Providing Validation Evidence for a Clinical-Science Module: Improving Testing Reliability with Quizzes

Michael J. Peeters, PharmD, PhD[1]; M. Kenneth Cor, PhD[2]; Erik D. Maki, PharmD[3]
[1]University of Toledo College of Pharmacy & Pharmaceutical Sciences
[2]University of Alberta Faculty of Pharmacy & Pharmaceutical Sciences
[3]Drake University College of Pharmacy

ABSTRACT

Description of the Problem: *High-stakes decision-making should have sound validation evidence; reliability is vital towards this. A short exam may not be very reliable on its own within didactic courses, and so supplementing it with quizzes might help. But how much? This study's objective was to understand how much reliability (for the overall module-grades) could be gained by adding quiz data to traditional exam data in a clinical-science module.*

The Innovation: *In didactic coursework, quizzes are a common instructional strategy. However, individual contexts/instructors can vary quiz use formatively and/or summatively. Second-year PharmD students took a clinical-science course, wherein a 5-week module focused on cardiovascular therapeutics. Generalizability Theory (G-Theory) combined seven quizzes leading to an exam into one module-level reliability, based on a model where students were crossed with items nested in eight fixed testing occasions (mGENOVA used). Furthermore, G-Theory decision-studies were planned to illustrate changes in module-grade reliability, where the number of quiz-items and relative-weighting of quizzes were altered.*

Critical Analysis: *One-hundred students took seven quizzes and one exam. Individually, the exam had 32 multiple-choice questions (MCQ) (KR-20 reliability=0.67), while quizzes had a total of 50MCQ (5-9MCQ each) with most individual quiz KR-20s less than or equal to 0.54. After combining the quizzes and exam using G-Theory, estimated reliability of module-grades was 0.73; improved from the exam alone. Doubling the quiz-weight, from the syllabus' 18% quizzes and 82% exam, increased the composite-reliability of module-grades to 0.77. Reliability of 0.80 was achieved with equal-weight for quizzes and exam.*

Next Steps: *Expectedly, more items lent to higher reliability. However, using quizzes predominantly formatively had little impact on reliability, while using quizzes more summatively (i.e., increasing their relative-weight in module-grade) improved reliability further. Thus, depending on use, quizzes can add to a course's rigor.*

Keywords: validation, reliability, generalizability theory, quizzes

DESCRIPTION OF PROBLEM

High-stakes decision-making is common in pharmacy education. For example, failure of a specific course may result in delays in students' progression through a PharmD program.[1] This is understandable given the importance of pharmacist competence to rationale medication use in the healthcare system. However, high-stakes decisions should be undergirded by sound validation evidence, and reliability is a vitally important component of this evidence.[1] Thus, evidence for elements that may result in course failure (and delays students' progression through a PharmD program) should have sound rigor. That is, the composite-reliability of course-grades, that includes all learning assessments within that course, should be scrutinized. In our litigious society, reliability can be a key vulnerability for legal challenges.[2] Another related and important reason for sound reliability is fundamental fairness to students; test-scores and decision based on test-scores should be fair for them.

During student pharmacists' education, quizzes are an often-used pedagogical technique within a diversity of classroom instructional approaches (e.g., traditional lecture-based, case-based, team-based learning or flipped classrooms).[3] Formatively, quizzes can promote test-enhanced learning.[4] Although summatively, test-scores from any single quiz have a notoriously poor reliability, and so are not often a summative focus for assessment of students' learning. In fact, we did not find any literature that describes a summative assessment role for quizzes (and it has not been discussed in our academic experiences either). Demonstrating a summative assessment role for quizzes through their ability to enhance rigor (reliability) could add another dimension to their use in many courses.

DESCRIPTION OF INNOVATION

This investigation was IRB-approved as exempt by Drake University, as analyses were all conducted retrospectively.

