A Dissertation

entitled

Using Generalizability Theory to Improve Assessment within Pharmacy Education

by

Michael Joseph Peeters

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Doctoral of Philosophy in Foundations of Education: Research and Measurement

_____

Dr. Gregory E. Stone, Committee Chair

_____

Dr. Noela A. Haughton, Committee Member

_____

Dr. M. Ken Cor, Committee Member

_____

Dr. Spencer E. Harpe, Committee Member

_____

Dr. Amanda Bryant-Friedrich, Dean
College of Graduate Studies

The University of Toledo

December 2019

An Abstract of

Introducing Generalizability Theory into Pharmacy Education
by
Michael Joseph Peeters

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the
Doctor of Philosophy Degree in
Educational Research and Measurement

The University of Toledo
December 2019

Validity is a foundational aspect of learning assessments; the process of validation is vital. Validation is context-specific and needs to be examined for test-scores from each learning assessment that will be used in decision-making—especially high-stakes decision-making which may affect progression in a pharmacy education program, graduation, and/or licensing. Kane's Framework for Validation has four inferences (scoring, generalization, extrapolation, and implications), of which generalization is an important first step. Generalizability Theory (G-Theory) can do this though it has been used very rarely in pharmacy education. This dissertation is focused on demonstrating use of G-Theory for reliability evidence (generalization inference) with three common applications of learning assessments in pharmacy education—a performance-based assessment, multiple examinations, and multiple quizzes in preparation for an examination. Reliability was reported for each application. As well, variance was found, and optimization of test parameters was explored for each. Rigor can and should be a focus for every high-stakes assessment, including (and especially with) performance-based assessments. In the unique educational context and resources of each college/school of pharmacy, reliability and the impact of alterations in test parameters

should be examined. Following from this, the impact on reliability in varying the number of exam questions, and varying the weight of quizzes was also demonstrated. At each institution and for various learning assessments, G-Theory provides validation evidence that can show rigor for use of learning assessment scores. G-Theory also allows exploration for customization of test parameters. It would be prudent for G-Theory to evolve to much wider use in pharmacy education.

Acknowledgements

The author gratefully thanks his committee and other colleagues that helped with various aspects of his coursework projects and this dissertation, including data acquisition as well as collaborating on various scholarly activities.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AERA.....American Educational Research Association
ANOVA .....Analysis of Variance (inferential statistical test)
APA .....American Psychological Association

D-Studies.....Decision-studies

G-coefficient .....Generalizability coefficient
G-Study .....Generalizability Study
G-Theory.....Generalizability Theory

KR-20.....Kuder-Richardson Formula 20

MCQs.....Multiple-choice questions

NCME .....National Council on Measurement in Education

OSCE .....Objective Structured Clinical Exam (testing format)

PharmD .....Doctor of Pharmacy

SOAP .....Subjective-Objective-Assessment-Plan (written communication in patient care)

UT .....University of Toledo

List of Symbols

Generalizability Theory nomenclature:

x.....crossed

: .....nested

° .....facet (such as quizzes) that is different across categories (is “fixed” in multivariate Generalizability Theory)

• .....facet (such as persons) that is same across categories (in multivariate Generalizability Theory)

Chapter One

Introduction

Background of the Study

Fairness to students is important in the assessment of learning. In addition to a discussion of the criticality of validity and reliability, the latest edition of *The Educational and Psychological Testing Standards* (called "*The Standards*" hereafter) introduced the concept of Fairness In Testing (AERA, APA & NCME, 2014). *The Standards* describe fairness as "a fundamental validity issue" and "[the] chapter addresses measurement bias as a central threat to fairness in testing" (pg.49). Thus, one key aspect of fairness in testing is test validation—the concept of providing evidence to support use and interpretation of learning assessment results.

Though not only for fairness, validity is an essential aspect of test score quality. Indeed, *The Standards* describe it as "*the most* fundamental consideration in developing and evaluating tests" (AERA et al, 2014, pg.11; italics emphasis added). In fact, one review of legal court cases regarding quality assessment programs showed "strong congruence between *The Standards* and how validity is viewed in the courts, and that testing agencies that conform to these guidelines [*The Standards*] are likely to withstand legal scrutiny" (Sireci & Parker, 2006, pg.27). Even more specifically, the reliability of a test has been a key vulnerability within litigation of health-professions education (Tweed & Miola, 2001).

Legal challenges have been most contentious when test scores were used in a high-stakes setting. According to *The Standards,* the stakes of a setting refers to the

importance of decisions that are made based on scores from a test (AERA et al, 2014). A low-stakes test represents an assessment of learning "used to provide results that have only minor or indirect consequences for individuals, programs, or institutions involved in the testing" (pg. 221). In contrast, a high-stakes test is one where "results have important, direct consequences for individuals, programs, or institutions involved in the testing" (pg. 219). Thus, in high-stakes scenarios, validation of test and assessment process become critical (Peeters & Cor, 2019).

Problem Statement

Tests within classroom settings of higher education are often used without rigorous validation. Having validation evidence is an aspect of fair testing and especially critical with high stakes testing (Peeters & Cor, 2019). The problem addressed in this study was a lack of rigorous validation evidence, especially for high-stakes testing, within pharmacy education.

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., every course is a pre-requisite for future courses). Failure of any single course can cause a student to fall out-of-sync with the next offering of that course in the following year, and so that student's graduation will be delayed by at least one year. Furthermore, health-profession accreditors require reporting attrition and any delay in students' progression. For instance, the Accreditation Council for Pharmacy Education deems it important for site reviewers to evaluate students' on-time graduation from a PharmD program, as well as colleges/schools of pharmacy to disclose this to the public via that college/school's

webpage (Accreditation Council for Pharmacy Education, 2015). With this need for timely progression and in this "Age of Accountability" (DeLuca, 2012), educators' responsibility for educational outcomes are under increased scrutiny from other stakeholders (e.g., administrators, lawyers, parents) that each course be conducted scrupulously and fairly, especially if failure of that course could halt progress in a curriculum and result in additional tuition and fee costs for a student. Thus, delaying students' progression within a health-professions degree program should be seen as a high-stakes situation, and so any testing used for decisions to delay progression (i.e., important, direct consequence for a non-progressing student) should meet standards for high-stakes testing.

The *Standards* provide guidance on how to seek validity evidence and what sources to use (AERA et al, 2014). In *Educational Measurement,* Kane (2006) presents a summary of validation, specifically describing evidential needs for making inferences from test scores. He defines validation as a process of generating evidence that enables investigators to feel confident in the validity of their inferences, specifically regarding use and interpretation of scores from a test. Kane's Framework for Validation is argument-based; it starts with stating an argument for a specific use and interpretation of scores from a learning assessment and is followed by generating evidence for inferences of *scoring*, *generalization*, *extrapolation,* and *implications*. This dissertation has used Kane's Framework for Validation.

As an accreditor of education in a health-profession, the Accreditation Council for Pharmacy Education (2015) notes the importance of validation, using terms such as *standardized*, *valid*, *reliable*, and *validating*. However, the reporting of validation

evidence within pharmacy education literature has been limited (Hoover, Jung, Jacobs & Peeters, 2013), and so improving the psychometric rigor of tests in pharmacy education seems needed. An apparent and continued lack of awareness among pharmacy academicians regarding the need for evidence for validation of tests has also been noted (Peeters & Martin, 2017; Peeters & Cor, 2019). Validation efforts would be beneficial for a number of stakeholders including colleges/schools of pharmacy (who would demonstrate their commitment to holding their students to high-quality standards for competence in the practice of pharmacy), faculty and administrators (who would be better assured of test quality), and students (who would benefit from an improved focus on fairness in testing and more accurate assessments of their abilities).

It is important to recognize and emphasize that validation should not be conceptualized as one validation study but as a series of investigations (Cook et al, 2015; Kane, 2006). Use of Kane's Framework for Validation has repeatedly been highlighted in reviews from medical and pharmacy education (Cook et al, 2015; Peeters & Martin, 2017). Validation can require several studies providing supporting evidence for *scoring*, *generalization*, *extrapolation,* and *implications*. In some instances, *scoring* evidence *may* be overlooked if adequate *generalization* evidence is demonstrated; however, if *generalization* evidence is insufficient, *scoring* should be examined. Thus, an important first step in validation of test score use is providing evidence to support *generalization* (Peeters & Martin, 2017), and one notable tool for generating this evidence is Generalizability Theory (G-Theory).

G-Theory has been around since 1972 and used extensively in medical education where it has become a standard for showing rigor of testing (Crossley et al, 2007). Noting

the array of contextual and implementation differences among over one-hundred medical schools with different implementation and versions of similar tests, Crossley and colleagues (2007) concluded that "Generali[z]ability theory 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" (pg.927). 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 been only three investigations that have reported use of G-Theory in the pharmacy education literature (Munoz et al, 2005; Peeters, Serres & Gundrum, 2013b; Cor & Peeters, 2015), and no studies that investigated Doctor of Pharmacy students. Because pharmacy education shares variety and complexity of assessments similar to medical education, G-Theory could also be particularly useful in pharmacy education.

Towards a goal of more commonly providing validation evidence in pharmacy education, one step forward could be providing worked examples of how to obtain *generalization* evidence using G-Theory. As many educators would attest, students can learn substantially better when given worked examples (Atkinson, Derry, Renke & Worthem, 2000; Renkl, 2002). Thus, worked examples of G-Theory use specific to pharmacy education *and* explicitly focused on Kane's Framework for Validation should be beneficial.

Purpose and Importance of the Study

This dissertation focused on generating key validation evidence for inference of test scores in pharmacy education. Within Kane's Framework for Validation, different

types of evidence should be sought to support inferences, of *scoring*, *generalization*, *extrapolation*, and *implications*, made using test scores. This should be done for each assessment of learning. As initial validation evidence for generalization, reliability can and should be analyzed following even the first administration of a learning assessment (Zibrowski, Myers, Norman, & Goldszmidt, 2011). Unlike extrapolation and implications evidence, educators do not need to wait to observe a future consequence for generalization evidence. Educators can analyze reliability post-administration with every cohort of test-takers. Thus, reliability is a crucial (initial) quality indicator for use and interpretation of test scores from a test (AERA et al, 2014; Peeters, Beltyukova & Martin, 2013a; Tavakol & Dennick, 2012). This should be evaluated for all types of learning assessments. High reliability is especially important for testing and decision-making in the health professions (Peeters & Cor, 2019; Tavakol & Dennick, 2012). As further evidence of its importance and noted earlier in this chapter, reliability of tests has been a key vulnerability in litigation for health-professions education (Tweed & Miola, 2001).

Scientifically-rigorous, high-stakes testing situations is a notable application of reliability to pharmacy education (Peeters & Cor, 2019). Unfortunately, while many colleges/schools of pharmacy often employ some tests that their educators may describe as “high-stakes”, reliability has been infrequently reported in the pharmacy education literature (Hoover et al, 2015). Furthermore, poor reliability of various performance-based assessments has often been documented in medical education (Brannick, Erol-Korkmaz, & Prewett, 2011). It can be challenging to implement a performance-based test that is sufficiently reliable (van der Vleuten, 1995). In fact, the overarching concept of validation in educational testing, which includes reliability, appears relatively new for

many pharmacy educators (who are not often formally-trained in education). As yet there does not appear to be any explicit examples of validation in the pharmacy education literature. However, at this time Peeters & Martin (2017) have provided an introductory primer on validation for pharmacy academics. The purpose of this dissertation was to demonstrate use of G-Theory to: (a) analyze the reliability of various assessments of learning in pharmacy education to create generalization evidence towards validation for each of these, and (b) optimize test parameters for future iterations of a learning assessment.

