory can be found in the literature (e.g., medicine[27-30] nursing,[37] psychology,[38] pharmacy[36]), a synopsis is included herein. This synopsis is not meant as an entire guide; Bloch & Norman,[29] as well as Streiner, Norman, & Cairney[30] (along with various G-Theory software manuals[39-41]) can provide helpful guidance to facilitate G-Theory use. Also within this Journal Issue, other articles highlight G-Theory within different pharmacy education applications—with an OSCE[42] (because performance-based assessments may show the G-Theory framework best), multiple knowledge-based exams[43] (because many courses use this traditional assessment structure), and quizzes[44] (because many revised courses have used quizzes at the start of class-time for 'flipped-classroom' or team-based learning pedagogies). These three applications started with evidence analysis for the generalization inference because scoring was straightforward. In a further application of case presentations[24] (a non-OSCE performance-based assessment wherein scoring was not straightforward), the scoring inference needed further evaluation before the generalization inference could be evaluated.

DOI: https://doi.org/10.24926/iip.v12i1.2131

G-Theory Fundamentals. Conceptually, readers have seen statistical analysis with two variables that have been extended to three variables (Table 2). Classical Test Theory (CTT) has traditional reliability indices of internal consistency (commonly reported by Cronbach's alpha) or inter-rater reliability (commonly reported by Cohen's kappa). Meanwhile, G-Theory can be seen as an extension from Classical Test Theory (CTT) that integrates these "separate" indices.

Table 2. Examples of Extensions in Statistics

2 comparators	3 comparators	Description
Correlation	Multivariable regression	From comparing bivariate association, to controlling for multiple (3+) variables[45]
Student's t-test	ANOVA	From comparing two groups, to comparing three or more groups[46]
Winsteps	Facets	From comparing persons versus items, to adding additional facets such as raters[47]
Simple ANOVA	Factorial ANOVA	From comparing main effects (within vs between), to also including interaction effects[48]
CTT's inter-rater reliability (e.g., Cohen's kappa)	CTT's Intraclass correlation	From comparing 2/binary outcomes, to comparing 3+/ordinal outcomes (e.g., ratings)[30]
CTT's internal consistency (e.g., Cronbach's alpha or KR-20)	*Generalizability Theory*	From characterizing one source of error between two test parameters (e.g., students and exam items), to multiple error sources with addition of more test parameters such as raters or testing occasions

ANOVA= analysis of variance, CTT= Classical Test Theory, KR-20= Kuder-Richardson formula #20

As an extension from CTT, G-Theory has also been described as a unifying theory for reliability.[30] Reliability indices of single error sources, such as internal consistency, inter-rater reliability, and test-retest stability are included and integrated within G-Theory. That is, G-Theory is a tool to combine multiple reliability indices into one composite reliability coefficient—one number instead of a handful of separate numbers from various internal consistencies among a handful of assessments, and various inter-rater reliabilities among a handful of raters. In fact, this is one difference between CTT and G-Theory. While CTT posits that there is one error source, and so all error is confounded into that one index of reliability, G-Theory acknowledges and parses out measurement error from multiple concurrent sources.[49,50] By imagining a performance-based assessment, we understand there can be multiple simultaneous sources of measurement error (e.g., raters, items, students). And so, multiple CTT indices of reliability will arise within any performance-based assessment; however, G-Theory will calculate these at the same time, for the same learning assessment, and integrate them into one composite reliability coefficient for the entire learning assessment.[48-50] Similarly, this integration can be done with multiple exams in a course (including a test parameter for multiple exam occasions), or even with multiple quizzes before a course exam (including a test parameter for multiple quiz occasions). Before further description of G-Theory, some foundational terms are in Table 3.

