Validation Evidence from using Generalizability Theory in a Basic-Science Course: Reliability of Course-Grades from Multiple Examinations

Michael J. Peeters, PharmD, PhD, FCCP, BCPS[1]; M. Kenneth Cor, PhD[2]; Sai HS Boddu, PhD[1,3]; Jerry Nesamony, PhD[1]
[1]University of Toledo College of Pharmacy & Pharmaceutical Sciences, Toledo, OH
[2]University of Alberta Faculty of Pharmacy & Pharmaceutical Sciences, Edmonton, AB
[3]Department of Pharmaceutical Sciences, College of Pharmacy and Health Sciences, Ajman University, Ajman, United Arab Emirates

ABSTRACT

Description of the Problem: *Reliability is critical validation evidence on which to base high-stakes decision-making. Many times, one exam in a didactic course may not be acceptably reliable on its own. But how much might multiple exams add when combined together?*
The Innovation: *To improve validation evidence towards high-stakes decision-making, Generalizability Theory (G-Theory) can combine reliabilities from multiple exams into one composite-reliability (G_String IV software). Further, G-Theory decision-studies can illustrate changes in course-grade reliability, depending on the number of exams and exam-items.*
Critical Analysis: *101 first-year PharmD students took two midterm-exams and one final-exam in a pharmaceutics course. Individually, Exam1 had 50MCQ (KR-20=0.69), Exam2 had 43MCQ (KR-20=0.65), and Exam3 had 67MCQ (KR-20=0.67). After combining exam occasions using G-Theory, the composite-reliability was 0.71 for overall course-grades—better than any exam alone. Remarkably, increased numbers of exam occasions showed fewer items per exam were needed, and fewer items over all exams, to obtain an acceptable composite-reliability. Acceptable reliability could be achieved with different combinations of number of MCQs on each exam and number of exam occasions.*
Implications: *G-Theory provided reliability critical validation evidence towards high-stakes decision-making. Final course-grades appeared quite reliable after combining multiple course exams—though this reliability could and should be improved. Notably, more exam occasions allowed fewer items per exam and fewer items over all the exams. Thus, one added benefit of more exam occasions for educators is developing fewer items per exam and fewer items over all exams.*

Keywords: occasion, course, reliability, validation, generalizability theory

DESCRIPTION OF PROBLEM

Educators want fair and rigorous assessments of students' learning in their course but may also want to make high-stakes decisions based on students' in-course performances. When a student fails a course during their PharmD coursework, increased stakes can result, if there becomes a delay in that student's progression through their PharmD program.[1] Increased stakes (including high-stakes) decisions need ample validation evidence to support educators' and administrators' inferences; with the higher the stakes, the more/stronger validation evidence needed.[1,2] Thus, examinations and other assessments of students' learning, should be sufficiently valid (providing accurate measures of content being assessed), including suitable reliability (consistently statistically-discriminating among students).[3,4] Put simply, evidence of reliability is needed for courses—for validation evidence in cases were high-stakes decisions may occur.

Reliability can be deceptively complex; there are multiple types and approaches to it.[3,4] Readers will be familiar with Classical Test Theory (CTT) from their experiences as educators and as students. In CTT, scores for correct items on an examination are summed into a total score for the examination. Typically, the coefficient of reliability (KR-20 or Cronbach's alpha for CTT's internal consistency) is estimated individually for each exam in a course. This coefficient assumes only one source of measurement error—an inter-item sampling error from the specific set of items sampled from a universe of possible items to be included on that examination.[4,5] Further, CTT only analyzes one source of error at a time.[4-6] Alternatively, Generalizability Theory (G-Theory), extends CTT to model *multiple* sources of measurement error.[4,5] Being able to model the contribution of different characteristics of a measurement process to the observed error variance affords educators the ability to make decisions to optimize measurement towards better reliability. G-Theory can also examine the trade-off for estimation of reliability of scores derived from combinations of number of exam items and number of exam occasions. (For a more detailed primer on G-Theory, see the companion to this article.[5])

Of note, we assumed Latent Trait Theory for this investigation. Much like quantifying temperature, weight, and serum sodium in the clinical sciences, Latent Trait Theory is extremely common in the social sciences, when and where researchers are trying to investigate and quantify entities that are non-physical such as empathy (psychology) or knowledge

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

DOI: https://doi.org/10.24926/iip.v12i1.2925

(education). Within this study, learners were assumed to have a latent trait of 'general pharmaceutics knowledge' that educators were trying to measure with the various exam items. With this assumed, all multiple items on all exams should align and "tap" into that latent trait of 'general pharmaceutics knowledge'; different exams should simply be different (repeated) measures.