While Cronbach's alpha and KR-20 (for dichotomous data) are very common reliability coefficients for internal consistency (e.g., among items on a multiple-choice test), G-Theory is a notable and powerful extension. G-Theory will analyze reliability from multiple sources of error, such as multiple items *and* multiple examinations on different occasions, or multiple performance-based assessment tasks *and* each task having its own (multiple) items and (multiple) raters. G-Theory can integrate the different sources of error and summarize reliability with these complex test designs. In short, G-Theory is a method for producing validation evidence for a generalization inference. To nudge use of G-Theory forward as a helpful tool for validation of high-stakes testing within pharmacy education, the aim of this dissertation was to show multiple applications of G-Theory as validation evidence specifically in pharmacy education. Figure 1 illustrates an overview of the three sub-studies in this dissertation. The first application with performance-based assessment should be most straightforward, however performance-based assessments are not the only assessments in pharmacy education. Assessments of learning within courses should also be rigorous. The second application

is the classic approach with multiple examinations within a course. With increased use of quizzes more recently (e.g., within Team-Based Learning or Flipped Classroom pedagogies), the third application investigates the integration of quizzes with course examinations.

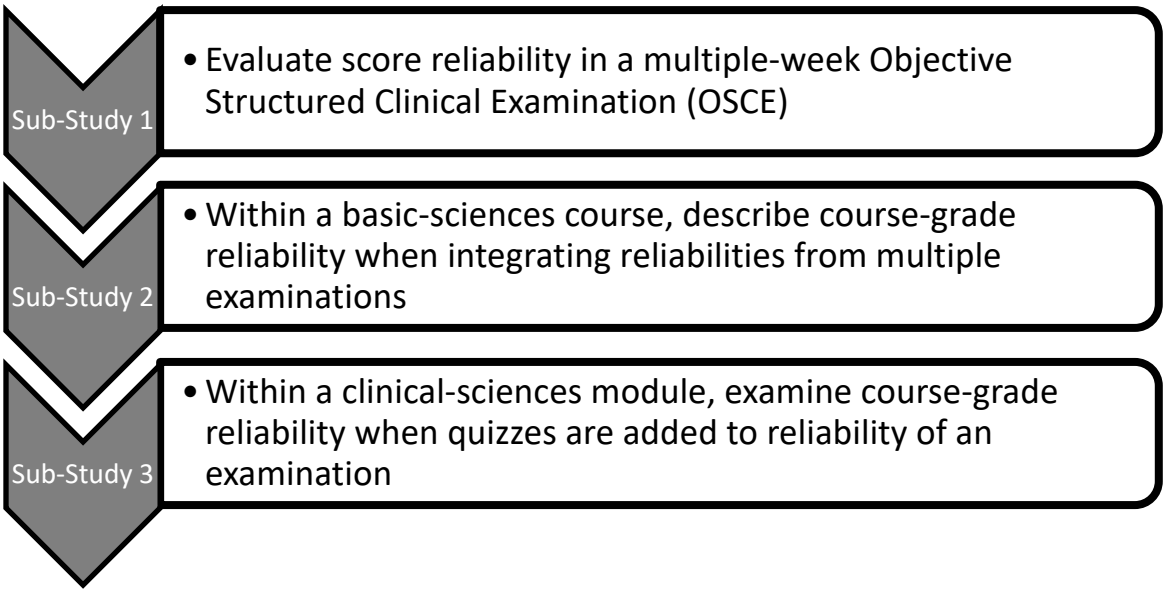

Figure 1. Outline of dissertation sub-studies to demonstrate application of Generalizability Theory in pharmacy education

Beyond G-Theory's use in characterizing reliability following administration of a test, another important benefit to its use is to extrapolate and determine the best use of (limited) resources for each local institution to provide acceptably-reliable tests. More background description of these insightful D-Studies is within the next dissertation chapter. Reader insights with demonstrating utility of G-Theory's D-Studies is central to Sub-Study 2 (course exams) and Sub-Study 3 (course quizzes).

Learning assessments can differ in different locations. Hodges (2003) points out the many variations of assessment formats that can be modified for different settings at various institutions, for a similar OSCE of clinical skills. For instance, one college/school

of pharmacy might have a dedicated space they could borrow from their medical school for an OSCE whereas others might need to be more creative and use faculty offices. Another example is whether an OSCE is administered over one-day, two-days or multiple-days. Further, one college/school of pharmacy might use multiple raters in each station while another uses just one. Thus, reliability of an OSCE administered at one institution cannot be assumed when considering an OSCE version at another institution; each needs its own distinct evaluation of reliability.

After a fundamental reliability study (i.e., a Generalizability-Study or G-Study) is conducted using G-Theory, it can also be helpful to further explore reliability. This can be explored for variance components that describe the data variance that is attributable to different test parameters (facets) that were specified in the G-Theory test design. Furthermore, using the variance components from G-Theory's initial G-Study, decision-studies (D-Studies) can estimate the reliability with changes to various test parameters; showing what effect changing those facets could have on the overall reliability. For example, the overall reliability of an OSCE will likely have dissimilar influences from increasing the number of test stations, increasing the number of raters, or increasing the number of testing occasions. D-Studies with those varied testing facets may help optimize reliability given the available resources (e.g., space, faculty, time) at a specific institution.

Definition of Terms

Within this dissertation, several critical terms are used. These terms include *validity, reliability, validation, learning assessments,* and *performance-based assessments*.

Simply put, *validity* refers to the accuracy of test score use and interpretation (AERA et al, 2014). *Reliability* is narrower than validity and is often regarded as "precision" (as opposed to "accuracy") of scores. It is concerned with consistency in statistically-discriminating among test-takers using the various items of the test (Haertl, 2006). Reliability is one aspect of validity, with validity extending beyond reliability (AERA et al, 2014). That said, reliability is a key component to scientifically-rigorous testing (Peeters et al, 2013a). While validity has numerous methods to describe aspects of it, reliability is often distilled to a coefficient (e.g., internal consistency, inter-rater reliability, g-coefficient). *Validation* of educational assessments and procedures is the process of generating evidence to support valid uses and inferences from scores of a test (AERA et al, 2014; Peeters & Martin, 2017). Noting this, a common misperception needs explicit mention. Neither tests nor test scores can be validated; instead, it is a claim (inference) and decisions from use of test scores that can be validated (Kane, 2006). Additionally, some literature (including pharmacy education) asserts "assessment" as programmatic assessment. However, this dissertation (like much other literature; e.g., Fielding & Regehr, 2017) is focused on "assessments" as tests of students' learning. Going forward in this dissertation, a "test" will be referred to as a *learning assessment* or simply *assessment*. Furthermore, a *performance-based assessment* is a specific type of learning assessment that, in health-professions education, is most often focused on skills. One format of performance-based assessment most often considered the gold-standard for this type in health-professions education is the *objective structured clinical examination* (OSCE) (Boursicot, Roberts, & Burdick, 2014; Cor & Peeters, 2015; Hodges, 2003; Khan, Ramachandran, Gaunt, & Pushkar, 2013; Newble, 2004).

Theoretical Frameworks

Two theoretical frameworks are used in this dissertation. The first, which is a main focus for this dissertation, is Kane's Framework for Validation of learning assessments. The second, Miller's Pyramid, categorizes learning assessments by use.

Validation is for educational assessments, as well as procedures used in decision-making within education. It is the process of generating evidence to support valid uses and inferences from test scores (Peeters & Martin, 2017). As in Figure 2, Kane provides a framework for validation (Cook et al, 2015; Kane, 2006; Peeters & Martin, 2017). Within, *scoring* evidence comes before *generalization* evidence, proceeds to *extrapolation* evidence, and then to *implications* evidence.

Figure 2. Levels of inference evidence within Kane's Framework for Validation (from Peeters & Martin, 2017)

Generalization is a key inference for validation evidence with assessments of students' learning within Kane's Framework for Validation (Cook et al, 2015; Peeters & Martin, 2017). Do scores from an assessment generalize to other similar students (who will subsequently take the same classes at that same college/school of pharmacy)? A reliability coefficient is a primary form for this generalization evidence. Notably, G-Theory is one tool that can be used to construct generalization evidence (i.e., reliability) for validation of a learning assessment (Peeters & Martin, 2017).

Because the three sub-studies of this dissertation employ different types of learning assessments, a framework of Miller's Pyramid can be helpful to categorize these

different types of learning assessments. While Miller's Pyramid is discussed at length in the next chapter, in short it has four levels (from its base) labelled as "knows", "knows how", "shows how", and then "does" at the top. Notable assessments within "knows" and "knows how" are multiple-choice examinations (e.g., this dissertation's Sub-Study 2 and Sub-Study 3), along with the "shows how" of performance-based assessments (e.g., this dissertation's Sub-Study 1) as especially important to health-professions education.

Research Questions

Since this proposed investigation involves three sub-studies of G-Theory use, there are nine research questions that will be addressed (i.e., three for each sub-study). Figure 3 provides a mapping of the learning assessments and research questions within this dissertation. As can be seen, each is at a different level of progression within Miller's Pyramid and each should have its own generalization evidence from Kane's Framework for Validation. The structure of the three research questions for each sub-study is: (a) define the reliability, (b) report variance components from the G-Study, and (c) identify patterns from the D-Studies.

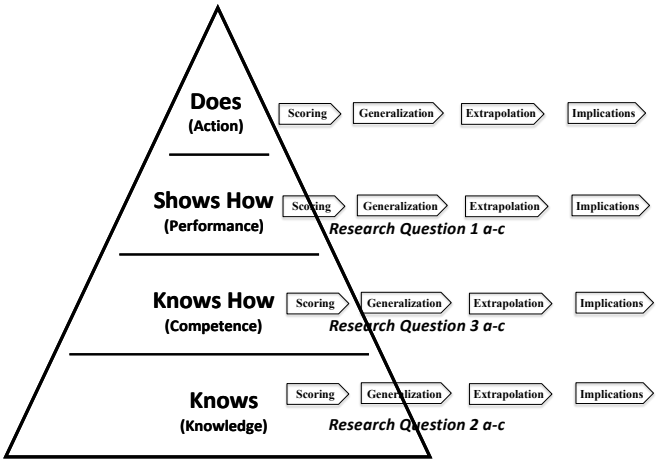

Figure 3. Model of Kane's Framework for Validation nested within each category of Miller's Pyramid (with dissertation research questions mapped)

These questions are as follows:

Research Question 1a. What was the reliability of an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo?

Research Question 1b. For an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo, what were the relative contributions of occasions and stations to examination score variance?

Research Question 1c. What would be the optimal number of stations for each of three weeks of an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo?

Research Question 2a. What was the composite course-level reliability for multiple examinations in a basic-science course that involved 1st-year Doctor of Pharmacy students at the University of Toledo?

Research Question 2b. What were the relative contributions to variance in scores from examination *occasions* and examination *items* that involved 1st-year Doctor of Pharmacy students in a basic-science course at the University of Toledo?

Research Question 2c. How would the number of examination *occasions* and the number of examination *items* be optimized for a basic-science course that involved 1st-year Doctor of Pharmacy students at the University of Toledo?

Research Question 3a. In a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University, did quizzes add to the reliability of an examination?

Research Question 3b. What were the contributions to variance of scores for quiz-items and exam-items in a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University?

Research Question 3c. How could the *number* and *weighting* of items from quizzes be optimized beyond items from the exam for module-grade reliability in a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University?