Table 3. Glossary of Terms for Generalizability Theory

Term	Definition
Facet	A set of similar conditions of assessment: a "variable", a test parameter, test sources of variation (e.g., students, items, occasions, raters, stations). A facet is a "factor" in Analysis Of Variance (ANOVA) language.
Fixed facet	A finite facet that is held constant and will *not* be generalized to a universe of infinite versions of this facet in Decision-Studies (D-Studies) (e.g., number of OSCE weeks or number of quizzes, if it is maxed out and cannot meaningfully change).
Random facet	A facet with many versions; a facet to generalize/extrapolate in D-Studies
Levels	Levels is ANOVA language, with each facet/factor having multiple configurations (e.g., item scored with 4-levels; 1, 2, 3, or 4; one, two, or three raters; 10 or 15 stations in an OSCE; 50 or 100 items on an exam).
G-Study	Generalizability Study: Initial analysis of data for variance components from different facets in the specified G-Theory design and discriminate contribution to score variance from different facets and interactions of facets.
D-Studies	Decision Studies: Extensions from a G-Study that use its analyzed score variance to examine "what if" situations for impact on reliability, to help decide on modifications to the next testing iteration (e.g., What if there were 3 raters instead 2? What if there was 1 rater instead of 2? What if there were 10 stations instead of 6?).

OSCE=objective structured clinical examination

DOI: https://doi.org/10.24926/iip.v12i1.2131

G-Theory Sources of Measurement Error. CTT is based on a model of 'true-score equals test-score plus error'. Different CTT reliability indices can reflect variability between test-scores and one other test parameter. G-Theory is an extension from CTT.[29,30,49] Each CTT reliability coefficient has only one error term, and so all error is within it. With G-theory, aside from a presumed variance in students' test scores because of differences in those students' ability, score variance can come from multiple other sources, and these are sources are termed error. For instance, CTT's internal consistency (e.g., commonly reported with Cronbach's alpha or KR-20) describes the variability of *scores* for each rater across multiple *items*. CTT's inter-rater reliability (e.g. reported with an intraclass correlation) describes *score* variability between *raters* on the same item. Furthermore, CTT's test-retest reliability can describe *score* variability with a learning assessment administered over multiple *occasions* (like multiple quizzes before an exam). [50] Meanwhile, G-Theory can simultaneously parse out and attribute the contribution to overall-error of multiple measurement error sources.[29,30,49,50]

G-Study Variance Components. As opposed to 'variables' in traditional studies, test parameter causing potential variance in test-scores are termed 'facets' in G-Theory. Facets, such as *students*, *items*, *raters*, *stations* and *occasions*, can be concurrent sources of variance in a test's scores. An OSCE is a common learning assessment in many health-professions that can exemplify multiple facets involved in total score variation.

To better express the notion of facets, let us imagine that *students* in a pharmacy education program have completed a 16-station OSCE, with eight *stations* each *week* for two weeks. At each station, two *raters* independently score each student performance, with raters using a rubric with multiple *items*. Multiple sources of variation within total-scores are notable herein (italicized above).

A G-Study is the initial analysis of data to obtain a reliability (g-coefficient), as well as variance from different facets. Variance in total-scores will not only come from differences among students' ability, but will also emerge from items, raters, stations, and weeks/occasions. If these facets are specified in the design, analysis in the G-Study will indicate the percentage of variance in total-scores that is accounted for by each of the facets, as well as interactions of facets (i.e., variance components). The relative size of measurement error attributed to various facets can help educators to focus more effort on facets with more variance as opposed to facets with less variance.

Multiple concurrent sources of score variance (and measurement error) can easily be seen within a performance-based assessment. There can be inter-rater error, if more than one rater was used, inter-item error, if the rubric used more than one item, and/or inter-task error, if more than one task was evaluated. While CTT would need to analyze these error sources separately, G-Theory can analyze them together. In a G-Study, educators use G-Theory's multi-way repeated-measures analysis of variance to calculate one composite reliability that integrates and summarizes numerous possible CTT indices, including internal consistency, inter-rater reliability, and test-retest reliability.[29,30]

Decision-studies. Following a G-Study analysis, investigators can perform further analyses termed decision-studies (or *D-Studies*). Within CTT, the Spearman-Brown formula may be used for a single exam. Based on the data from an internal consistency analysis, the Spearman-Brown formula can estimate reliability changes through extrapolation to more related items, that is, to understand how many more exam items would be needed to improve the overall exam reliability to an acceptable threshold (like 0.80[12]). In G-Theory, D-Studies are an extension of this concept. In a D-Study, the associated estimate of reliability changes due to an adjustment to one or more of the facets, such as more or fewer raters, stations and/or items in a performance-based assessment.