DESCRIPTION OF INNOVATION

This study was IRB-approved as exempt by the University of Toledo; all analyses were retrospectively conducted.

From this report, three innovations are notable. First, we illustrate combining multiple assessments of learning into a composite-reliability for an overall course-grade. Second, we demonstrate use of G-Theory to do this combining and compare with traditional CTT indices. Third, we more specifically examine the interaction of exam items with multiple exam occasions.

Approaches to Reliability

In this investigation, reliability was described using both CTT and G-Theory approaches, so that these could be compared. Moreover, we used a conventional high-stakes reliability cutoff of 0.8.[1,3]

Using Classical Test Theory for Individual Exams. Using the KR-20 for internal consistency, a reliability coefficient for students' scores was reported for each individual exam by ExamSoft (ExamSoft Worldwide, Dallas TX). Of note, using this CTT approach, with its multiple KR-20 reliability coefficients for each of the multiple individual exams, all items could be calculated into one combined KR-20; however, that would ignore the multiple separate exam occasions (e.g., midterm exam, final exam on different days). That is, a single combined KR-20 would simply be in error and misleading.

Course-Grades via Generalizability Theory. With addition of an *occasion* test parameter (*occasion* facet in G-Theory terminology[4,5]), G-Theory could analyze student performances over multiple exams to construct the composite-reliability of course-grades. Our G-Theory assessment design was *students crossed with items nested in occasions* (*p x i : o*). In this G-Theory design, all facets were random and variation in observed course-grades was explained by potential differences from: isolated students' pharmaceutics ability (student variance; *p*), difficulty of different exams (occasion variance; *o*), difficulty of items on the different exams (items nested in occasions variance; *i : o*), the interaction between students and the different occasions (variation in how dissimilar students performed from one test to the next, such as changes in test-taking circumstances; *p x o*), as well as the interaction between students and items nested in occasions (variation in how some students performed on different sets of items from one exam occasion to the next; *p x i : o*).

Participants & Course Design

One-hundred and one 1st-year PharmD students took this basic-science (pharmaceutics) course at the University of Toledo College of Pharmacy & Pharmaceutical Sciences. Of the 101 students, 36 were males and 65 were females, with an average age of 21 years (standard deviation of 1.6 years). In 2017, this was a 15-week course with 12 weeks of instruction and three weeks set aside for examinations (i.e., two midterms and one final-exam). The pharmaceutics course was designed to introduce students to basic concepts of dosage forms and the materials, methods, and technology used in the preparation of manufactured/compounded pharmaceutical products. Students took all examinations in this course using ExamSoft.

Reliability Analyses

As commonly accepted practice, the course instructor adjusted PharmD students' performance scores using data from item analysis (e.g., percent correct, point biserial) both before and double-checking with student appeals. Any adjustments were prior to this investigation's analyses.

Internal consistency reliability (by KR-20) was computed for each examination separately in ExamSoft and confirmed using SPSS version 25 for Mac (Armonk, NY). G-Theory was used to estimate the reliability of the course-grades derived from these examination occasions; we used G-String IV (Hamilton, ON, Canada). In line with best practice reporting guidelines for G-Theory, description of the measurement facets, reliability, variance components, and decision-studies (Table 1 and Figure 1) have been provided.[5]

CRITICAL ANALYSIS

Individual Examinations via Classical Test Theory

Used to calculate individual exam reliability with CTT, the estimated KR-20 reliability for each of the three exams separately were: Exam 1 (50 questions) KR-20=0.69, Exam 2 (43 questions) KR-20=0.65, Exam 3 (67 questions) KR-20=0.65.

Composite-Reliability via Generalizability Theory

To estimate the composite-reliability of course-grades, this G-Theory model analyzed the change in reliability as a function of number of occasions and items within occasions. Based on the three exams, this composite-reliability was 0.71. This was higher than any KR-20 from an exam alone.

Importantly, the amount of variation from multiple sources of variation that impacted reliability was estimated for students in this course. From the G-Study analysis, we found that only 2% of variance in course-grades was attributable purely to student ability differences alone, suggesting more difficult items appear needed on all exams. Meanwhile, as evidence of good assessment design, close to zero percent was from differences attributable to exam occasion alone, suggesting that exams were of similar difficulty. Moreover, 1% was explained by differences in how *some* students performed from occasion to occasion, suggesting that most students had similar

DOI: https://doi.org/10.24926/iip.v12i1.2925

performances on all exams and were not making up for a poor performance on one. However, approximately 21% was attributable to differences in item difficulty within occasions. The majority (76%) was attributable to differences in student performance on the different items that were nested in each occasion.

