Validation Evidence using Generalizability Theory for an Objective Structured Clinical Examination

Michael J. Peeters, PharmD, PhD, BCPS[1]; M. Kenneth Cor, PhD[2]; Sarah E. Petite, PharmD, BCPS[1];
Michelle N. Schroeder, PharmD, BCACP, CDE[1]
[1]University of Toledo College of Pharmacy & Pharmaceutical Sciences, Toledo, OH
[2]University of Alberta Faculty of Pharmacy & Pharmaceutical Sciences, Edmonton, AB

ABSTRACT

Objectives: *Performance-based assessments, including objective structured clinical examinations (OSCEs), are essential learning assessments within pharmacy education. Because important educational decisions can follow from performance-based assessment results, pharmacy colleges/schools should demonstrate acceptable rigor in validation of their learning assessments. Though G-Theory has rarely been reported in pharmacy education, it would behoove pharmacy educators to, using G-Theory, produce evidence demonstrating reliability as a part of their OSCE validation process. This investigation demonstrates the use of G-Theory to describes reliability for an OSCE, as well as to show methods for enhancement of the OSCE's reliability.*

Innovation: *To evaluate practice-readiness in the semester before final-year rotations, third-year PharmD students took an OSCE. This OSCE included 14 stations over three weeks. Each week had four or five stations; one or two stations were scored by faculty-raters while three stations required students' written responses. All stations were scored 1-4. For G-Theory analyses, we used G_Strings and then mGENOVA.*

Critical Analysis: *Ninety-seven students completed the OSCE; stations were scored independently. First, univariate G-Theory design of students crossed with stations nested in weeks (p x s:w) was used. The total-score g-coefficient (reliability) for this OSCE was 0.72. Variance components for test parameters were identified. Of note, students accounted for only some OSCE score variation. Second, a multivariate G-Theory design of students crossed with stations (p• x s°) was used. This further analysis revealed which week(s) were weakest for the reliability of test-scores from this learning assessment. Moreover, decision-studies showed how reliability could change depending on the number of stations each week. For a g-coefficient >0.80, seven stations per week were needed. Additionally, targets for improvements were identified.*

Implications: *In test validation, evidence of reliability is vital for the inference of generalization; G-Theory provided this for our OSCE. Results indicated that the reliability of scores was mediocre and could be improved with more stations. Revision of problematic stations could help reliability as well. Within this need for more stations, one practical insight was to administer those stations over multiple weeks/occasions (instead of all stations in one occasion).*

Keywords: objective structured clinical examination, reliability, validation, generalizability theory

DESCRIPTION OF PROBLEM

Assessing clinical competencies is a vital aspect for education in the health-professions; although, building rigorous performance-based assessments can be challenging.[1,2] As opposed to measuring and reporting a single source of measurement error with written examinations (i.e., using internal consistency with Cronbach's alpha or KR-20), a performance-based assessment has multiple sources of measurement error. Thus, an internal consistency coefficient (such as Cronbach's alpha or KR-20), or inter-rater reliability coefficient (such as an intraclass correlation or Cohen's kappa) will not be enough to adequately describe the reliability of total-scores from these more complex assessments.[3]

The Objective Structured Clinical Exam (OSCE) is one popular performance-based assessment method to assess competencies in the healthcare professions by direct observation.[1-3] Compared to other performance-based assessment methods, the OSCE method can more easily control complexity and variables in an exam (i.e., standardization).[2] A typical OSCE structure has students rotate around a series of timed stations, with each station assessing a different skill related to a clinical competency. That said, the OSCE method has the *potential* for stronger reliability and validity support than with other performance-based assessment methods.[2] Developing and implementing an OSCE should not assume nor imply that reliability and validity will be sufficient;[4] these need to be examined, especially for high-stakes testing.[5]

Doctor of Pharmacy (PharmD) programs should have both valid and reliable assessment mechanisms, and evidence for validation is needed. The extent of this evidence will depend on the stakes of a learning assessment's scores in decision-making. The higher the stakes, the more validation evidence that is needed.[5] Generalizability Theory can be used for some validation evidence and standards for its reporting have been discussed.[6]

As a conceptual framework, Kane's Framework for Validation can provide guidance and structure for validation of interpretations from test scores for learning assessments in any PharmD program.[7] Furthermore, reliability of student learning assessments can be justified using Generalizability Theory (G-