D-Studies have been highlighted by Streiner et al, "herein lies one of the real strengths of generalizability theory; the potential to make significant gains in reliability within a fixed number of total observations, by optimally distributing the numbers of observations over various sources of error."[30] For example, using D-Studies for an OSCE can help determine the change to reliability associated with altering the number of levels for one or more facets (examples of levels in Table 3) within the learning assessment. In this example, educators could explore what effect changes to the number of raters, number of items used by raters, and/or number of stations would have on reliability. If a reliability threshold of 0.8 is used, how an educator arrives at that 0.8 can be derived from the D-Studies.

Within this Journal Issue, D-Studies are highlighted for the classroom-based applications of multiple exams[43] and quizzes.[44] Within the exam application, the D-Studies table of number of items and number of exams can be helpful for an instructor determining how many of exam and how many items they will have in their course (noting differences among courses in weeks of contact time and amount of time for each class-session). Likewise, D-Studies from the quiz application can help instructors with how many quiz items are "enough" and what weighting an educator might consider allocating to quizzes versus exams in calculating course grades.

Of note in both applications, the specifics of what one educator at one institution may consider "optimal" from trade-offs between facets may differ from another educator's context. Thus, the importance of each educator doing this for their own learning assessments in their educational context can be most insightful and most helpful for them.

Evidence from many empirical studies have demonstrated that increasing the number of levels for each of the four facets in the

DOI: https://doi.org/10.24926/iip.v12i1.2131

previous example (number of raters, number of items used by raters, and/or number of stations) can improve reliability.[51] However, increasing the number of levels for *some* facets in this OSCE example above could contribute more to overall variance than increasing the number of levels for other facets. Using G-Theory, educators can examine the trade-offs of different situations. For instance, by examining the impact on reliability of eight raters in an OSCE, would reliability be better if two raters scored within four stations, or instead, if those same eight raters were dispersed singly to score within eight stations? For expensive, time- and resource-intensive testing (such as an OSCE), this can help with important decision-making evidence for future test iterations to optimize reliability, while balancing rigor (fairness) with available resources (e.g., staffing, space, fiscal).[27,30] Through constructing generalization evidence for validation, optimizing a learning assessment's reliability can bolster its validity.

While G-Theory is one method, it can model numerous assessment designs. Table 4 describes a number of different types of G-Theory design features.

Table 4. G-Theory Design Features

Design Feature	Description
Crossed facet	Every facet is sampled at all levels with one another (e.g., in an OSCE, raters are crossed with student and crossed with stations *if* the same raters score all students within all stations) Outside of research, this is less common.
Nested facet	One or more facets occur only within certain instances of another facet (e.g., in an OSCE, raters are nested in stations and crossed with students *if* different raters score various students in different stations) In educational assessment, this design is very common.
Balanced design	A design that has equal amounts for all facets (e.g., all exam occasions have same number of questions, all stations use same items, all occasions have same number of stations).
Unbalanced design	A design with an unequal number within any facet (e.g., multiple quizzes with different numbers of items on each quiz, multiple exams have different numbers of items on each exam, an OSCE with different number of raters in various stations or different items used by OSCE raters within different stations).
Univariate design	A conventional design with random facets as crossed or nested facets. This type of design is this vast majority of literature.
Multivariate design	This is an alternative to the popular variant of univariate design, wherein one facet is fixed. Only mGENOVA (at the time of this writing) can analyze a multivariate design and has 13 pre-determined designs.