Computing reliability for assessment of learning (e.g., exams) overwhelmingly uses Classical Test Theory's internal consistency, with the Kuder-Richardson Formula 20 (KR-20) most frequent. The KR-20 is a special case of Cronbach's alpha for dichotomous (right/wrong) data and is commonly used/reported by testing software such as ExamSoft™.[5] A KR-20 is limited to calculation for only one testing episode at a time; KR-20s cannot be combined. That said, it is well-known that test reliability will likely improve with a greater number of items on a single exam (i.e., scores from a longer exam with more questions should be more reliable than scores from a shorter exam).[6] It is not surprising that test-scores from one

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.v12i1.2235

quiz have a notoriously poor reliability, when conventional reliability analyses such as KR-20s are performed, because that single quiz by definition has few items (and so inadequate sampling of items).

As a result, a Classical Test Theory perspective seems too simplistic and inadequate for understanding course grade reliability. Quizzes are not isolated independent events but are dependent on a course exam, as they build towards it. As an alternative to Classical Test Theory, G-Theory combines the multiple reliabilities from the multiple quizzes and exam into one composite-reliability for a course-grade. If needed for high-stakes decision-making, this course-grade reliability could be scrutinized instead of the exam reliability alone.

Given the need for validation evidence in high-stakes decision-making, Kane's Framework for Validation provided the theoretical framework used in this investigation.[8] Within Kane's Framework for Validation, reliability evidence is integral for its generalization inference.[8] Additionally, this report follows with other examples in a series of articles demonstrating uses for G-Theory within pharmacy education, and this series begins with a primer on G-Theory.[7]

G-Theory Assessment Design

To calculate the composite-reliability of grades from this module (from combining quiz and exam scores), G-Theory was used (mGENOVA; University of Iowa, Iowa City, IA). While G-Theory's computation is complicated and greatly assisted by computer software, G-Theory provides a means to combine multiple assessments into one overall reliability.

While the numerous G-Theory assessment designs have mathematical foundations, no mathematical formulae are included here; for brevity, an interested reader could review these elsewhere.[6] The associated primer with this article can also provide more G-Theory background, prominent resources, general methodology, and nomenclature.[7]

In this analysis, *students* were *crossed* with *items* that were *nested* in the different *testing occasions*. In G-Theory, nomenclature this is a person x item : occasion ($p^{\bullet} \times i^{\circ}$). This design was multivariate for *testing occasion* because each occasion was seen as rating a distinct yet related aspect of cardiovascular therapeutics such that each was considered a separate (though related) variable of student performance. The design was also unbalanced, as the quizzes and exam had different numbers of items. Beyond composite reliability, the G-Study would estimate the amount of variance in course-grade that is attributable to the person and item facets (i.e., variance components). Further, percent-weights were used for the quizzes and exam according to the syllabus (18% quizzes, 82% exam). In a subsequent decision-study, changes in percent-weight were compared, with the impact of each on reliability examined. Following G-Theory reporting guidance, the facets, design, variance components, reliability, and decision-studies have been described.[7]

Innovation

As with other learning assessments, quizzes can have both formative and summative assessment roles. With potential for high-stake decision-making in various PharmD coursework, the summative assessment role is important. The innovation of this investigation was to describe, summatively, how integration of multiple quizzes with the reliability for a course exam can enhance a composite-reliability of course-grades. Of note, enhanced reliability is important validation evidence that can greatly bolster instances of high-stakes decision-making, including a decision to delay a student's progression through a PharmD program.

CRITICAL ANALYSIS

Participants & Course Design

One-hundred and one 2nd-year PharmD students took a clinical-science (cardiovascular therapeutics) module. Students averaged 23.6 years-old (with standard deviation of ± 1.7 years-old) and 65.3% were female.