Organization of this Dissertation

Through the three sub-studies, this dissertation describes methods and results from three different applications of learning assessments that are common in pharmacy education. Each application is intended to provide a different type of example of how G-Theory could be used. The first sub-study demonstrates how G-Theory could calculate, along with optimize, the reliability of a performance-based assessment of skills within an OSCE. (In the 'bible' of G-Theory, Brennan (2010) asserts that "Generalizability Theory is particularly well suited to evaluating assessments that are based on ratings of human performance." (pg. 117). That is, analyzing a performance-based assessment may most easily demonstrate G-Theory initially, before progressing to other applications.) The second sub-study demonstrates how G-Theory could calculate the composite reliability of an entire course's final grades, through combining reliabilities from multiple examinations into one composite reliability for course-grades. (This is a very common teaching scenario and the subsequent D-Studies are a highlight from it.) The third sub-study demonstrates how G-Theory can integrate short quizzes (i.e., brief assessments of students' learning) with a longer exam, resulting in a different example of composite

reliability for course-grades. (In some coursework in some institutions, use of quizzes is increasing either at the beginning of class meetings during a flipped-lecture or team-based learning that require outside preparation for class, or even to stimulate active-learning during lecture.)

Chapter Two

Review of the Literature

Introduction

For over 100 years, a debate has taken place within the Social Sciences between quantitative and qualitative research approaches (Onwuegbuzie & Leech, 2005). The arguments center around the use of quantitative research methods, within positivist and post-positivist paradigms, versus the use of qualitative research methods, typically aligned with constructivist or critical theory paradigms (Bergman, 2012). Among the many differences between these methods, lies generalizability. Quantitative researchers view broad generalization as a key goal, while qualitative researchers, focused primarily on description and do not generally attempt to generalize beyond their limited participants. Thus, within education, researchers follow many pathways, using either or both quantitative and qualitative designs, with some attempting to generalize while others focus on their unique situation and leave the reader to infer if described aspects of a researcher's experience will transfer to that reader's setting. While this lack of a single direction may be positive when considering the broad nature of research, it also leads to some substantive issues.

Educational research faces a difficult challenge. Discussing the power of context among educational settings, Berliner (2002) stated, "we do our science under conditions that physical scientists find intolerable" (pg. 19). He goes on to suggest that "because of the myriad interactions, doing educational science seems very difficult, while science in other fields seems easier" (pg. 19). In a more recent editorial, Norman (2017) discussed this quantitative-qualitative debate and the reproducibility of study findings. He pointed

out that findings from some disciplines and their predominant quantitative research methods appear more replicable and generalizable than other qualitative approaches. For instance, 53% of cognitive psychology investigations with more experimental quantitative study designs were replicable, versus 29% of social psychology investigations with more non-experimental quantitative and qualitative study designs (Open Science Collaboration, 2015). It appears that findings from quantitative methods may be more reproducible than qualitative methods. However, as Van Bavel and colleagues (2016) discovered, context-specificity (i.e., unique specifics within one context that other contexts may not have) was associated with reproducibility of study findings regardless of disciplines, study designs or methods. Indeed, it would seem that no discipline provides 100% generalizable or reproducible findings. Context-specific factors are found in every discipline and thus all disciplines appear context-bound to some degree. The social sciences, including education, are particularly prone to contextual factors. Findings are much less transferable to other settings, as those social settings become increasingly different through various cultures and times. The nature of educational research and its use of multiple models along with a general lack of ability to develop experimental models may exacerbate this issue within the field. Often, the specifics associated with any context, limit the ability of other educators to replicate the findings study authors have reported from their educational institution. Outside of logistical limitations (e.g., time, space, staff resources, different curricula), education environment factors (e.g., characteristics of the educator, students, institution, course content) will inevitably differ from one setting and institution to another. Thus, there appears to be context-specific (qualitative features) within quantitative analyses.

Just like context-specificity in educational research, educational *testing* also has contextual challenges. In parallel with educational research, educational testing can be assumed to have generalizable (quantitative) and context-specific (qualitative) aspects. As quantitative standards, evidence for validity and reliability are evaluated, while qualitative standards of transparency, responsiveness and transferability are explored (Cook et al, 2015). With integration of quantitative and qualitative, generalization in educational testing should best be viewed as generalizing to future cohorts at the same educational institution as opposed to other institutions (Peeters & Martin, 2017). Questions of whom will be tested, over what content, where and how a test will be administered, as well as how a test's scores might be used afterwards are specific issues of validity and reliability that can be particular for each test administration at every individual institution. Because these context-specific aspects can differ with each testing occasion (and especially between different institutions), the validity and reliability with use of any learning assessment should be clarified at each institution. In fact, validity and reliability are not properties of a learning assessment but are characteristics that describe a specific use of assessment scores from a particular group of test-takers (Peeters et al, 2013a). The process of generating this evidence is termed *validation* (Peeters & Martin, 2017), and is a process specific for each learning assessment in every educational institution. Validation is mainly a local process and rarely a national process. This process is essentially empirical (i.e., data-driven)—providing evidence to support any use and/or interpretation of scores from a learning assessment.

This picture should stimulate a need for each institution's own analysis for some aspects of their educational testing. This chapter will explore the key concept of validity,

its relationship to validation, along with a framework for validation (especially for use within high-stakes testing). Nested within this discussion of validity and validation, generalization is one important evidence for validation, with reliability as the prominent generalization evidence. As a key element in the establishing validity, limitations with traditional coefficients of reliability will be discussed. Furthermore, a prominent framework for understanding different types of learning assessments will be described, along with its application to pharmacy education (including its rigor therein). A discussion of what *high-stakes testing* implies follows. Based on limitations with traditional use of reliability, the author will subsequently propose consideration of Generalizability Theory (G-Theory) as a helpful tool when learning assessments have multiple sources of variance. The abundant use of G-Theory in medical education will be summarized, along with its currently rare use in pharmacy education including discussing its opportunities for future expansion. Lastly, barriers to G-Theory use will be described, such as its mythical complexity and handful of specialized software programs. Of note, this dissertation has focused on issues relevant to pharmacy education and was not meant to cover all disciplines within education as a broader field.

Validity and Validation

"Validity refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (AERA et al, pg.11). Boorsbom and colleagues (2004) remind us that validity is an ontological claim. Meanwhile, validation as the process to gather validity evidence, is an epistemological concept. A practical difference between an ontological claim and an epistemological

concept is that conclusive evidence for validity (ontological claim) cannot be gathered, although tangible evidence for the activity of validation (epistemological concept) can. "Validation is the kind of activity researchers undertake to find out whether a test has validity. Validation is more like theory testing: the muddling around in the data to find out which way to go" (Boorsbom et al, 2004, pg.1063). Thus, validation is a practical extension of validity, and validation should be a practical focus for rigorous learning assessments.

Said another way, validity (an ontological claim) can be seen as making suggestions about generalizing. There can be strong, weak, or a lack of validity evidence to support a learning assessment. Evidence for validity can be helpful for a new or novel learning assessment—one that is envisioned for use in multiple settings. Alternatively, there is validation (an epistemological concept). Validation can be seen as a practical extension that tests the generalizable validity in a specific, localized use by a single institution. Validation uses real-world data to confirm or disconfirm validity for an individual institution's contextual experience. Thus, validation is where continued efforts should be targeted for each use of an already-developed learning assessment, and with the differing contexts inherently among institutions, it needs to be done at every college/school of pharmacy.

Validation Framework. Kane has described a four-stage framework for validation, that includes evidence from *scoring, generalization, extrapolation, and implications* (Figure 2 reproduced below from Chapter One of this dissertation). Kane's Framework for Validation begins with asserting (i.e., making an argument) for a specific use of scores from a learning assessment, and then collecting evidence to support that

argument (Cook et al, 2015, Kane, 2013; Peeters & Martin, 2017). It is important to recognize and emphasize that validation should not be conceptualized as one "validation study" (Cook et al, 2015; Kane, 2013). Each of Kane's stages of evidence to support an argument's inference could be a separate investigation. These evidences for argument can and should build on one another. Cook et al (2015) summarizes this concept with "how to prioritize evidence?" (pp.560-561). Evidence for argument from *scoring* is based on translating observations or responses into one or more scores. Evidence for *generalization* involves how test-items perform towards creating rigorous total-scores following administering a learning assessment. Evidence for the *extrapolation* is focused on how scores from this assessment are reflections of real-world performance. Lastly, evidence for *implications* involves how scores from an assessment are applied to inform a decision or action.

Figure 2 (reproduced from Chapter One). Levels of inference evidence within Kane's Framework for Validation (from Peeters & Martin, 2017)

Validation is data-driven and may require several studies providing supporting evidence for an argument, with a separate investigation for each evidence (*scoring*, *generalization*, *extrapolation, implications)*. Many times, evidence for *scoring* can be overlooked if adequate *generalization* evidence is demonstrated; however, if *generalization* evidence is insufficient, *scoring* needs evaluation (Kane, 2013). Thus, an important first step in validation is providing evidence to support *generalization* (Peeters & Martin, 2017); reliability is central to this. While *implication* and associated decisions

may be, overall, the most important inferences in the validity argument (Cook et al, 2015), evidence for the *generalization* argument can be generated with every administration of a learning assessment (including the first) and is key initial evidence demonstrating rigor in measurement.

As medical education continues to strive at improving the quality of its learning assessments, Kane's Framework for Validation has received considerable attention (Cook et al, 2015; Taveres, Brydges, Myre, Prpic, Turner, Yelle, & Huiskamp, 2018). This approach to validation appears germane and helpful for pharmacy education as well (Peeters & Martin, 2017). The Accreditation Council for Pharmacy Education has specified that Doctor of Pharmacy programs must provide evidence of validation for their learning assessments (Accreditation Council for Pharmacy Education, 2015). Unfortunately, the Accreditation Council for Pharmacy Education does not suggest how this might be accomplished. This critical omission, in part, sparked the present investigation. Learned guidance may be helpful.

Because each learning assessment is context-bound (i.e., specific for one group of learners in one environment), validation should be accomplished at every institution using their own data. Frequently misunderstood in the health-professions (as well as other educational settings) is the idea that evidence for validation (such as reliability) can be generalized to other institutions and other contexts. Validation is evidence from one specific group learners in one specific educational context at one specific institution. Within this misconception, one myth among some health-profession educators is that educational tests and other psychometrics instruments have fixed, immutable characteristics, and that a reliability observed with one sample, in one setting, is

generalizable to all other samples elsewhere (Zibrowski et al, 2011). Current understanding of validity and reliability differs from this fixed-concept, suggesting that these measures can and do change, and should be more carefully monitored as the sample of test-takers, programmatic uses, and interpretations of assessment results change. A learning assessment that is reliable in one cohort should be generalizable for similar use in similar future cohorts at the same institution, but this reliability may not replicate to different learners at other institutions. As a result, each institution should generate their own validation evidence for how they are using their learning assessment scores.

According to Kane's Framework for Validation, *generalization* is an important initial evidence towards a validation argument. A perceived lack of awareness of need for this within pharmacy education sparked the present investigation. A first step in educating pharmacy academics was a primer on validation of learning assessments for pharmacy education from this current dissertation author (Peeters & Martin, 2017). A next step could be to provide worked examples of validation evidence for common learning assessments in pharmacy education. Worked examples appear to be effective for initially learning a new process (Atkinson et al, 2000; Renkl, 2002). For many pharmacy academicians not formally trained in education, validation can be a new concept and process.