Building upon and using these variance contributions from the various sources, Table 1 shows specific estimates about how reliability could change as a function of the number of occasions and the number of items nested in occasion. With a high-stakes reliability cutoff of 0.8, using multiple exam occasions enabled acceptable reliability. Although, this differed with the number of questions and exam occasions. It was notable that a larger number of exam occasions could mean fewer items were needed on each exam. Interestingly, use of more exam occasions appeared to save on the overall number of items needed over all exams.

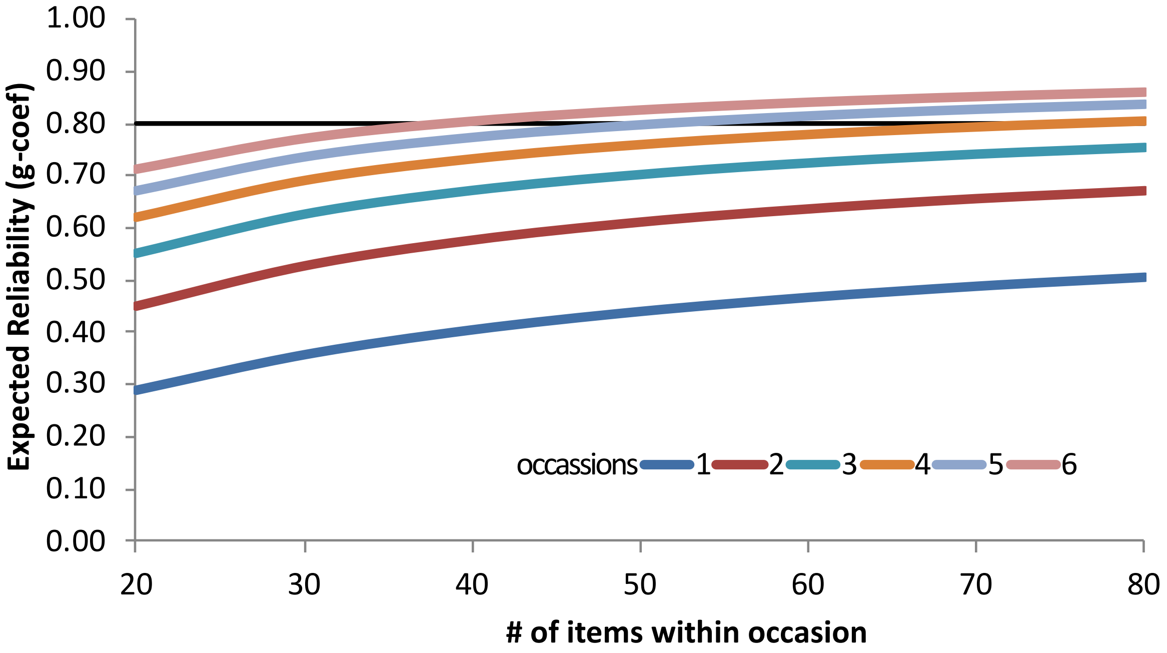
Table 1. Decision-studies of estimated reliability (via G-coefficients) for various numbers of items and various numbers of exam occasions for a first-year PharmD basic-science course

		Number of Items						
		20	30	40	50	60	70	80
Occasions	1	0.29	0.36	0.41	0.44	0.47	0.49	0.51
	2	0.45	0.53	0.58	0.61	0.64	0.66	0.67
	3	0.55	0.63	0.67	0.70	0.73	0.74	0.76
	4	0.62	0.69	0.73	0.76	0.78	0.79	***0.81***
	5	0.67	0.74	0.77	0.79	***0.82***	***0.83***	***0.84***
	6	0.71	0.77	***0.80***	***0.83***	***0.84***	***0.85***	***0.86***

Note: Bold meets an acceptable threshold of 0.80[1,3]

Figure 1 illustrates Table 1 in a graphical format. As seen, one to three exam occasions did not approach the 0.8 threshold for high-stakes testing. Although, four exams could—with many exam items.

Figure 1. Course-grade reliability as a function of number of testing occasions and items nested in occasion



Note: Line at 0.8 as threshold for acceptable reliability (for high-stakes decision-making) [1,3]

DOI: https://doi.org/10.24926/iip.v12i1.2925

This investigation has limitations. It was context specific. This analysis was from a single cohort from a single year. More specifically, it was from one PharmD course in one PharmD curriculum at one institution. The precise reliability numbers from this analysis are sample-dependent. Further, while many items had been used previously (and had acceptable item analysis then), other items were new. Lastly, we took a high-stakes decision-making approach to this investigation. While an exam's summative role is clearly a concern, learning assessment can have formative roles as well (and are beyond this investigation).