Crossed Versus Nested G-Theory Facets. The most straightforward G-Theory assessment designs have all facets as crossed. That is, all levels of each facet interact with all levels of other facets. For example, if all students are evaluated on five tasks by the instructor, then students are crossed with the instructor. In another example, if all students sit for the same series of seven quizzes, then student is crossed with quizzes. From the standpoint of G-Theory, crossed facets are preferred, as the measurement error can be best partitioned into its variance components. Alternatively, there are nested facets in G-Theory. As opposed to crossed facets interacting with all levels of each other, a nested facet is "inside" of another facet; all levels of one facet do not interact with all levels of another facet. For instance, if *all* students are evaluated by a series of faculty on five tasks such that a different, "random" evaluator scores *some* students on a task, then students are crossed with tasks, while evaluators/faculty are nested within tasks. In another example, *all* students sit for the same series of quizzes, but quizzes have a different number of items on them (e.g., five items on *some* quizzes versus nine items on other quizzes), then students are crossed with quizzes, while items are nested within quizzes.

From the standpoint of G-Theory, nesting is less ideal, as the measurement error for the nested facet cannot be completely isolated from the facet it is nested within. While a nested design is less "clean" from a technical, G-Theory standpoint, it is exceedingly common in actual educational assessments. Even between different classrooms, there are bound to be differences (however perceived as small) in learning assessments.

Balanced Versus Unbalanced G-Theory Designs. Two exams (e.g., a midterm exam and a final exam) are commonly used in various courses and can help illustrate the difference between balanced and unbalanced designs. If the two exams have the same number of items (e.g., both have fifty multiple-choice questions), then it is termed a balanced design. However, most designs in education are unbalanced (e.g., the midterm exam may have forty questions while the final exam has sixty questions. Other examples of unbalanced designs are in Table

DOI: https://doi.org/10.24926/iip.v12i1.2131

4). This is especially so with the variety of situational differences within a performance-based assessment.[27] Within a performance-based assessment, an unbalanced design could include use of, for example, an uneven number of concurrent raters for various stations (e.g., some stations with no raters, other stations with one rater, and other stations with two raters).

Univariate Versus Multivariate G-Theory Designs. Many designs are univariate and involve random facets. Central to a multivariate design is that one of the facets is fixed (e.g., content categories). That is, most facets can be generalized to many levels of those facets (i.e., random). However, if a facet has a limited/set number of categories, it should be fixed in the assessment design. Once a facet is fixed, it cannot be extrapolated in decision-studies for other situations. For example, a univariate design for an OSCE may have random facets for station and occasion. In decision-studies, these facets can be extrapolated to more and fewer stations, as well as varying the number of occasions. On the other hand, if the occasion facet is fixed (e.g., there is a set number of quizzes that can be administered in a module—once per week and this will not likely change) then the occasion facet cannot be extrapolated and only the number of items per quiz can be extrapolated. However, variance can be analyzed within each quiz separately from one another—and so each quiz can be explored independently. Multivariate G-Theory can be seen in the OSCE application,[42] the Quiz application,[43] as well as the Case-Presentation application.[24]

In all of these reports, the variance is separated into each category of the fixed facet (other examples of a fixed facet are in Table 3), and so the reliability for each week, each quiz, or each rubric item can be determined. Table 2 in the OSCE[42] application followed after multivariate G-Theory analysis for a fixed number of weeks. Table 1 in the Quiz[44] application came from multivariate G-Theory analysis for seven different (fixed) quizzes within a short PharmD module. Table 2 in the case-presentation application followed after multivariate G-Theory analysis for a fixed number of items on the case-presentation evaluation rubric.[24] In each of these examples (and most clearly seen in Table 1 of quizzes application[24]), variance components were reported separately for all levels of each fixed facet. This cannot be done using a univariate G-Theory design.