Reliability as Validation's Generalization Evidence

In Kane's Framework for Validation, generalization evidence has a prominent role for helping determine initial quality of a learning assessment (Peeters & Martin, 2017). As generalization evidence, reliability is key. Reliability is even more important for fairness with higher-stakes testing (Peeters & Cor, 2019). This reliability coefficient

provides a snapshot of consistency in measurement, though it may change (somewhat) with each new testing situation. Reliability can and should be calculated following each test administration (Zibrowski et al, 2011). Using other investigators' reported reliability is problematic.

Reliability is often described as an essential yet insufficient requirement for the establishment of validity (Cook et al, 2015; Downing, 2003; Kane, 2006; Peeters et al, 2013a; Peeters & Martin, 2017). That is, while acceptable reliability is required, it is not the sole requirement for establishing validity. Other validity evidence is subsequently needed as well. That said, reliability is a key attribute to fair testing (AERA et al, 2014), and can help support psychometric rigor of a testing process (Peeters et al, 2013a). There are a number of approaches to assessing reliability, with most common approaches using Classical Test Theory. Within Classical Test Theory, commonly-used indices include internal consistency, test-retest stability, and inter-rater reliability. A notable assumption of Classical Test Theory is that measurement error can only consider one source at a time (Peng, 2007). However, a performance-based assessment with multiple tasks (items), two raters scoring students over multiple weeks of testing; this would include at least three concurrent sources of measurement error in a single assessment (from tasks, raters and weeks/occasions).

Another framework in which reliability may be evaluated is Generalizability Theory (G-Theory). G-Theory is a notable extension of Classical Test Theory based on an Analysis of Variance (ANOVA) model. Recall that an ANOVA differentiates within-group variance from variance due to between-group differences. Researchers within an experimental design are interested in minimizing within-group variation, to investigate

differences between groups. This involves F-tests and p-values. Alternatively, the psychometric G-Theory model investigates an individual subject over multiple tasks and is not worried with F-tests and p-values. Instead, the psychometrician is interested in intra-individual variation over those repeated tasks, and with reliability among those repeated scores.

Furthermore, G-Theory has been described as a unifying theory for reliability within Classical Test Theory (Streiner et al, 2015). Reliability indices 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 process-level 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 Classical Test Theory and G-Theory; while Classical Test Theory posits that there is one source and one index of reliability, we know there are multiple sources of measurement error and so multiple indices of reliability can be calculated. G-Theory provides a means to integrate them into one process-level coefficient (Brennan, 2011; Peng, 2007).

Classifying Learning Assessments in Health-Professions Education

Focused towards developing competent practitioners, Miller's Pyramid is used frequently to classify learning assessments within health-professions education including pharmacy education (Figure 4). Miller's Pyramid is a framework that sorts learning assessments into four categories based on their intended purpose. It begins with a base of "knows", proceeds to "knows how", then "shows how", and finally "does". Each category builds on those below it. That is, the pyramid structure suggests that "knows" is

a foundation for “knows how”, and those are foundational for “shows how” and then “does.”

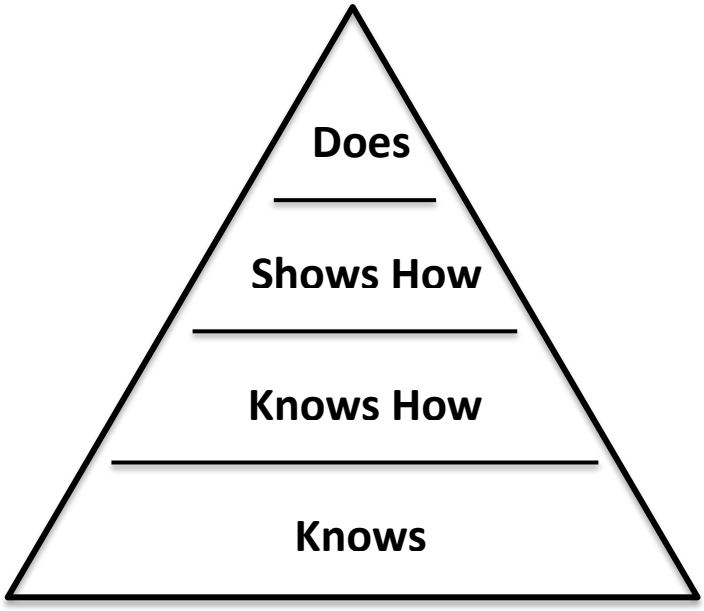

Figure 4. Miller’s Pyramid for the assessment of clinical skills, competence, and performance (from Miller, 1990)

Within Miller's Pyramid, knowledge tests are said to examine students at the “knows” and possibly the “knows how” levels. At these levels in Miller’s Pyramid, learning assessments with multiple-choice questions are commonly used in health-professions education. For many years, “knows” has been widely used in higher education. A long-standing criticism of “knows” (i.e., *recall*) is that many knowledge tests do not assess understanding beyond recall. The next step in Miller’s Pyramid, “knows how” contrasts with recall-oriented “knows”, to further extend knowledge assessments from simple recall to *application* of knowledge. Typical “knows how” assessments are structured as case-based questions. While all item types require

significant attention and skill, “knows how” case-based questions are often particularly difficult to write (Lane, Raymond & Haladyna, 2016). These categories can be overlaid with the cognitive domain of Bloom’s Taxonomy (Anderson & Krathwohl, 2001).

The third category in Miller’s Pyramid, "shows how" is skills-based, and is thus connected much more frequently with performance-based assessments[1]. This category is focused on competence and is outside of the cognitive domain of Bloom’s Taxonomy. Introduced by Harden in 1978, the *objective structured clinical examination* (OSCE) has become the gold-standard method for performance-based assessment in medical and pharmacy education (Khan, Ramachandran, Gaunt & Pushkat, 2013; Sturpe, 2010). This method uses multiple independent sampling (Hanson, Kulasegaram, Coombs, & Herold, 2012; Swanson, Norman & Linn, 1995), with a structured rotation of students through standardized stations (like a traditional “bell-ringer” exam in some anatomy courses). While an OCSE appears to have the highest potential for psychometric rigor among performance-based assessment methods (Boursicot et al, 2014), it also has the potential for the least psychometric rigor, whether nationalized or local. For instance, a locally-developed clinical examination in any health-profession education may rely on incomplete (and shoddy) psychometrics, such as using only an index of interrater reliability for an entire OSCE (which has many more sources of measurement error than only error between raters). This mistaken use may give false-confidence when another unanalyzed aspect of the performance-based assessment could have enormous measurement error, and undermine precise, accurate measurement of students’ abilities.

---

[1] In health-professions education, a *performance-based assessment* is literally an assessment based on performance of a skill (i.e., observation). Because they involve observations of examinees performing skills, a *skills-based assessment* is synonymous.

At the top of Miller's Pyramid, "does" involves in-practice assessments such as those on students' clinical rotations and is the most difficult to assess rigorously. Medical education has attempted to inject rigor into the process through rotation evaluations (Al-Jarallah, Moussa, Shehab & Abdella, 2005), clinical encounter cards (Sherbino, Kulasegaram, Worster & Norman, 2013), and 360-degree workplace-based assessments (Moonen-van Loon, Overeem, Donkers, van der Vleuten & Driessen, 2013).

Reliability of Learning Assessments in Miller's Pyramid. In terms of assessments, the most reliable format for "knows" and "knows how" is using multiple-choice questions (Worthen, 1993). Reliability from a single administration of one multiple-choice-based learning assessment may be sufficiently characterized by a KR-20 for internal consistency. At the "shows how" level, OSCEs have the most promising potential for acceptable reliability compared with other performance-based assessments in medical education (Boursicot et al, 2014), though with an OSCE, increases in sources of potential measurement error are manifold. An internal consistency, inter-rater reliability, and/or inter-task reliability coefficient(s) would not adequately describe the multiple concurrent sources of measurement error. However, G-Theory could provide an overall reliability coefficient that would include all other single-source reliability indices. This dissertation will concentrate primarily on multiple-choice testing and OSCE formats within pharmacy education because of their frequent and justified use within pharmacy education settings. It is notable that regardless of which category of Miller's Pyramid that a specific learning assessment is classified into, Kane's Framework for Validation is needed for use of scores from each learning assessment. Integration of these theoretical framework is illustrated in Figure 5.

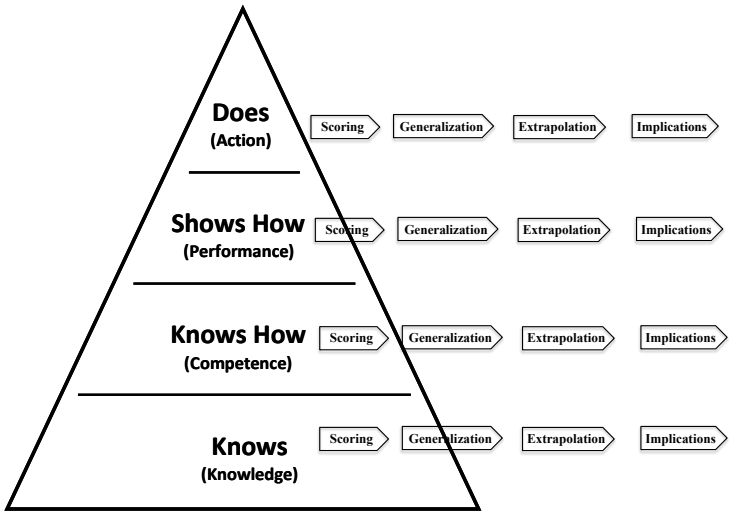

Figure 5. Integration of Miller's Pyramid with Kane's Framework for Validation

Learning Assessments in Pharmacy Education

To become a competent pharmacist, numerous learning assessments of knowledge and skills in various content areas are needed during pharmacy education. As seen in Figure 6, the framework of Miller's Pyramid helps categorize these numerous learning assessments in pharmacy education (Cor & Peeters, 2015). Common learning assessments in pharmacy education in the "knows" category are traditional multiple-choice knowledge-based examinations in various content areas, and is a category most often described with rote, recall-based knowledge. Moving up the pyramid, "knows how" involves application of "knows" knowledge and can be described as "higher-order thinking". Commonly in pharmacy education, this category is assessed using long-answer written cases, including the commonly-framed "SOAP" notes (Subjective/Objective/Assessment/Plan). The third category is "shows how" and this involves assessing skills. In pharmacy education, the gold-standard format to these performance-based assessments is an OSCE. Of note, every "shows how" OSCE is locally-developed within pharmacy education in the United States at this time, with most

conducted in a questionable, suboptimal manner (Sturpe, 2010). The fourth and final category in this Pyramid is "does". As opposed to the standardized, highly-structured setting needed for assessing "shows how", "does' involves assessing learners' performance in the workplace. Most often in pharmacy education, this includes clinical rotation evaluations from preceptors.