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

DOI: https://doi.org/10.24926/iip.v12i1.2110

Theory), which is a widely-accepted psychometric model for quantifying reliability.[3] Of note, G-Theory is especially useful (and some experts would say *essential*) for analyzing a performance-based assessment (e.g., an OSCE).[3,8] An introductory review, as well as other examples in using G-Theory in pharmacy education, are in this issue of the Journal.[9]

Prior studies in pharmacy education have rarely reported use of G-Theory to produce their reliability evidence as part of their validation process.[7] Innovations of this report are demonstrating the use G-Theory to compute reliability for a complex performance-based assessment in pharmacy education, as well as showing the use of decision-studies to illustrate changes in reliability depending on number of stations over multiple weeks/occasions. This article is intended to demonstrate use of G-Theory with an example of performance-based assessment in pharmacy education.

DESCRIPTION OF INNOVATION

In this OSCE iteration, the 14 stations were: device counseling, over-the-counter counseling, knowledge of top 200 medications, compounding calculations, prescription checking, obtaining a medication history, medication reconciliation, drug information presenting, renal dosing, adverse drug events, adherence barriers, pharmacokinetic calculations, intravenous compatibility, and drug interactions. These stations were planned, created, and developed by a team of five practicing faculty pharmacists. The fourteen skills-based OSCE stations were divided into 4-5 stations per week. Thus, one OSCE spanned three weeks to include all of the different stations.

The OSCE circuit format was roughly 9-minutes per station with a 1-minute break in-between (though some stations took twice as long and so were scheduled as "double-length" stations). One or two stations each week used faculty raters, while the other stations (three per week) were written and scored afterwards. Every station was scored independently using a holistic 4-point scale (whether 4-point rating-score of a rater-based station or from a 4-point grading rubric of a written station). All stations were equally-weighted in students overall score. The addition of written stations has previously been shown to improve the reliability of an entire OSCE, with suitable validity, if the written stations can adequately address the skills being assessed.[2]

Assessment

Pharmacy Practice Faculty volunteered to be raters and did not receive training beyond an email description of that week's stations, and some instruction on using the associated scoring rubric for those stations. Because participating faculty differed each week (and were not the same for all weeks), raters in the G-Theory assessment designs were nested in and not crossed with stations.[9] Measurement error from raters and stations could not be completely parsed from each other.

Students' success on each station was assessed for that station (pass/fail). The requirements to satisfactorily pass an OSCE station were integrated into the criteria of rubrics for each station and developed by a team of four pharmacy practice-based faculty. Students could fail one station in all three weeks of the initial OSCE and still pass the entire OSCE. However, if students failed two or more stations, they needed to successfully remediate and repeat those specific failed stations in a following week, in order to pass the entire OSCE. This study was IRB-approved as exempt at the University of Toledo.

Statistical Analysis

Generalizability Theory (G-Theory) was used in analyzing this performance-based assessment. G-Theory is a statistical modeling technique that estimates reliability when multiple factors are identified as contributing to observed score variance.[8] G-Theory is especially well-suited for evaluating performance-based assessments.[3,8,9] Two G-Theory designs were analyzed. First, a univariate design of students were crossed with stations that were nested in weeks (p x s:w). For this univariate G-Theory design, G_String software was used (McMaster University, Hamilton ON). Second, and to better understand which station(s) may be of more concern in revisions of the learning assessment after used, a multivariate G-Theory design of students (random facet) crossed with station (random facet) was used ($p^{\bullet} \times s^{\circ}$). (Note: The third facet of 'number of weeks' is not identified in the multivariate design equation; this facet was fixed as opposed to random in our attempt to constrain the error into smaller categories—instead of station variance spread over three weeks, it was forced to three one-week categories.) For this multivariate G-Theory design, mGENOVA software was used (University of Iowa, Iowa City IA). Furthermore, suggested reporting practices for G-Theory were used, including description of facets, reliability, variance components, and decision-studies.[9]

CRITICAL ANALYSIS

In 2017, 97 PharmD students took part in this OSCE. On average, these students were 23-years-old (standard deviation of ±1.8-years) and 65% (63/97) were female. Ninety-two percent of stations (1259 of 1358 attempts) were satisfactorily passed. Our g-coefficient (reliability) for the total-score, based on 14 stations attempted over three weeks, was 0.72. A threshold of 0.80 is often considered acceptable for high-stakes testing.[5] G-Theory estimates of the variance in observed scores attributable to each of the modelled components of the OSCE total-score are shown in Table 1. As noted, variance in the measured ability of students accounted for almost two-thirds of overall variance in the total-score. Meanwhile, context specificity (a common limitation in learning assessments) accounted for close to one-quarter of score variance.