At Drake University, PharmD students completed three required, clinical-sciences (pharmacotherapy) courses during the 2nd and 3rd years of their curriculum. Within one of these courses during students' 2nd-year of their 4-year PharmD curriculum, we investigated a cardiovascular therapeutics module that spanned five weeks and used quizzes as part of its active-learning. Topics covered in this module included: Advanced Cardiovascular Life Support, Venous Thromboembolic Disease, Atrial Fibrillation, Acute Coronary Syndrome, Heart Failure and Chronic Stable Angina. There were eight learning assessments, which included seven quizzes and one modular exam. The course instructor has taught these cardiovascular topics for more than ten years; the items had been used previously, refined over time, and have been relatively stable. Students completed all assessments on paper Scantron™ forms. These were scanned using a Benchmark 3000 (Apperson Education Products, Cerritos, CA) and then graded electronically using DataLink Connect software (v4.4.02, Apperson Education Products, Cerritos, CA).

Reliability Analyses

As was typically done in the course, the instructor adjusted PharmD students' performance scores using data from item analysis (e.g., percent correct, point biserial) in the context of written student appeals. From G-Theory, the composite-reliability for grades from this module was 0.73. Table 1 reports the amount of overall test-score variance that can be attributed to sources of student, item, and the interaction of student with item (including residual error). Moreover, an internal consistency (by KR-20) is reported for each learning assessment for comparison with the G-Study's g-coefficient.

Table 1. G-Study variance components estimates by testing occasion, along with KR-20 coefficients for comparison

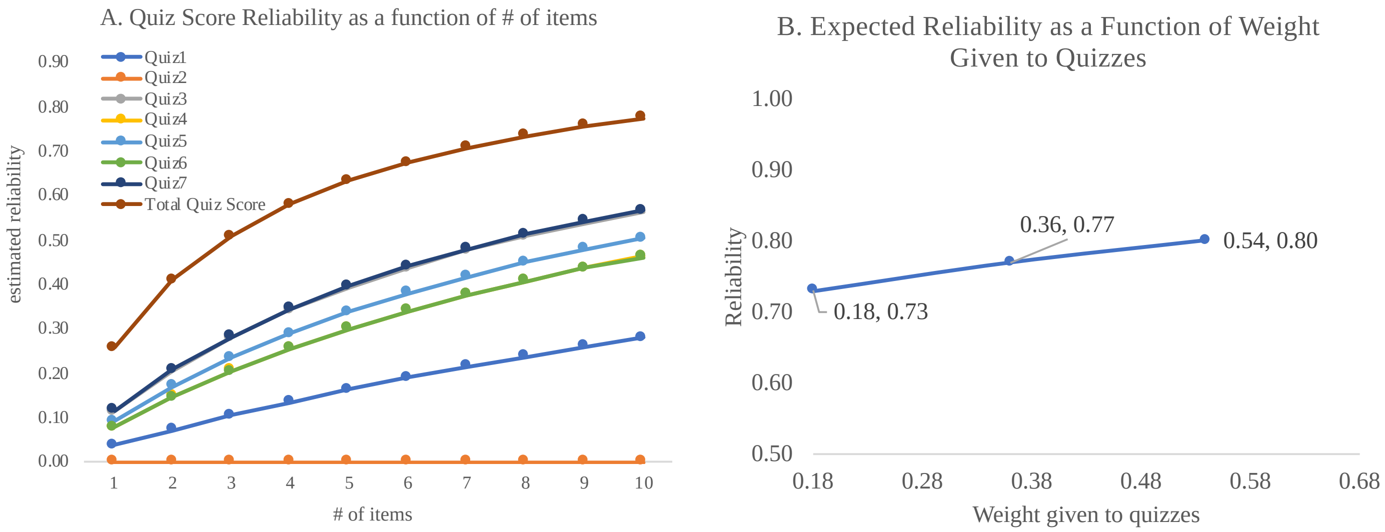
	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Quiz 6	Quiz 7	Exam	Overall
# of items	9	9	9	5	5	5	8	32	82
student	0.004 (3%)	0 (0%)	0.017 (9%)	0.011 (7%)	0.014 (9%)	0.011 (8%)	0.014 (9%)	0.007 (5%)	0.006 (6%)
item	0.038 (25%)	0.035 (21%)	0.032 (18%)	0.013 (9%)	0.013 (8%)	0.002 (2%)	0.029 (19%)	0.025 (18%)	0.017 (16%)
student x item (+error)	0.112 (73%)	0.128 (79%)	0.128 (73%)	0.129 (84%)	0.136 (84%)	0.124 (90%)	0.11 (72%)	0.111 (78%)	0.082 (78%)
KR-20 reliability*	0.26	0.00	0.54	0.30	0.34	0.30	0.51	0.67	