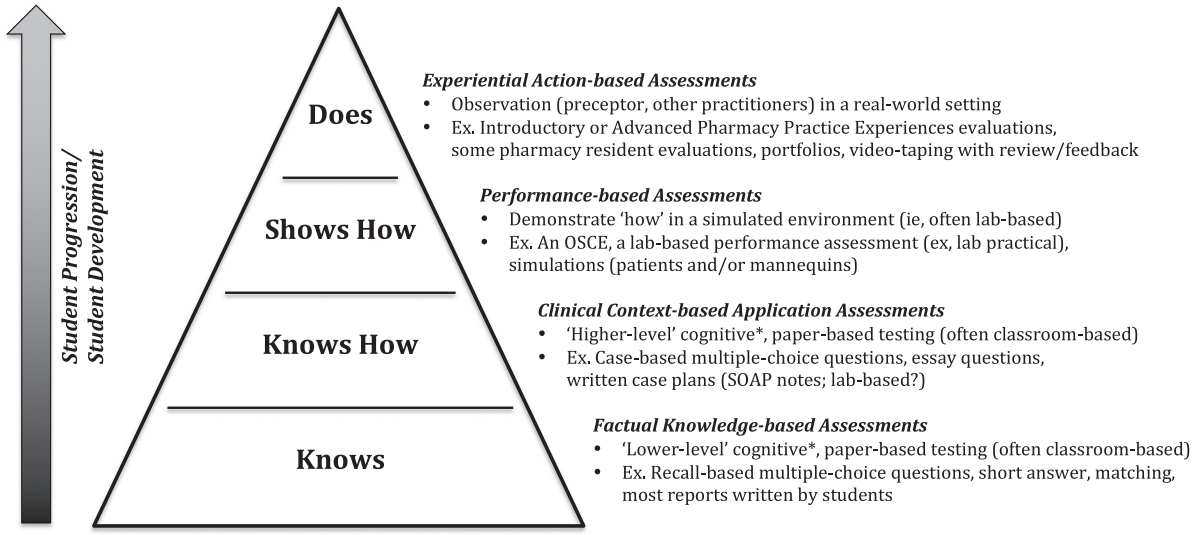

Figure 6. Miller's Pyramid applied to pharmacy education (from Cor & Peeters, 2015)

Questionable Reliabilities. At a national-level, pharmacy education has three notable psychometrically-monitored national written examinations—the Pharmacy College Admission Test, the Pharmacy Curriculum Outcomes Assessment, and the North American Pharmacist Licensure Examination. These standardized written examinations show excellent reliabilities with their intended populations for high-stakes purposes. However, reliability of local examinations in pharmacy education can be another matter. While national-level written examinations are monitored by a psychometrician, local assessments are rarely closely examined by someone with this training. As coefficients of internal consistency, the KR-20 and Cronbach's alpha are the most frequent reliability

indices used in pharmacy education (Hoover et al, 2013). When a written examination is administered on a single occasion (thus having one major source of measurement error), one index of reliability from internal consistency can suffice to describe precision with measurement error for that assessment. Fortunately, this is often reported by a computer-based examination software (e.g., ExamSoft) for a single written examination. Thus, through internal consistency, reliability with a single written examination can easily be identified for a locally-developed written learning assessment.

Reliability with local performance-based assessments in pharmacy education can become a more complex issue and more suspect for reliability, as multiple measurement errors (from multiple skills performed, multiple raters, and/or multiple items rated for each skill) can impact these learning assessment scores. Put simply, internal consistency is not enough for the complexity of performance-based assessment.

Furthermore, very few faculty members at a college/school of pharmacy typically closely and fervently follow the internal consistencies of their single written examinations. Indirectly, evidence for this is from limited reliability reporting in the pharmacy literature (Hoover et al, 2013). Even more so, this pertains to performance-based assessments, as performance-based assessments have multiple sources of error beyond internal consistency. These skills-based assessments do not have a readily and easily computed reliability from examination software, and so the vast majority of performance-based assessments in pharmacy education are without evidence of reliability.

With many colleges/schools of pharmacy prominently relying on a performance-based OSCE evaluation of their students, this seems problematic. Problematic because

reliability is a key initial evidence for validation and test quality, as well as reliability as a central advantage of OSCEs over other performance-based assessments methods (van der Vleuten, 1996). These performance-based examinations can be quite variable among colleges/schools of pharmacy (Sturpe, 2010), and their reliability is being ignored.

Performance-based assessments have an important role in pharmacy education, and the quality of these assessments in pharmacy education needs to improve. It would seem that the validity of these learning assessments could improve with calculated reliability coefficients known, and that changes made to improve reliability could substantially improve the fairness of testing with these learning assessments.

High-Stakes Testing in Pharmacy Education. *High-stakes testing* is a term used and many-times misused in pharmacy education (Peeters & Cor, 2019). Clarity seems needed herein. According to the *Standards for Educational and Psychological Testing* (AERA et al, 2014), "stakes in testing considers the importance of the results to decisions. When the stakes for an individual are high and important decisions depend substantially on test performance, the responsibility for providing evidence supporting a test's intended purposes is greater than might be expected for tests used in low-stakes settings (pg. 188)." Admission to a PharmD program, ability to progress within a lockstep PharmD curriculum where there are prerequisite and corequisite course needs, graduation from a PharmD program, and licensing as a pharmacist, are gradations within high-stakes testing. Thus, even though common language is "high-stakes testing", it is not the testing method that is high-stakes but is the use and interpretation of the test-scores for that testing. No testing method *requires* a high-stakes approach, although some testing methods appear to be more commonly so in pharmacy education.

A common occurrence in pharmacy education is learning assessment used for high-stakes testing without any validation evidence to support this use. An example of a learning assessment used commonly for high-stakes testing in pharmacy education, is a performance-based OSCE (Sturpe, 2010). Notably, an OSCE takes considerable resources to develop and administer properly. The tremendous resource use and substantial effort may at least be two reasons why an OSCE is so often high-stakes in pharmacy education. That said, an acceptable reliability is a vital criterion for fair testing (AERA et al, 2014); determining the reliability of a high-stakes OSCE administration is prudent to fair testing (Peeters & Cor, 2019). The Standards for Educational and Psychological Testing remind that "although it is never possible to achieve perfect accuracy in describing an individual's performance, *efforts* need to be made to *minimize* errors of measurement" (AERA et al, 2014, pg.188; italics added for emphasis).

It is interesting to note that large educational testing companies have developed numerous written standardized tests for high-stakes purposes that focus on "knows" and "knows how" (e.g., Scholastic Aptitude Test, ACT, Graduate Records Examination, Pharmacy College Admission Test, Medical College Admission Test). Fewer large-scale "shows how" assessments have been developed, however there are barriers such as cost, administrative difficulty, and psychometric matters including the establishment of reliability and validity (Gao, Shavelson & Baxter, 1994; Shavelson, Baxter & Gao, 1993; Swanson, Norman, Linn, 1995). For these reasons, locally-developed "shows how" performance-based assessments are often being used at educational institutions. For those performance-based assessments that impose a high-stakes use (e.g., not graduating), psychometrics should be evaluated for fairness and rigor.

Generalizability Theory

G-Theory is an extension of Classical Test Theory. Introduced by Lee Cronbach in 1972, G-Theory uses a factorial repeated-measures ANOVA structure. This can be contrasted with the simple ANOVA of CTT (with 'between' and 'within' variance partitions). Meanwhile, G-Theory's factorial design has multiple factors and interactions among factors. For example, a G-Theory design could be persons x tasks x raters, with variance components of: person, task, rater, person x task interaction, person x rater interaction, task x rater interaction, and person x task x rater interaction (this highest order interaction is also termed *residual*).

Within G-Theory, a generalizability-study (G-Study) is an initial G-Theory analysis that provides two notable findings. First, one reliability coefficient is calculated for the entire process in a learning assessment. Second, it calculates variance components for multiple sources of variances (i.e., facets), including students, items, content ease/difficulty, and evaluators. That is, it parses the variation in total scores from an assessment, into the contributions associated with the assessment's multiple facets. Following the G-Study results of reliability and variance components, further decision-studies (D-Studies) can extrapolate the consequences on reliability of changes to facets (e.g., more or fewer items, raters, occasions or skills-stations). D-Studies can help educators to decide, for instance, how many stations are needed for an acceptable reliability, or whether they might obtain lower measurement error with two raters in three stations versus one rater in six stations.

Rios, Li, and Faulkner-Bond (2012) reviewed G-Theory studies in education and found that most used sample sizes <100. A year later, Atiligen (2013) empirically

suggested that a sample size of 50-300 adequately provided unbiased estimations of generalizability indices. This size of sample could easily be obtained at each institution. Thus, G-Theory has potential to improve fairness and rigor of complex academic testing at each institution, by calculating a process-level reliability and parsing total score variation into variance attributable to multiple sources with a G-Study, as well as then facilitate analyses that can optimize reliability through further D-Studies. With better understanding of reliability trade-offs by using D-Studies, G-Theory has the potential to improve the reliability and rigor of local educational testing (Cor & Peeters, 2015).

G-Theory can be applied to an examination. With one examination, the G-coefficient is the same as internal consistency. However, one limitation of Classical Test Theory is that it allows for consideration of only *one* reliability index (internal consistency in this case). But what if we wanted to combine the internal consistencies from multiple examinations into one course-level reliability coefficient? This reliability would characterize the course-level letter grade. With multiple exams, the items would need to be combined, but there are multiple examination occasions. The internal consistency from Classical Test Theory could only be analyzed for each separate examination with no means to combine exams, or an internal consistency could be calculated from all items together while disregarding the fact these came from multiple examination occasions. Instead, after introducing a facet for occasions, G-Theory can combine those multiple examinations into one course-level reliability coefficient.

As noted previously, health-professions education requires performance-based assessments that are different from knowledge-based written examinations (i.e., “shows how” in Miller’s Pyramid). These performance-based examinations have more sources of

variance than a written examination. While a written examination has variance from the test-taker and examination items, a performance-based assessment has variance from test-takers and items (same as a written exam), as well as from raters, stations, and occasions. A model for reliability needs to take these all into account simultaneously. G-Theory can do this. G-Theory can provide one composite reliability coefficient from the multiple sources of measurement error.

Generalizability Theory Fundamentals. That said, the complexity of G-Theory requires a primer of fundamentals. In Table 1 are common terms used within G-Theory.

Table 1

Glossary of terms in Generalizability Theory

Term	Definition
Generalizability Theory (G-Theory)	Analysis of Variance (ANOVA)-based tool to analyze variation over repeated measures from various test sources (e.g., items, occasions, raters, stations)
Facet	A set of similar conditions of assessment; a "variable"; test sources of variation (e.g., students, items, occasions, raters, stations)
D-facet	Facet of determination; object of assessment (usually students). There is only *one* D-facet in a G-Study.
G-facet	Generalization facet; most facets are this type; each has multiple levels within and are assumed to be sampled from an infinite universe (e.g., often raters, stations, items)

(table continues)

Table 1

Glossary of terms in Generalizability Theory (continued)

Term	Definition
S-facet	Stratification facet; categorization facet that is not a G-facet (e.g., year, site, gender, occasion)
Fixed facet	A finite facet that is *not* generalized to a universe of infinite versions of this facet (i.e., held constant); not a facet for generalization (in D-Studies)
Random facet	A facet with many versions; a facet to generalize/extrapolate in D-Studies
Levels	Following Analysis of Variance (ANOVA) language, each variable/facet has multiple configurations (e.g., item scored with 4-levels; 1, 2, 3, or 4)
Crossed design	Two facets are crossed when any levels of one facet can interact with any levels of the other facet (e.g., student is crossed with exam items when all students take all exam items)
Nested design	One facet is nested within another facet when the nested facet has different levels for certain facets—it is not crossed; this happens in all unbalanced designs (e.g., rater is nested in station when different raters are in each station; if *same* rater for all stations, it would be rater crossed with station)
G-Study	Initial analysis of data for variance components from different facets in the specified G-Theory design; provides contribution to measurement error from different facets and interactions of facets.