Multivariate G-Theory can be a helpful diagnostic tool to narrow down which station(s), quiz(zes), or rubric item(s) were least helpful from a reliability standpoint. In fact, Brennan (who authored the seminal *Generalizability Theory*[31]) has stated that, "G-Theory is best viewed as a multivariate theory".[52] And so through poor reliability coefficients, this multivariate theory can contribute further diagnostics as to where a problem might have occurred. This should focus more effort on Week 2 in the OSCE application,[42] Quiz 2 in the Quiz application,[44] or the poor performance for Items 3-6 in the Case-Presentation application.[24]

G-Theory Use in Medical Education. While use of G-Theory was limited during the 1980's, it has become much more widely used in medical education.[27,30] In fact, OSCEs are used in high-stakes national licensure for physicians in Australia, Canada, South Korea, Switzerland, Taiwan, the United Kingdom, and the United States,[53] with multiple prominent medical educationalists affirming that use of G-Theory is absolutely required for an OSCE.[29,30,53,54]

The format of OSCEs was introduced first into medical education and has become an important performance-based assessment format in medical and other health-professions education.[55,56] The validity, reliability, feasibility, and educational impact for OSCEs is noteworthy.[55,56] Instead of describing "the" OSCE, inferring that this is a single entity for testing in all situations, it may be better termed as "an" OSCE to describe a format that can be applied for a number of purposes. Different OSCEs can evaluate skills with history-taking, physical examination, surgical procedures, other procedures, teamwork, and communication. Thus, an OSCE can have many concurrent sources of variance that may differ between an OSCE at one institution versus an OSCE at another institution.[57] The OSCEs will not be exactly the same. Each university has their own unique mixture of resources (number of faculty/support staff, faculty expertise, faculty workloads), assessment philosophy, and financial commitments. G-Theory provides a flexible design structure to analyze different OSCEs at different institutions.

That is, OSCEs have multiple complex relationships of performance rating items, multiple raters, different scenarios, and other variables. Currently, G-Theory is considered the best means to characterize reliability, analyze variance components, and optimize that reliability for these multivariable assessments.[29,30,52,53] With each university needing to analyze, evaluate, and validate their own use of learning assessments,[17] G-Theory appears helpful for determining an assessment's reliability. As noted earlier, this is especially important if an OSCE is used for high-stakes testing.[12] In other coursework, G-Theory can help to improve the rigor of course-level assessment.

How Does the Academy Move Forward?

Similar to other health-professions programs, pharmacy curricula are often rigidly, "lock-step" structured. That is, a student must successfully complete all coursework at one level before progressing to more advanced coursework (i.e., many courses are pre-requisites for future courses). Failure of any single course can cause a student to fall out-of-sync until the next offering of that course (often the following year); that student's graduation may be delayed by at least one year. Furthermore, delay in students' progression is required reporting by the Accreditation Council for Pharmacy Education.[58] This accreditor deems it important enough for public disclosure on a college/school of pharmacy's webpage.[58] With this progression need and in this Age of Accountability,[59]

DOI: https://doi.org/10.24926/iip.v12i1.2131

educators have received more scrutiny from other stakeholders (e.g., students, administrators, lawyers, parents) with educators' responsibility for guiding students to attain educational outcomes. Scrutiny from various stakeholders has increased most when a learning assessment can halt a student's progress in a curriculum (and as a result increase that student's tuition costs and time-to-graduation). Thus, delaying students' progression within a PharmD degree program should be seen as a high-stakes situation; testing that is used for decisions to delay progression should meet standards for high-stakes testing. And so, these high-stakes learning assessments need to be conducted scrupulously and fairly. That is, these high-stakes learning assessments should have validation evidence—G-Theory can be an important tool to generate evidence for validation's generalization inference.

As pharmacy education moves forward, the methods being used for high-stakes testing need to advance and improve. Validation of learning assessments requires action at a local level (at each institution specific to their uses and interpretations of scores from their learning assessment). Notably, these methods should become more complex and may need assistance from a specialist. Just as some colleagues may need assistance with statistical analyses in their scholarship, colleagues need to raise their awareness of validation. This increased awareness may include finding psychometric support at their college/school of pharmacy. With G-Theory's substantial evidence, pharmacy education should advance towards its greater use. That said, pharmacy education is not alone in its slow uptake; G-Theory had also in recent years been forwarded as a "new" tool in nursing education.[37]

G-Theory has been employed for three examples in pharmacy education.[34-36] However, routine use and reporting is very sparse. Authors in medical education, appear to have been early adopters of this powerful tool (e.g., in the 1980s[60]). With the first of the three reports in 2006, pharmacy education has been a much later adopter of G-Theory.