KEY ISSUES

Herein, we demonstrated G-Theory's integrated reliability coefficients from multiple examination occasions (e.g., midterm exams, final exam) into a composite-reliability. The improvement was small (by .02-.06), but not inconsequential. Reliability of a course-level grade should better reflect the rigor of an entire course, as opposed to just looking at a reliability index for one exam occasion.

Not surprisingly, more items lent to higher reliability. But it was not linear nor easy to add KR-20s together. Instead, G-theory analyzed how the multiple exams measured an underlying pharmaceutics ability of students. The composite-reliability was an improvement over any individual exam. One notable insight was that there was a balance (or trade-off) of exam length (number of items) and number of exam occasions, with the shorter each exam, the more occasions that are needed. Thus, acceptable composite-reliability could be accomplished with different numbers of items over the different numbers of exam occasions.

Table 1 showed various combinations for number of exam occasions and number of exam item within each exam occasion that can be used to achieve an acceptable reliability. For example, four 80-item exams (320 items total) could be used to achieve approximately the same level of acceptable reliability (of 0.8) as with six 40-item exams (240 items total). Considering the amount of effort that it takes to create a single high-quality MCQ item, a difference of 80 total questions can represent a significant time and effort savings for exam developers. Thus, at least one more exam occasion appears needed for the context of this current pharmaceutics course being studied. Of note, these item-saving findings are similar to authors' unpublished experiences observing similar over many years in multiple settings.

Our study appears innovative in examining a composite-reliability at a course-level. Elsewhere in health-professions education, multiple components (e.g., MCQ, extended matching items, short-answer items, essay responses, OSCE history taking cases) of one assessment have been combined but each component only administered on one occasion.[7] In addition, this idea was expanded with a theoretical basis for evolving from a single learning assessment to a program (involving numerous learning assessments).[8] It is at this higher program-level that rigorous (reliable), meaningful interpretations can better be made.

Furthermore, studies outside health-professions education have demonstrated that including an *occasion* facet in analyzing reliability gave an improved estimation of reliability for the entirety of the multiple occasion learning assessment.[9] Therein, using only a single test occurrence was insufficient. Thus, in instances where testing is over multiple occasions, such as for an entire course as opposed to one high-stakes exam occasion like a licensing exam, it would seem that addition and consideration of an occasion facet should be recognized in analysis.

NEXT STEPS

A prior review of the pharmacy education literature by Hoover and colleagues documented that reliability was reported only sometimes (<20%).[10] Notably, all of those reliability coefficients were from a single learning assessment used on a single occasion (personal communication, 2019). None describe reliability of a course-grade. If course advancement is seen as a high-stakes situation,[1] the reliability of course-grades should matter most. Its reliability will come from the entire *collection* of learning assessments in that course and not from reliability of scores for any single exam occasion. That is, an *occasion* facet, as used in Generalizability Theory, can better estimate reliability of a course-grade, as opposed to simply a single exam.

Conflicts of Interest: None
Funding/support: None

REFERENCES

1. Peeters MJ, Cor MK. Guidance for high-stakes testing within pharmacy educational assessment. *Curr Pharm Teach Learn*. 2020; 12(1):1-4. doi: 10.1016/j.cptl.2019.10.001
2. 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
3. Peeters MJ, Beltyukova SA, Martin BA. Educational testing and validity of conclusions in the scholarship of teaching and learning. *Am J Pharm Educ*. 2013;77(9):article 186. doi: 10.5688/ajpe779186
4. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales*, 5th ed. New York, NY: Oxford University Press, 2015.
5. Peeters MJ. Moving beyond Cronbach's alpha and inter-rater reliability: A primer on Generalizability Theory for pharmacy education. *Innov Pharm. 2021; 12(1):Article 14.*
6. Brennan RL. Generalizability Theory and Classical Test Theory. *Appl Meas Educ*. 2011;24(1):1-21. doi: 10.1080/08957347.2011.532417

DOI: https://doi.org/10.24926/iip.v12i1.2925

7. Webb NM, Schlackman J, Sugrue B. The dependability and interchangeability of assessment methods in science. *Appl Meas Educ*. 2000; 13(3):277-301. doi:10.1207/S15324818AME1303_4
8. Van Der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Ed*. 2005; 39(3):309-317. doi: 10.1111/j.1365-2929.2005.02094.x
9. Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Med Educ*. 2001; 35(4):326-30. doi: 10.1046/j.1365-2923.2001.00929.x
10. 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. doi: 10.5688/ajpe7710213

DOI: https://doi.org/10.24926/iip.v12i1.2925