(table continues)

Table 1

Glossary of terms in Generalizability Theory (continued)

Term	Definition
D-Study	Decision study; extension of G-Study that uses its analyzed variance to yield generalizations of "what if" situations for impact on reliability (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?)
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 facets (e.g., multiple quizzes have different numbers of items, multiple exams have different numbers of items, each OSCE occasion has a different number of stations; different items are 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. There are 13 pre-determined designs and only mGENOVA can analyze a multivariate design.

Sources of Variance. As opposed to variables in traditional studies, G-Theory terms these *facets*. Facets are sources of variance in a test's scores (e.g., students, items,

raters, stations). There are different types of facets. There is a *facet of determination* (D-facet). There is only one D-facet in any G-Study, and this is usually students in many education studies. There are *facets of generalization* (G-facets). Most facets in a G-Theory study are G-facets. G-facets have multiple levels and are assumed to randomly come from a universe of possibilities (i.e., generalize back to that universe). A third type of facet are *facets of stratification* (S-facets). S-facets are categorization facets and are not G-facets. They do not generalize and are non-random "buckets" into which other facets are placed.

Common to a number of health-professions an OSCE exemplifies a learning assessment with multiple facets involved in total score variation. To better express the notion of facets, let us specify 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 score student performance across six areas. Raters score using a set of items that are common to all stations*. Multiple sources of variation within total scores are noticeable. Variance in overall scores will not only come from differences among students' ability but will also emerge from items, raters, stations, and occasions. These are all facets in G-Theory, with students as the D-facet, items, raters, and stations potentially as G-facets, as well as occasions potentially as an S-facet. Different reliability indices in Classical Test Theory can reflect different aspects of this score variance, but these are all happening at the same time. How the raters score the items, for instance, is analogous to internal consistency (i.e., similar to Cronbach's alpha). Variation may also emerge from between raters (i.e. inter-rater consistency), among stations (case specificity), or over the different weeks (a variant of test-retest reliability). The G-Study

uses a multi-way repeated-measures analysis of variance (with variables of station, rater, item, and week), to calculate one process-level reliability that integrates and summarize the plethora of possible internal consistency, inter-rater reliability and test-retest reliability indices (Streiner et al, 2015). Additionally, the G-Study analysis would indicate the amount of variation in total scores that is accounted by each facet or interaction of facets (i.e., variance components).

Crossed and Nested Designs. Aside from determining facets, an investigator must also determine the G-Theory design. In G-Theory, there are a multitude of possibilities for design. By convention, the D-facet (usually students) is termed *p*. In a simple design of students taking an examination, this will be crossed with items, *i*. The term *crossed* means that all students take all items. In another example, the number of examinations could be increased from one to two. The resulting G-Study design will need to allow for these multiple occasions, *o*. This could be done in one of two ways. First, students are crossed with items which are crossed with occasions (short-hand: *p x i x o*). Second, a *nested* design could be used. This design would have items nested in occasions crossed with students (short-hand: *p x i : o*). An advantage of this second nested design is that if there are a different number of items on the two examinations, the items will *not* be weighted equally—the examination occasions would be instead.

Balanced Versus Unbalanced Designs. As in the examination example in the paragraph above, if the two examinations have the same number of items then it is termed a balanced design. However, most designs in education and especially with situational differences for performance-based assessments, designs are unbalanced. That is, in

unbalanced designs there is an uneven number of, in this example, items on the two examinations.

Univariate Versus Multivariate Designs. The vast majority of designs are univariate and involve random facets. Central to a multivariate design is that one content categories are considered fixed. That is, most facets can be generalized to many categories of those facets, however if a facet has a limited/set number of categories, it should be fixed. Once fixed, it cannot be extrapolated in decision-studies to other situations. For example, a univariate design for an OSCE may have random facets for station and week of station. In decision-studies, these facets can be extrapolated to more and fewer stations each week as well as number of weeks. On the other hand, if the week facet is fixed (i.e., there is a set number of weeks and this will not likely change) then the week facet cannot be extrapolated and only the number of stations per week can be extrapolated. However, the variance can be constrained to each week separately from one another—and so each week can be explored separately. This can be seen in the OSCE application,[40] as well as the quizzes application.[42] In both reports, the variance is constrained with each category of the fixed facet, and so each week or quizzes reliability can be defined and reported. This can be seen in Table 2 for OSCE weeks,[40] and Table 1 for the different quizzes.[42] While a stratification facet in a univariate is also fixed and has distinct limited categories, a univariate design cannot constrain other facets to analyze them at each stratification category. That is, Table 2 in the OSCE example[40] followed after multivariate G-Theory analysis, while the different variance components in Table 1 of the Quiz example[42] followed from multivariate G-Theory analysis; neither could be replicated using only a univariate G-Theory design.

Decision-Studies. Following a G-Study analysis, investigators can perform further analyses termed *decision-studies* (or *D-Studies*). Within Classical Test Theory, the Spearman-Brown formulae may be used for a single examination. 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 items would be needed to improve reliability sufficiently. In G-Theory, D-Studies are an extension of this concept. In a D-Study, associated reliability changes due to the adjustment of the facets (such as more or fewer raters, stations and/or items) are estimable.

D-Studies were highlighted by Streiner et al (2015) as, "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" (pg. 212). For example, using D-Studies with an OSCE can help determine the change to reliability associated with altering the N of one or more facets within the learning assessment. Investigators could examine what effect changes to the number of raters, number of items used by raters, and/or number of stations would have on reliability.

Evidence from many empirical studies have demonstrated that increasing the N associated with each of the four facets in the previous example can improve reliability (van der Vleuten & Schuwirth, 2005). However, increasing the N of *some* facets in this OSCE example will contribute more to overall variance than increasing the N of *others*. Educators could examine the trade-offs of different situations, for instance, by examining the effect on reliability if two raters scored four stations, or if the same eight raters

instead were dispersed singly to score eight stations. For expensive, time- and resource-intensive testing, this can help with important decision-making evidence for future test iterations that optimize reliability, rigor and fairness (Crossley, Davies, Humphris, & Jolly, 2002). Through constructing generalization evidence for validation, optimizing a learning assessment's reliability can bolster its validity.

Generalizability Theory Use in Medical Education. During the 1980's, G-Theory saw limited use in medical education (Streiner et al, 2015). With the subsequent further development and increased power of personal computers that facilitated its many onerous calculations, G-Theory found increased utility. As a result, G-Theory has become much more widely used in medical education (Crossley et al, 2007). "Generalisability theory 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." (pg.927). 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 (Swanson & van der Vleuten, 2013), with multiple prominent medical educationalists declaring that use of G-Theory for an OSCE is "absolutely required" (pg.S23; also in Bloch & Norman, 2012; Streiner et al, 2015; Swanson, Clauser & Case, 1999).

As Khan et al (2013) and Boursicot et al (2014) account, the OSCE was introduced in medical education and has become an important performance-based assessment format in medical and other health-professions education (including pharmacy, nursing, and physical therapy). The validity, reliability, feasibility, and

educational impact for OSCEs is noteworthy (Khan et al, 2013; Newble, 2004). 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 from an OSCE at one institution to an OSCE at another institution (Boursicot et al, 2014). Each university has their own unique mixture of resources (number of faculty/support staff, faculty expertise, faculty workload), assessment philosophy and financial commitments.

OSCEs have multiple complex relationships of performance rating items, multiple raters, different scenarios, and other variables to account for. Currently, G-Theory is considered the best means to analyze variance components, characterize reliability, and optimize that reliability for these multi-variable assessments (Streiner et al, 2015; Swanson et al, 1999; Swanson & van der Vleuten, 2013). With each university needing to analyze, evaluate, and validate their own use of learning assessments (Peeters & Martin, 2017), 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 (Peeters & Cor, 2019).

Generalizability Theory Use in Pharmacy Education. Because of similarities with conceptualization of educational assessments among health-professions, G-Theory seems useful to pharmacy education as well. Unfortunately, G-Theory's use in pharmacy education has been very limited and quite recent compared to medical education. Only three published studies using G-Theory in pharmacy education exist (Munoz et al, 2005;

Peeters et al, 2013b; Cor & Peeters, 2015).

In 2005, Munoz and colleagues reported generalizability of a 26-station OSCE for 153 entry-to-practice candidates and 37 pharmacists (Munoz et al, 2005). This report described pharmacist licensing in Ontario, Canada.

A 2013 investigation by Peeters and colleagues is the only study with data from an American institution. However, this investigation was focused on an approach to interviewing candidates for post-graduate pharmacy practice residency positions and was not evaluating students in a PharmD curriculum (Peeters et al, 2013b).

While demonstrating use of G-Theory, the most recent article by Cor & Peeters (2015) used learning assessment data from student pharmacists in Alberta, Canada. Its focus within performance-based assessments was outside of an OSCE format.

The first important step to validation involves generating generalization evidence towards a validation argument. G-Theory appears to be a promising tool for this. Additionally, G-Theory has been used successfully in medical education and appears promising in transfer to pharmacy education. It would seem that pharmacy education could benefit from G-Theory's ability to consolidate multiple sources of variation into a process-level reliability coefficient, parse total score variance into variance from different sources, and then optimize reliability with D-Studies (including economizing raters and task sampling).

Barriers to Generalizability Theory Implementation. Two notable barriers to using G-Theory are its mystique of difficulty and its available specialized software.

Complicated and Abstract. G-Theory can seem abstract and complicated. After

all, the term “theory” is in its name, and “theory” can make many clinicians’ eyes glaze over. Despite instruction through software manuals, a primer on using G-Theory was needed (Bloch & Norman, 2012). Admittedly, navigating within its many steps and multitude of designs requires practice. Even with medical education’s widespread use of G-Theory, it is not understood by all medical education researchers. When Pell, Fuller, Homer & Roberts (2010) discussed test quality and use of G-Theory, their description suggested limited understanding of G-Theory. Additionally, Allalouf, Klapfer & Fronton (2008) appeared to completely miss that previously-reported G-Theory coefficients would parsimoniously include their many internal consistencies and interclass (inter-rater) coefficients. Together, this provides some evidence that G-Theory can be complicated enough. Every educator need not be familiar with its intricacies; however, it is hoped that all educators might gain some appreciation of its potential for helpful use.

Generalizability Theory Software. Software programs available for G-Theory need evaluation on route to its broader use in pharmacy education. Similar to other statistics-based software applications, personal computers have shaped the landscape of G-Theory use. At present, there are a handful of specialized G-Theory software programs, some as stand-alone programs while others are syntax for broader statistics programs. Most are not user-friendly. Table 2 gives a summary of the G-Theory software currently available.

Table 2.

Comparison of Generalizability Theory software

Software	Max Facets	Nested	Fixed	Unbalanced	User-Friendly	Software Format	Data Format	Graphs	D-studies	Item Weights
GENOVA	6	yes	no	no	no	FORTRAN syntax	Short	no	yes	yes
urGENOVA	6	yes	no	yes	no	ANSI C syntax	Short	no	no	no
mGENOVA	6	yes	yes	yes	no	ANSI C syntax	Short	no	yes	yes
SPSS syntax	G1 2, G2 3+	yes	no	G1 no, G2 yes	yes if SPSS-user	SPSS syntax	Short or Long	yes	yes	no
R syntax	R1 2, R2 3+	yes	no	G1 no, G2 yes	yes if R-user	R syntax	Short or Long	yes	yes	no
Gtheory (Rpkg)	2+	no	yes	yes	yes if R-user	R package	Long	no	yes	no
EduG	8	yes	yes	no	yes	Windows	Short (like mGENOVA)	no	yes	no
G_String	infinite	yes	yes	yes	yes	Windows (and Mac)	Short or Long	no	yes	no

While each software program has advantages and disadvantages, no single program stands out for all situations. While using numerous G-Theory program options from Table 2, two notable programs are G_String and mGENOVA. With Windows (and Mac in its newest version), G_String seems most versatile and user-friendly for many unbalanced assessment designs in G-Theory. It still requires prior instruction and plenty of practice to use adequately. Although in situations where some stations might be weighted differently than others, it cannot adjust station-weights. For those situations, mGENOVA would seem best (though not nearly as user-friendly).