One barrier to using G-Theory in the past has been availability and ease of use for computer software.[30] In fact, substantial improvement in the computing power of personal computers over the past few decades has made G-Theory much more approachable.[29] Table 5 lists the multiple standalone G-Theory software programs available. Of note, all were freely available at the time of this writing. Aside from these, major statistical programs (SPSS, SAS, R, Stata) with sufficient programming functions will also allow univariate G-Theory analyses.

Table 5. Available Generalizability Theory Software Programs

Software	Company (Location)	Platform Compatibility	Unbalanced Designs?	Univariate (U) / Multivariate (M)
EduG[39]	IRDP (Neuchatel, Switzerland)	Windows	No	U
G_String[40]	McMaster University (Hamilton, ON, Canada)	Windows & Mac	Yes	U
GENOVA[41]	University of Iowa (Iowa City, IA, USA)	Windows	No	U
urGENOVA[41]		Windows	Yes	U
mGENOVA[41]		Windows	No	M

With every newly-used method, there are accepted ways to report and peer-review.[61] When evaluating a G-Theory report, it is important to appraise the descriptions of facets, the G-Study design, the reliability, and variance components, as well as findings from any D-studies (which often includes a figure summarizing the D-Studies).[62]

Lastly, with the number of associated studies herein,[24,42-44] it is important to emphasize that validation should not be conceptualized as one validation study, but as a series of investigations.[15-17] One investigation may generate evidence for the scoring, generalization, extrapolation, and/or implications inferences. However, no single study needs to gather all inference evidence at once. Depending on the extent of stakes with a learning assessment, the higher the stakes, the larger the need for stronger evidence to support inferences.

Conclusion

As pharmacy education moves forward, the methods being used for high-stakes testing need to improve. Validation of learning assessments requires action at a local level (at each institution, specific to their uses and interpretations of scores from their learning assessment). Notably, with enhanced awareness by pharmacy educators, these more-complex methods may need assistance from a specialist. G-Theory can help educators to generate evidence for the generalization inference of validation for more-complex learning assessments; pharmacy education should advance to using it far more often.

DOI: https://doi.org/10.24926/iip.v12i1.2131

Conflicts of Interest: None
Funding/support: None

Acknowledgements: I would like to thank Drs. Kimberly Schmude and Robin Zavod, for their reviews with various drafts of this article.