DOI: https://doi.org/10.24926/iip.v12i1.2110

Table 1. Variance Component Estimates from a Practice OSCE of 3rd-year PharmD Students using a p x (s: w) G-Theory design

	Variance Component Description	OSCE Score Variance
student *(p)*	Variance from difference in ability of students	0.13 (64%)
station *(s)*	Variance from difficulty of stations for all students	0.015 (7%)
station nested in week *(s : w)*	Variance from difficulty in stations from each week	0.009 (4%)
student x week *(p x w)*	Variance from some students (but not all) finding a week more difficult than other weeks	0.001 (1%)
student x station nested in week *(p x (s : w))**	Variance from some students finding some stations more difficult than others (context specificity)	0.049 (24%)
Total Variance		0.204

*This also includes residual error

Based on the multivariate G-Theory analysis, Table 2 shows the expected g-coefficient (reliability) estimates from decision-studies for varying numbers of stations in each of the three weeks. Scenarios with an estimated g-coefficient below the accepted 0.80 are in gray. From this Table, it can be noted that stations in week 2 (obtaining a medication history, analyzing for medication reconciliation, presenting a drug information case, renal dosing, identifying and resolving an adverse drug reaction) were far less reliable than stations in other weeks.

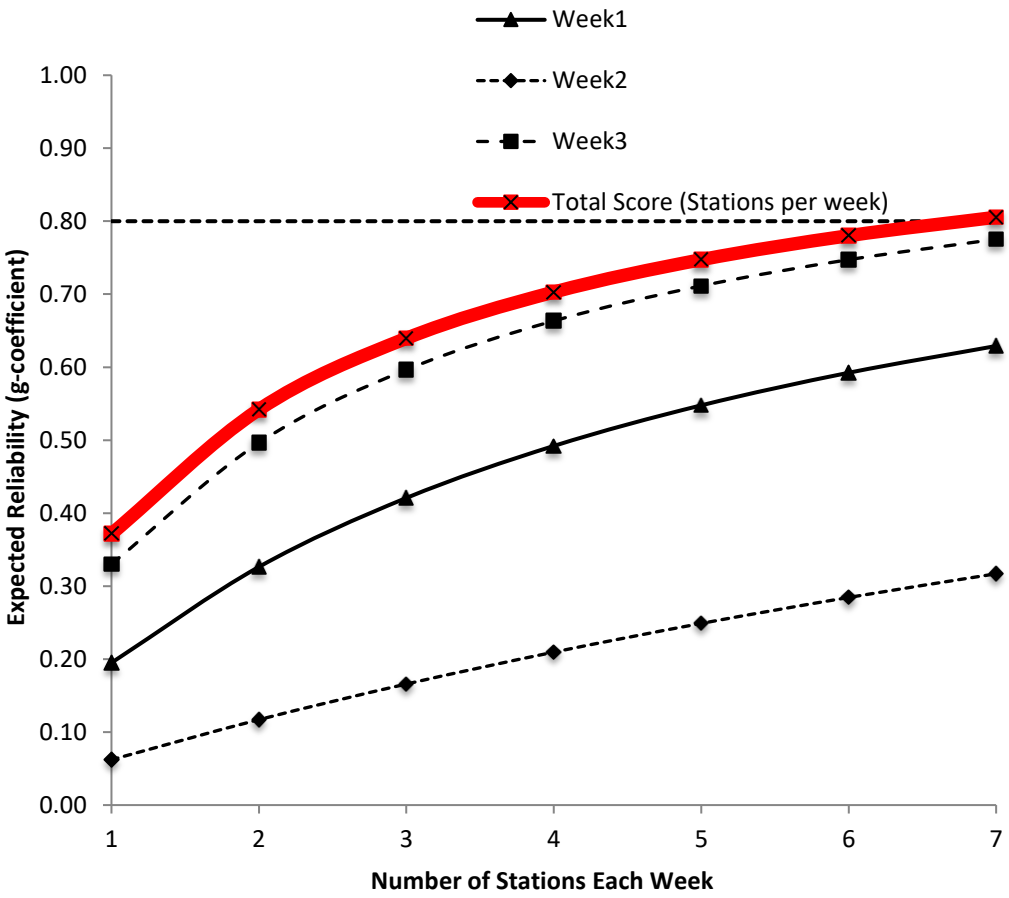
Table 2. Estimated G-Coefficients for Stations in 3rd-year PharmD OSCE