Summary

Validation is needed for learning assessments in pharmacy education and is especially needed for high stakes testing (Peeters & Cor, 2019). G-Theory appears to be one promising tool to generate rigorous generalization evidence for a sound validation argument. As opposed to G-Theory's ubiquitous use in medical education, pharmacy education has seen only sporadically rare use of G-Theory in three applications (with the majority outside the United States). From this literature review, G-Theory appears promising to contribute to fair learning assessments in pharmacy education. "Evidence concerning reliability (generalizability) and validity must be obtained so that inferences about educational achievement made from assessment scores are appropriate and

meaningful. Otherwise, the future of performance assessments could be threatened" (Gao et al, 1994, pg.324).

Chapter Three

Methods and Procedures

This chapter reports methods and procedures from three separate sub-studies that make up this dissertation. Recall that this dissertation involved three applications of G-Theory in three independent sub-studies. As reported in Chapter 1, each learning assessment was from a different category of Miller's Pyramid; although each is a different use and needs generalization evidence of its own (within Kane's Framework for Validation). This process at its core is data-driven. The first sub-study demonstrated how G-Theory could calculate, along with optimize, the reliability of a performance-based assessment of skills within an OSCE. In Brennan's (2010) seminal text *Generalizability Theory*, Brennan (2010) asserts that "Generalizability Theory is particularly well suited to evaluating assessments that are based on ratings of human performance." (pg. 117). That is, analyzing a performance-based assessment may most easily demonstrate G-Theory initially, before progressing to other applications. The second sub-study demonstrated how G-Theory could calculate the composite reliability of an entire course's final grades, through combining reliabilities from multiple examinations into one composite reliability for course-grades. This second sub-study is a very common teaching scenario and the subsequent D-Studies are a highlight from it. The third sub-study demonstrated how G-Theory can integrate short quizzes (i.e., brief assessments of students' learning) with a longer exam, resulting in a different example of composite reliability for course-grades. In some coursework in some institutions, use of quizzes is increasing either at the beginning of class meetings during a flipped-lecture or team-based learning that require outside preparation for class, or even to stimulate active-learning during lecture.

Purpose and Research Questions

Purpose Statement. The *purpose* of this proposed study is to demonstrate use of G-Theory for three common applications of learning assessment in pharmacy education, and in so doing, generate key validation evidence for the generalization inference within Kane's Framework for Validation.

Research Questions. This dissertation involved three applications of G-Theory in three independent sub-studies. As reported in Chapter 1, each learning assessment was from a different category of Miller's Pyramid; although each is a different use and needs generalization evidence of its own (within Kane's Framework for Validation). The first sub-study demonstrated how G-Theory could calculate, along with optimize, the reliability of a performance-based assessment of skills within an OSCE. In the 'bible' of G-Theory, Brennan (2010) asserts that "Generalizability Theory is particularly well suited to evaluating assessments that are based on ratings of human performance." (pg. 117). That is, analyzing a performance-based assessment may most easily demonstrate G-Theory initially, before progressing to other applications. The second sub-study demonstrated how G-Theory could calculate the composite reliability of an entire course's final grades, through combining reliabilities from multiple examinations into one composite reliability for course-grades. This second sub-study is a very common teaching scenario and the subsequent D-Studies are a highlight from it. The third sub-study demonstrated how G-Theory can integrate short quizzes (i.e., brief assessments of students' learning) with a longer exam, resulting in a different example of composite reliability for course-grades. In some coursework in some institutions, use of quizzes is

increasing either at the beginning of class meetings during a flipped-lecture or team-based learning that require outside preparation for class, or even to stimulate active-learning during lecture.

Furthermore, each sub-study had three research questions. The structure of the three research questions for each sub-study were: (a) define the reliability, (b) report variance components from the G-Study, and (c) identify patterns from the D-Studies. These research questions followed the phases of a G-Theory analysis (Bloch & Norman, 2012; Streiner et al, 2015), as well as guidance for reporting G-Theory results (Briesch, Chafouleas & Johnson, 2016; Hendrickson & Yin, 2010). Thus, there were nine research questions addressed overall (i.e., three for each application).

The research questions are as follows:

Research Question 1a. What was the reliability of an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo?

Research Question 1b. For an Objective Structured Clinical Examination of 3rd-year Doctor of Pharmacy students at the University of Toledo, what were the relative contributions of occasions and stations to examination score variance?

Research Question 1c. What would be the optimal number of stations for each of three weeks for an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo?

Research Question 2a. What was the composite course-level reliability for multiple examinations in a basic-science course that involved 1st-year Doctor of Pharmacy students at the University of Toledo?

Research Question 2b. What were the relative contributions to variance in scores from examination *occasions* and examination *items* that involved 1st-year Doctor of Pharmacy students in a basic-science course at the University of Toledo?

Research Question 2c. How would the number of examination *occasions* and the number of examination *items* be optimized for a basic-science course that involved 1st-year Doctor of Pharmacy students at the University of Toledo?

Research Question 3a. In a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University, did quizzes add to the reliability of an examination?

Research Question 3b. What were the contributions to scores variance from quiz-items and exam-items in a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University?

Research Question 3c. How could the *number* and *weighting* of items from quizzes be optimized beyond items from the examination towards a module-grade reliability in a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University?

Proposed Research Design and Methodology

The research design for this dissertation is post-hoc educational psychometric analyses for each of the learning assessment sub-studies. In all of its three sub-studies, the investigation will be considered a secondary analysis of archived data—post-hoc analysis after use within these PharmD students' coursework. For these analyses, investigators used G-Theory to quantify reliability, evaluate facet variance components, and explore

different facet scenarios within three common applications for learning assessments in pharmacy education.

Each application's three research questions were based on common aspects of G-Theory analyses (Bloch & Norman, 2012; Crossley et al, 2002; Crossley et al, 2007; Streiner et al, 2015). First, reliability is a notable result of the generalizability-study (G-Study). Reliability can summarize one important aspect of testing rigor and is also considered to be validation evidence for the generalization inference. Second, a notable G-Study contribution is its ability to parse data into its variance components based on the facets specified. This can give investigators an idea of which facets are prominent sources of variance and which may be less helpful to focus on improving. Third, as a practical extension from the G-Study, additional decision-studies (D-Studies) can extrapolate the variance into different practical situations, as the number of levels for a facet are altered. For example, if multiple raters were used in some stations of an objective structured clinical examination (OSCE), what impact might varying the number of raters have on that OSCE's reliability? D-Studies can help with practical decision-making towards a more efficient future assessment with acceptable reliability and rigor.

Three applications of learning assessments that are common in pharmacy education were investigated. The first sub-study illustrated G-Theory use for a high-stakes performance-based assessment (e.g., OSCE). The second sub-study examined integration of multiple examinations from a traditional lecture-based basic-science course. The third sub-study investigated integration of quizzes and an examination during a didactic clinical-science course.

Generalizability Theory. As proposed in Chapter Two, G-Theory can be seen as a flexible model with different extensions based on the testing situation specifics (Bloch & Norman, 2012; Crossley et al, 2002; Crossley et al, 2007; Streiner et al, 2015). The first part of a G-Theory analysis is conducting a G-Study. This G-Study analyzes and apportions the variance due to different test-related facets in the data set. An important initial step towards a G-Study is to define the G-Study's design that best describes the design of a specific learning assessment. Designs can vary among tests and their specifics. After the G-Study provides an overall reliability index (i.e., g-coefficient) for the data, the second part of a G-Study is to understand the magnitude and impact on reliability of different sources of variation (with their potential measurement error). This is an understanding of the variance attributable to the various G-Study facets (i.e., components) and their interactions. This was demonstrated for all three sub-studies, although it may be most helpful with Sub-Study 1 (OSCE).

A third part that can be very helpful and practical corollary to a G-Study is D-Studies. Similar in concept to the Spearman-Brown Prophesy, D-Studies extrapolate for changes in variation of facets, given the variance found within the prior G-Study. Thus, by imputing changes to the number of facets, test developers can explore the influence of altering various learning assessment administration characteristics (e.g., number of items on an examination, the number of examination occasions, or the number of stations in a performance-based examination). Exploration of all of these can allow the investigator to optimize reliability based on an institution's specific needs and available resources. Both Sub-Study 2 and Sub-Study 3 can highlight the helpfulness of D-Studies in assisting with decision-making for future iterations of written learning assessments.

Sample and Sampling Method

Test validation, with its requirement for multiple types of evidence, is fundamental to high-stakes testing in pharmacy education (Peeters & Martin, 2017; Peeters & Cor, 2019). Recall that in Kane's Framework for Validation, evidence is needed for inferences of *scoring*, *generalization*, *extrapolating* and *implications*. If used, G-Theory can help provide evidence for *generalization*. There will be three separate pharmacy education sub-studies within this investigation, and so three separate samples of Doctor of Pharmacy (PharmD) students. Table 3 summarizes the archived data for the three sub-studies.

TABLE 3.

Summary of Secondary Archived Data for Sub-Studies One through Three

	Learning Assessment Type	PharmD Program Year	Semester of Data
Sub-Study One	OSCE	Third	Spring 2017
Sub-Study Two	Course Exams	First	Spring 2017
Sub-Study Three	Course Quizzes	Second	Spring 2018

Sub-Study One (OSCE). From archived data following Spring Semester 2017, this sub-study was a retrospective analysis from data of an OSCE that was administered to third-year PharmD students at the University of Toledo College of Pharmacy & Pharmaceutical Sciences during these students' last semester of classroom-based coursework. This dissertation author was the course-coordinator and so had access to this data. This OSCE was intended as a high-stakes, skills-based assessment of PharmD

students' pharmacy practice skills prior to their advanced year-long clinical rotations. Assessing students' "readiness for clinical rotations" is an important accreditation standard for pharmacy education institutions (Accreditation Council for Pharmacy Education, 2015).

Sub-Study Two (Course Exams). This sub-study was a retrospective analysis of archived data from Spring Semester 2017, for the multiple examinations (i.e., two midterm exams and one final exam) that make up a final course grade within a traditional PharmD didactic basic-science course. Collaborating with pharmaceutics colleagues in the University of Toledo College of Pharmacy & Pharmaceutical Sciences, this analysis was conducted with matched data from a first-year PharmD course that is focused on the basic-science discipline of pharmaceutics.

Sub-Study Three (Course Quizzes). This sub-study was a retrospective analysis of archived data, from Spring Semester 2018, from a clinical-science (cardiovascular pharmacotherapy) course of second-year PharmD students. This data was acquired through collaboration with an instructor from a different institution (Drake University College of Pharmacy). The course instructor employed multiple low-stakes quizzes integrated with a higher-stakes exam into course grades. Similar in format to Sub-Study 2, students' responses were matched on multiple quizzes and an examination, with subsequent G-Theory analysis.

Instrumentation and Procedures

Three applications were analyzed using G-Theory. Each used essentially the same G-Theory instrumentation, though with different designs (and spreadsheet columns) for specifics of each sub-study. These sub-studies were chosen is to demonstrate using G-

Theory for a variety of common learning assessment applications among multiple years of PharmD students. As noted in the prior chapter, there are a number of G-Theory programs available (Table 2).