* These Classical Test Theory KR-20 coefficients are only to compare with the G-Study's g-coefficient of 0.73; they were not part of the G-Theory analysis.

The results showed the KR-20 reliability of the quiz-scores varied from 0.00 to 0.54, with the exam-score reliability highest at 0.67. Using the G-Study variance-estimates and the actual weights applied in the course (18% for quizzes and 82% for the exam), the overall-grade reliability was estimated to be 0.73.

Two sets of Decision-Studies were explored. First was the influence of increasing the number of items per quiz on test-score reliabilities for individual quizzes. Second was the influence of changing the weight of quizzes relative to the exam on overall-grade reliability. Both are shown in the two panels of Figure 1.

Figure 1. Decision-Studies for the number of items and weight of quizzes/exam during a second-year PharmD course



DOI: https://doi.org/10.24926/iip.v12i1.2235

In panel A of Figure 1, as the number of items increased, the reliability of test-scores from individual quizzes increased as well. Scores from Quiz7 were the most reliable, while scores from Quiz2 did not appear to measure reliably at all. (For Quiz2, this might have reflected students' focus, as it was right before students' Spring Break.) As in Table 1, a composite-reliability with the quizzes and exam aggregated was much more reliable than scores from any quiz or the exam. Finally, the curve in panel B of Figure 1 illustrates that increasing the weight associated with quizzes toward the total score, was estimated to improve the reliability of module-grades. Of note, reliability met the threshold of 0.8 for high-stakes testing when quiz-items had equal weight with exam-items.

Importantly, this investigation was limited by context. It was from one iteration of one class at one institution. Undoubtedly, quizzes have different content and are used differently in various courses at many institutions. Expectedly, reliabilities from those other courses at other institutions will differ in their specifics. However, a premise of potentially using quizzes to bolster reliability of course-grades appears logically and empirically sound; a summative impact from quizzes is generalizable. Examining reliability should be done at each institution; a precise reliability is context-dependent and it will be specific to each course at those institutions. Moreover, it bears mentioning that reliability, although of considerable importance, is secondary to and should not be allowed to drive test content or format. In addition, the intent of decisions from quiz-scores (e.g. high-stakes vs low-stakes) should also be considered with increasing weight of quizzes to improve reliability.

KEY ISSUES

Practically speaking, the findings from this investigation have two notable implications. The first is conceptual. It is well-known that reliability (via KR-20) will improve for tests if more test-items are used.[6] By definition, quizzes have few items and so each quiz will have poor reliability using traditional measures (e.g., KR-20). Although quizzes have historically been mis-analyzed as independent learning assessments on their own, they are dependent on and focused toward an exam (that is more reliable because it has many more exam-items). In other words, quality quiz-items aligned with the later exam-items can be seen as "exam-items" administered "earlier" than most of the exam (and thereby increasing the overall number of exam-items). Instead of analyzing each independently with KR-20, G-Theory allowed educators to combine different occasions of learning assessment (e.g., quizzes and an exam). In our investigation, increasing both the number of quiz-items (to 10 or 15), as well as increasing the percent-weight of quizzes could help bolster this composite-reliability for scores within this module. While it might seem counterintuitive that quizzes (each with a poor reliability by KR-20) can improve the stronger reliability of an exam (also by KR-20), quality quizzes can simply be seen as adding more test-items to an exam and thereby improving reliability of the entirety.