	Stations/week						
Week #	1	2	3	4	5	6	7
Week 1	*0.19*	*0.33*	*0.42*	*0.49*	*0.55*	*0.59*	*0.63*
Week 2	*0.06*	*0.12*	*0.17*	*0.21*	*0.25*	*0.28*	*0.32*
Week 3	*0.33*	*0.50*	*0.60*	*0.66*	*0.71*	*0.75*	*0.78*
Total-Score	***0.37***	***0.54***	***0.64***	***0.70***	***0.75***	***0.78***	**0.81**

Also recommended for reporting findings from a generalizability-study,[6] Figure 1 illustrates estimated g-coefficients from a set of decision-studies that varied the number of stations per week. As seen in Figure 1, the total-score was greater than any of the weeks alone. In addition, the results showed that seven stations per week were estimated in order to achieve a total-score reliability greater than .80.

DOI: https://doi.org/10.24926/iip.v12i1.2110

Figure 1. Graph of Estimated G-Coefficients with Increased Stations in Each of Three Weeks



Note: A threshold of 0.80 is often considered acceptable for high-stakes testing[5]

KEY ISSUES

The aim of this report was to demonstrate use of G-Theory in pharmacy education to provide reliability evidence in the validation process for performances of third-year PharmD students in an OSCE (i.e., performance-based assessment). The g-coefficient for the composite reliability over 3 weeks (14 attempted stations) was mediocre; it should be improved to a commonly accepted 0.80 with high-stakes testing.[5] Students' ability accounted for a moderate percentage of score variation (64%), though not all variation. Notably, other variation accounted for a substantial portion of the variation in scores (36%). This is consistent with many other studies of an OSCE.[1-3]

Decision-studies, estimating the g-coefficients with increasing stations, found that increasing the number of stations each week improved the composite reliability. This is also consistent with other studies of an OSCE.[1-3] Based on estimates from decision-studies within the current investigation, use of seven stations per week should provide a g-coefficient greater than 0.8 (g-coefficient=0.81). Additionally (though not to the exclusion of trying to add stations), stations in week 2 (obtaining a medication history, analyzing for medication reconciliation, presenting a drug information case, renal dosing, identifying and resolving an adverse drug reaction) should be closely reviewed (see Table 2), as these were far less reliable than stations in other weeks. (Note: this uses results from the multivariate G-Theory as a diagnostic strategy for identifying poor stations.)

While OSCEs are one method to assess PharmD students' practice-readiness for APPE experiences, these assessments should each be individually validated.[7] That is, important decisions such as to promote a PharmD student to their advanced pharmacy practice experiences, is high-stakes; sound validation evidence should support it. However, developing suitable, rigorous performance-based assessments can be challenging.[1,2] Furthermore, it is not enough to assume that a learning assessment is suitable, rigorous, and fair to students (i.e., no validation evidence), or to only use validation evidence from another institution. The current report demonstrates the necessary validation evidence could look like, when using G-Theory in evaluating an OSCE.

This research is not without limitations. The specific number of stations is sample-dependent; it is based on the specific data from administering this OSCE. Analyses of performance-based

DOI: https://doi.org/10.24926/iip.v12i1.2110

assessments (e.g. OSCEs) at other institutions may differ. That said, the general finding will not change—a larger number of stations will improve reliability of OSCEs everywhere.[1-3] Additionally, a single faculty-rater was used in some stations of our OSCE. This assessment design did not explore if a second rater would add a practical advantage. However, using more raters in any station has most often been less-advantageous than increasing the number of stations.[1]

NEXT STEPS

Within the framework for validation of an OSCE, this report provides evidence towards the generalization inference using reliability. For the foreseeable future, our curriculum will continue to offer a high-stakes OSCE during the P3 year. Our results indicate that our reliability of test-scores was mediocre (though decent when compared to others reported elsewhere[4]); however, its reliability could be improved with addition more stations each week. Additionally, targeting revision efforts to poorly performing week 2 stations may also improve reliability (aside for only increasing the number of stations).