References

1. Guyatt G, Rennie D, Meade M, Cook D, eds. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* 3rd ed. New York NY: McGraw-Hill Education; 2015.
2. Brownson RC, Gurney JG, Land GH. Evidence-based decision making in public health. *J Public Health Manag Pract*. 1999; 5:86-97.
3. Scott S, Webber CF. Evidence-based leadership development: The 4L framework. *J Educ Adm*. 2008 Sep 26;46(6):762-76.
4. Pfeffer J, Sutton RI. Evidence-based management. *Harv Bus Rev*. 2006; 84(1):62.
5. Pawson R. *Evidence-Based Policy: A Realist Perspective*. Thousand Oaks, CA: Sage Publications; 2006.
6. Davies P. What is evidence-based education? *Br J Educ Stud*. 1999; 47(2):108-121.
7. Harden RM, Grant J, Buckley G, Hart IR. BEME guide no. 1: Best evidence medical education. *Med Teach.* 1999; 21(6):553-562.
8. Hammer DP, Sauer KA, Fielding DW, Skau KA. White paper on best evidence pharmacy education (BEPE). *Am J Pharm Educ*. 2004; 68(1):Article 24.
9. Piascik P, Bird E. Creating and sustaining a culture of assessment. *Am J Pharm Educ*. 2008; 72(5): Article 97.
10. Accreditation Council for Pharmacy Education. Accreditation Standards and Key Elements for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree ("Standards 2016"). Published February 2015. Available at: https://www.acpe-accredit.org/pdf/Standards2016FINAL.pdf. Accessed 1 Sep 2020.
11. Banta TW, Blaich C. Closing the assessment loop. *Change.* 2010; 43(1):22-27.
12. Peeters MJ, Cor MK. Guidance for high-stakes testing within pharmacy educational assessment. *Curr Pharm Teach Learn*. 2020; 12(1):1-4.
13. Kane MT. Validation. In Brennan RL, ed. *Educational Measurement*. 4th ed. Portsmouth NH: American Council on Education; 2006: 17-64
14. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013; 50(1):1-73.
15. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ.* 2015; 49(6):560-575.
16. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simul*. 2016 Jan;1(1):31.
17. Peeters MJ, Martin BA. Validation of learning assessments: a primer. *Curr Pharm Teach Learn*. 2017; 9(5):925-933.
18. Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE guide no. 66. *Med Teach.* 2012; 34(3):e161-e175.
19. Zibrowski EM, Myers K, Norman G, Goldszmidt MA. Relying on others' reliability: challenges in clinical teaching assessment. *Teach Learn Med.* 2011; 23(1):21-27.
20. Hoover MJ, Jung R, Jacobs DM, Peeters MJ**.** Educational testing validity and reliability in the pharmacy and medical education literature. *Am J Pharm Educ*. 2013; 77(10):Article 213.
21. Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. *Acad Med*. 2016;91(6):785-795.
22. Pieris D. A critical perspective on the modified personal interview. *Perspect Med Educ*. 2019;8(1):33-37.
23. Hanson MD, Woods NN, Martimianakis MA, Rasasingham R, Kulasegaram K. Multiple independent sampling within medical school admission interviewing: an "intermediate approach". *Perspect Med Educ*. 2016;5(5):292-299.
24. Byrd JS, Peeters MJ. Validation evidence for an assessment of clinical case presentations. *Innov Pharm*. 2021; 12(1):Article 18.
25. Rajaratnam N, Cronbach LJ, Gleser GC. Generalizability of stratified-parallel tests. *Psychometrika*. 1965; 30(1):39-56.
26. Cronbach LJ, Gleser GC, Nanda H. Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York, NY: Wiley; 1972.
27. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ.* 2002; 36:972-978.
28. Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, Norcini J. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ.* 2007; 41:926-934.
29. Bloch R, Norman G. Generalizability Theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. *Med Teach.* 2012; 34(11):960-992.
30. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales*. 5th ed. New York, NY: Oxford University Press; 2015.
31. Brennan RL. *Generalizability Theory*. New York, NY; Springer-Verlag; 2010.

32. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Thousand Oaks, CA: Sage Publications; 1991.
33. Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability Theory: A practical guide to study design, implementation, and interpretation. *J Sch Psychol*. 2014; 52(1):13-35.
34. Munoz LQ, O'Byrne C, Pugsley J, Austin Z. Reliability, validity and generalizability of an objective structured clinical examination (OSCE) for assessment of entry-to-practice in pharmacy. *Pharm Educ,* 2004; 5.
35. Peeters MJ, Serres ML, & Gundrum TE. Improving reliability of a residency interview process. *Am J Pharm Educ.* 2013; 77(8):Article 168.
36. Cor MK, Peeters MJ. Using generalizability theory for reliable learning assessments in pharmacy education. *Curr Pharm Teach Learn.* 2015; 7(3):332-341.
37. Prion SK, Gilbert GE, Haerling KA. Generalizability theory: an introduction with application to simulation evaluation. *Clin Sim Nurs*. 2016 Dec 1;12(12):546-54.
38. Vispoel WP, Morris CA, Kilinc M. Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *J Pers Assess*. 2018;100(1):53-67.
39. Swiss Society for Research in Education Working Group. *EduG User Guide*. Neuchatel Switzerland: IRDP 2010. Available from: https://www.irdp.ch/data/secure/1968/document/EduGUserGuide.pdf. Accessed 1 Sep 2020.
40. Bloch R, Norman G. G String IV User Manual (Version 6.1.1), 2011. Available from: http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf. *Accessed 1 Sep 2020.*
41. GENOVA Suite Programs. Available from: https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs. Accessed on 1 Sep 2020.
42. Peeters MJ, Cor MK, Petite SE, Schroeder MN. Validation evidence from Generalizability Theory for an Objective Structured Clinical Examination. *Innov Pharm*. 2021; 12(1):Article 15.
43. Peeters MJ, Cor MK, Boddu SHS, Nesamony J. Validation evidence from Generalizability Theory for a basic-science course: Reliability of course-grades from multiple examinations. *Innov Pharm*. 2021; 12(1):Article 16.
44. Peeters MJ, Cor MK, Maki ED. Validation evidence from Generalizability Theory for a clinical-science module: Improving test reliability with quizzes. *Innov Pharm*. 2021; 12(1):Article 17.
45. Olsen AA, MacLaughlin JE, Harpe SH. Using multiple linear regression in pharmacy education scholarship. *Curr Pharm Teach Learn.* 2020; 12(10):1258-1268.
46. Norman GR, Streiner DL. *Biostatistics: The Bare Essentials*. 3rd ed. Hamilton ON Canada: B.C.Decker Inc; 2008.
47. Linacre JM, Wright BD. Construction of measures from many-facet data. *J Applied Meas*. 2002; 3(4):486-512.
48. Shavelson RJ, Webb NM, Rowley GL. Generalizability Theory. *Am Psychol*. 1989; 44(6):922-932.
49. Peng, S.K.L.P.Z. Classical versus Generalizability Theory of measurement. *Exam Meas. 2007; 4*:009.
50. Brennan RL. Generalizability theory and classical test theory. *Appl Meas Educ.* 2011; 24(1):1-21.
51. van der Vleuten CP, Schuwirth LW. Assessing professional competence: From methods to programmes. *Med Educ*. 2005; 39(3):309-317.
52. Brennan RL. (Mis) Conceptions about Generalizability Theory. *Educational Measurement, Issues and Practice.* 2000; 19(1):5-10.
53. Swanson DB, van der Vleuten CPM. Assessment of clinical skills with standardized patients: state of the art revisited. *Teach Learn Med.* 2013;25(S1):S17-S25.
54. Swanson DB, Clauser BE, Case SM. Clinical skills assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security. *Adv Health Sci Educ Theory Pract.* 1999;4(1):67-106.
55. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide. 81. Part I: an historical and theoretical perspective. *Med Teach.* 2013;35(9):e1437-e1446.
56. Boursicot KAM, Roberts TE, Burdick WP. Structured assessments of clinical competence. In Swanwick T, ed. *Understanding Medical Education: evidence, theory and practice*. Chichester, UK: Wiley Blackwell: 2014:293-304.
57. Hodges B. OSCE! Variations on a theme by Harden. *Med Educ*. 2003 Dec;37(12):1134-40.
58. Accreditation Council for Pharmacy Education. Policies and Procedures for ACPE Accreditation of Professional Degree Programs. Published June 2019. Available at: https://www.acpe-accredit.org/pharmd-program-accreditation. Accessed 1 Sep 2020.
59. DeLuca C. Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action Teach Educ*. 2012; 34(5-6):576-91.
60. Mitchell K, Anderson J. Reliability of holistic scoring for the MCAT essay. *Educ Psychol Meas*. 1986; 46(3):771-775.
61. Hancock GR, Stapleton LM, Mueller MO, eds. *The reviewer's guide to quantitative methods in the social sciences.* New York, NY: Routledge; 2019.
62. Hendrickson A, Yin P. Generalizability Theory. In Hancock GR, Stapleton LM, Mueller MO, eds. *The Reviewer's Guide to Quantitative Methods in the Social Sciences.* New York, NY: Routledge; 2019:123–131.