Second and as a result of this change in conception, quizzes can supplement a course-grade's reliability (if needed). Quizzes can have both formative and summative assessment roles; formative through focusing learners towards an upcoming exam and summative through the extent of weighting in course-grade calculations. With delaying a student's progression through their PharmD program seen as one potential high-stakes decision-making scenario, validation evidence by way of reliability can be foundational.[1,8]

Practically speaking, recall that the commonly-accepted threshold for high-stakes testing is 0.8;[1] a reliability below this should be supplemented with other learning assessment data. For instance, if a student failed a course (a course that delays PharmD program progression), it would only be fair to that student that they were adequately and fairly assessed in that failed course. Reliability is a key quality indicator for learning assessments (and is also important validation evidence). So if an educator, in their teaching, has a learning assessment with a reliability >0.8, they have achieved this evidence. However, if over multiple course iterations an exam reliability remains <0.8, than more assessment of a student's learning should be sought. Traditionally, more than one exam may be used in a course (e.g., one or two midterms and a final exam); this can help administer more related exam-items, if the multiple exams are aligned. Our study suggests another option—quizzes. If quizzes are aligned with an exam, these quiz-items can be used to supplement reliability of that exam. However, the weight of quizzes in calculating course-grades should be substantial compared to weight of exam (i.e., weight of quiz-items and exam-items should be similar) to improve reliability.

Notably, findings from this study in pharmacy education provide a similar though expanded picture to those from medical education. Wass, McGibbon, and van der Vleuten[9] showed that combining and altering the weighting of various learning assessments affected the composite-reliability of scores for a multipart exam. Although, Wass and colleagues' analysis did not involve quizzes and only involved different exam formats conducted during a single testing occasion.

NEXT STEPS

It appears that when testing is distributed over multiple testing occasions, *testing occasions* becomes another source of variance (and measurement error) that should be accounted for within the reproducibility of course grades.[10] In fact, including testing occasions as a source of variance, altered the contribution of other sources of variance—with a traditional internal consistency reliability being incorrect and overestimated.[10] In the present study, quizzes showed dissimilar and poor KR-20 reliabilities; however, when combined over multiple occasions, the combined reliability was higher than the exam KR-20 reliability on its own. Distributed over multiple occasions, precision in measuring student's learning (reliability) was improved.

DOI: https://doi.org/10.24926/iip.v12i1.2235

Quizzes over multiple testing occasions can improve the composite-reliability of course-grades (over an exam alone) and can bolster validation evidence for the generalization inference, when a grade is used in a high-stakes decision-making situation. Some courses may not have course-time for more or longer exams; quizzes can be another means to improve reliability of a course's letter-grades. This enhanced reliability (if inadequate from an exam alone) can help high-stakes decision-making when assessment of students' learning from a course is backed by sound and fair validation evidence. Quizzes can have both formative and summative assessment roles.

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. Tweed M, Miola J. Legal vulnerability of assessment tools. *Med Teach.* 2001; 23(3):312-314. doi: 10.1080/014215901300353922
3. Gleason BL, Peeters MJ, Resman-Targoff BH, Karr S, McBane S, Kelley K, Thomas T, Denetclaw TH. An active-learning strategies primer for achieving ability-based educational outcomes. *Am J Pharm Educ* 2011; 75(9):article 186. doi: 10.5688/ajpe759186
4. Larson DP, Butler AC, Roediger III HL. Test-enhanced learning in medical education. *Med Educ*. 2008; 42(10):959-966. doi: 10.1111/j.1365-2923.2008.03124.x
5. Exam Quality Through the Use of Psychometric Analysis. Examsoft.org. https://examsoft.com/resources/exam-quality-use-psychometric-analysis. Published October 24, 2019. Accessed January 3, 2021.
6. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to their Development and Use.* New York, NY: Oxford University Press; 2015.
7. 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.*
8. 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
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. 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

DOI: https://doi.org/10.24926/iip.v12i1.2235