It is inevitable that other small changes in test-content, as well as differences with administering our OSCE, will occur each year. These small changes may also alter the reliability of scores for our performance-based assessment. Thus, we will need to continue to monitor the reliability of test-scores from our OSCE.

This OSCE report also builds on prior findings from a non-OSCE performance-based assessment using a multiple-item rubric.[10] While illustration of G-Theory was helpful to provide scoring evidence there, it provided generalization evidence here. With soundness of reliability, using G-Theory within learning assessment validation can increase confidence in subsequent decision-making based on a learning assessment's scores.

Within Kane's Framework for Validation, three validation studies could be sequential future steps. First, while G-Theory appears to be most appropriate for generalization evidence for complex learning assessments such as performance-based assessments,[3] it may prove helpful to explore scoring for some more problematic stations (especially from our Week 2). Rasch Measurement analysis may also be helpful with this scoring evidence.[7] Second, a further study could examine extrapolation evidence from associations with non-OSCE PharmD student outcomes, like scores on the pharmacist licensing exam. Third, it seems prudent to evaluate the decision-rules that were used—namely, the cut-score for passing versus remediating stations, as well as our decision-rule for determining the performance required to satisfactorily complete the entire OSCE.

Regardless, it should remain clear that Cronbach's alpha and other coefficients of internal consistency should only be used for a simple written assessment on a single occasion (with only students and items as sources of variance). Introducing more occasions, such as more weeks of OSCE stations, introduces another test parameter. Using internal consistency alone would be insufficient in evaluating reliability for test-scores from this more complex learning assessment. Similarly, using inter-rater reliability alone would also be insufficient for characterizing reliability of test-scores from an OSCE (or other performance-based assessment). That said, both internal consistency and inter-rater reliability might be helpful with scoring evidence of single stations—but not for the composite reliability from multiple stations. However, G-Theory can accurately analyze this three-parameter composite reliability for *students*, *stations*, and *weeks*.

CONCLUSION

In test validation, evidence of reliability is vital for the generalization inference; G-Theory provided this for our OSCE. Our results indicated that the reliability of our scores was mediocre and could be improved with more stations. A practicality of this need for more stations is that multiple occasions (e.g., weeks) appeared appropriate for scheduling this learning assessment.

Funding/Support: None
Conflicts of Interest: None

REFERENCES

1. van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996; 1(1):41-67. doi: 10.1007/BF00596229.
2. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ*. 2004 Feb;38(2):199-203. doi: 10.1111/j.1365-2923.2004.01755.x.
3. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales*. 5th ed. New York, NY: Oxford University Press; 2015.
4. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ*. 2011; 45(12):1181-1189. doi: 10.1111/j.1365-2923.2011.04075.x.
5. Peeters MJ, Cor MK. Guidance for high-stakes testing within pharmacy education. *Curr Pharm Teach Learn*. 2020; 12(1):1-4. doi: 10.1016/j.cptl.2019.10.001
6. Hendrickson A, Yin P. Generalizability Theory. In Hancock GR, Stapleton LM, ed. *The Reviewer's Guide to Qualitative Methods in the Social Sciences.* 2nd ed. New York, NY: Routledge, 2019:123-131.
7. Peeters MJ, Martin BA. Validation of learning assessments: a primer. *Curr Pharm Teach Learn*. 2017; 9(5):925-933. doi: 10.1016/j.cptl.2017.06.001
8. Brennan RL, Johnson EG. Generalizability of performance assessments. *Educ Meas*. 1995; 14(4):9-12. doi: 10.1111/j.1745-3992.1995.tb00882.x.
9. Peeters MJ. Moving beyond Cronbach's alpha: Improving decision-making in pharmacy education. *Innov Pharm*. 2021; 12(1):Article 14.
10. Cor MK, Peeters MJ. Using generalizability theory for reliable learning assessments in pharmacy education. *Curr Pharm Teach Learn*. 2015; 7(3):332-341. doi: 10.1016/j.cptl.2014.12.003.