Sub-Study One (OSCE). After IRB approval, retrospective archived data were obtained from the course coordinator. These archived data were from a third-year PharmD OSCE that was administered over 3 weeks during Spring Semester 2017.

Over three weeks, students completed 14 stations, with five in week1, five in week2, and four in week3. In this OSCE iteration, the 14 stations that all students rotated through and completed 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. In this OSCE, all 14 stations were equally weighted in students' overall score.. These data were analyzed using a $p^{\bullet} \times s^{\circ}$ design. Sub-study One had a fixed three-week format (that will not vary), with four or five stations each week. The mGENOVA software was used for this fixed design.

Third-year PharmD students were assessed across various pharmacy practice skills. These skills were planned, created, and developed by a team of five practicing faculty pharmacists. These skills-based OSCE stations were focused into 4-5 stations per week. Roughly, the circuit format was 9-minutes per station with 1-minute in-between (though some stations took twice as long and so were de facto "double stations"). One or two stations each week used faculty graders, while the other stations (three per week) were written and scored afterwards. All stations were scored independently over the three

weeks of this OSCE. 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). 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 (Newble, 2004). Students could fail one station on the initial OSCE and pass the entire assessment. However, if two or more stations were failed, that student needed to remediate all stations that they did not pass.

Sub-Study Two (Course Exams). These archived data were from a basic-science (pharmaceutics) course for first-year PharmD students during Spring Semester 2017. After IRB approval, these retrospective archived data were obtained from the course coordinator.

In this 15-week course, there were 12-weeks of lecture and three examinations—two midterms and a final. The examination data were item-level data (1=right, 0=wrong) for each student, with students matched on their exam one, exam two, and exam three responses, prior to G-Theory analysis. With different numbers of questions on each exam in Sub-study Two, this G-Theory design is unbalanced but not fixed to a specific number of examinations. These data were analyzed with an unbalanced design of $p \times i : o$. For this Sub-Study Two, *G_String* software was used.

Sub-Study Three (Course Quizzes). After IRB approval by both institutions' IRBs, retrospective archived data were obtained from the course instructor. Collected during Spring Semester 2018, these archived data were from a second-year PharmD course (at Drake University College of Pharmacy) that used frequent quizzes prior to an examination.

While this therapeutics course was 15-weeks with three modules, only one module was be evaluated in this sub-study. The module was five weeks and had seven quizzes ranging from five to nine items, along with a 32-item exam. For the quiz and examination data, these data were item-level data (1=right, 0=wrong) for each student. Before any G-Theory analysis could be undertaken, each student's seven quizzes and examination were matched so that all each student's tests are a single row of data. This sub-study had a fixed number of quizzes and an examination, a different number of questions on each quiz (i.e., unbalanced), and different item-weightings for quizzes versus the exam (in the syllabus). Thus, these data were analyzed with an unbalanced, fixed design of $p^{\bullet} \times t^{\circ}$. Because it can account for fixed, unbalanced and different weightings, *mGENOVA* software was used.

Data Analysis

Using the G-Theory software (mGENOVA for Sub-Study 1 & Sub-Study 3, or G_String for Sub-Study 2), the various sub-study data were analyzed. From the subsequent data analytic reports, information was pulled out to answer the research questions. Below, description of report information was grouped by type of research question (a, b, c).

Data Analyses Part A. For each sub-study, the first question had similar information from data analytic reports. Herein, the reliability (via g-coefficient) for each G-Theory design was reported. This g-coefficient provided a reliability co-efficient for that learning assessment iteration.

Research Question 1a. What was the reliability of an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo?

Research Question 2a. What was the composite course-level reliability for multiple examinations in a 1st-year Doctor of Pharmacy students in a didactic basic-science course at the University of Toledo?

Research Question 3a. In a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University, did quizzes add to the reliability of an examination?

Data Analyses Part B. For each sub-study, the second question of each sub-study sought similar information from the data analysis reports. This information was elsewhere in the same data analytic report. That is, the G-Study has both a g-coefficient (research questions 1a, 2a, and 3a) and provided the relative contribution to variance in scores from different facets (research questions 1b, 2b, and 3b).

Research Question 1b. For an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo, what were the relative contributions of *occasions* and *stations* to examination score variance?

Research Question 2b. What were the relative contributions to variance in scores from examination *occasions* and examination *items* that involved 1st-year Doctor of Pharmacy students in a basic-science course at the University of Toledo?

Research Question 3b. What were the contributions to variance of scores for *quiz-items* and *exam-items* in a clinical-science module that involved 2nd-year Doctor of Pharmacy students at Drake University?

Data Analysis Part C. For each sub-study, the third research question of each sub-study needed similar information found elsewhere from the other research questions, within the same data analytic reports. While the information to answer the first two questions with each sub-study were from the G-Study, this information was from the D-Studies in the same report. Following after and from the variance components obtained from the G-Study, these D-Studies explored the impact of different choices to the number of facets in that specific learning assessment.

Research Question 1c. What would be the optimal number of *stations* for each of three weeks for an Objective Structured Clinical Examination that involved 3rd-year Doctor of Pharmacy students at the University of Toledo?

Research Question 2c. How would the number of examination *occasions* and the number of examination *items* for a basic-science course that involved 1st-year Doctor of Pharmacy students at the University of Toledo?

Research Question 3c. How could the *number* and *weighting* of items from quizzes be optimized beyond items from the examination towards a module-grade reliability in a clinical-science course that involved 2nd-year Doctor of Pharmacy students at Drake University?

Limitations, Delimitations and Constraints

Both limitations and delimitations of this dissertation need discussion.

Limitations. As with any data analysis from a single institution, generalization to other institutions was one notable limitation. The last fifty years of social science research have shown that context matters. That is, each high-stakes learning assessment within the specific institutional context at each college/school of pharmacy should be

analyzed itself (Peeters & Cor, 2019). Educational assessments are context-dependent and will vary in reliability (both in each cohort of learners, and at different institutions). Reliability deals most with generalization to future cohorts of students at the *same* institution. Although it can differ in subsequent cohorts, it should often remain quite similar (if reasonably stringent) unless other future changes make future cohorts of students different than the current students (e.g., using different teaching and learning methods, or a different curriculum). Further, evidence for validation should be specific to the use and interpretation of test scores at each college/school of pharmacy. For example, with contexts (e.g., curricular structure, admissions, faculty, educational environment) being similar, a test of first-year PharmD students in 2014 at a College of Pharmacy should generalize to the 2017 version of that same test also among first-year PharmD students at that same College of Pharmacy. Therefore, very specific findings from sub-studies in this dissertation are most relevant to the University of Toledo College of Pharmacy (or Drake University College of Pharmacy for the third sub-study). G-Theory, as an extension to Classical Test Theory, is sample-dependent; its results are characteristic of the participants that made up the sample in that G-Study. However, when accumulated with evidence from other institutions, generalizable concepts can surface. Beyond demonstrating these uses of G-Theory in pharmacy education, common key concepts will also be illustrated (along with discussion including corroborating evidence from the wider literature).

Additionally, optimizing test quality necessitates a retrospective review of testing questions, tasks, and conditions (Pell et al, 2010; Tokovol & Dennick, 2012). G-Theory

can help provide generalization evidence from each learning assessment, but this generalization is foremost targeted to future classes/cohorts at the same institution.

The inferences from Kane's Framework for Validation of learning assessments are *scoring*, *generalization*, *extrapolation*, and *implications* (Peeters & Martin, 2017). While G-Theory can provide evidence for optimizing *scoring* and (especially) *generalization*, this evidence is not *extrapolation* nor *implications* evidence. That is, it is not evidence of relationships to other variables outside of this study, nor is it evidence of consequences from this testing. Validation is not a single study, but a program of research (Cook et al, 2015; Kane, 2006; Peeters & Martin, 2017); this dissertation provides initial generalization evidence, but further evidence for extrapolation and implications inferences would also be helpful.

Furthermore, the results from G-Theory are sample-dependent. If, at a single institution, administration factors (such as timing, prepping students, training raters, or even revising content of items or tasks) change, the G-Study variance components may change. Thus, specific numbers within D-Studies can also change. Because G-Theory is sample-dependent, it is best used in more than one cohort of similar students, for validation of each learning assessment at every institution.

Delimitations. While many learning assessment principles apply more generally, this dissertation's intended audience is pharmacy education. This dissertation is not intended to be comprehensive of all situations or every type of learning assessment in pharmacy education, but it is hoped to provide common useful examples. As pharmacy educators seek to provide validation evidence for their uses of learning assessments, demonstration of G-Theory within these sub-studies may prove helpful.

Two major assumptions within this dissertation are the importance and need for rigorous psychometric validation evidence, and that reliability is a key aspect of validity. If either assumption is invalid, the helpfulness of this dissertation is more limited. However, the psychometric premise for learning assessments has met legal standards and been upheld by courts in the American legal system (Sireci & Parker, 2006). Additionally, with court cases in the health-professions, reliability of learning assessments has been instrumental, and poor reliability of learning assessments has been a key vulnerability (Tweed & Miola, 1991). Reliability is an increasingly important concern as the stakes of an assessment increase, with low-stakes testing not nearly as important as uses in higher-stakes testing (Peeters & Cor, 2019). When reliability can be analyzed and verified, it should not be assumed.

A notable delimitation is that G-Theory is an extension of Classical Test Theory and uses its method of summing right answers into a continuous total score with similarly-weighted item. Item-Response Theory demonstrates this is not entirely correct. In very high-stakes testing, this difference can make a difference in test efficiency (i.e., how many items are needed to precisely measure a student's ability). However, notable reasons for using a Classical Test Theory approach for validation are the complexity of IRT analyses, along with prior use and understanding of Classical Test Theory by the vast majority of educators and students (De Champlain, 2010; Hambleton & Jones, 1993). For sake of validation evidence for generalization, Classical Test Theory seems preferred with the possible exception that an Item-Response Theory-based approach would continue to be used within future iterations of a learning assessment that

specifically using Item-Response Theory. Thus, G-Theory is very similar and extends a Classical Test Theory approach.

Constraints. As with each program of higher-learning, there were notable constraints for each course within the three sub-studies. In the first sub-study of an OSCE, the number of weeks was specified to be three (and so not two or four). For the foreseeable future, this OSCE would be set at three weeks and so the optimal number of stations each week would be more helpful than theoretically expanding the number of weeks. Within each week, the time length of stations, as well as the time between stations could vary to support a greater or lesser number of stations each week. During this initial OSCE, these were set as 9-minutes per station and 1-minute in-between stations. Within the second sub-study of combining examinations, the number of items (questions) on each examination would need to fit into the classroom time allowed. For instance, a 100-minute class should be fine for an exam of <100-items, but 100-minutes may not be enough for a 150-item examination. In the third sub-study of quizzes, a constraint on the quizzes was to increase the number of quizzes. There was already a quiz within most sessions, and so unless there became more lecture blocks, there could not feasibly be more quizzes. However, it could be feasible to manipulate the number of items on each quiz, as well as the relative-weighting given to quizzes and exams could be varied and so investigated.

Chapter Four

Results

This chapter reports findings from three separate sub-studies that make up this dissertation. Recall that this dissertation involved three applications of G-Theory in three independent sub-studies. As reported in Chapter 1, each learning assessment was from a different category of Miller's Pyramid; although each is a different use and needs generalization evidence of its own (within Kane's Framework for Validation). This process at its core is data-driven. The first sub-study demonstrated how G-Theory could calculate, along with optimize, the reliability of a performance-based assessment of skills within an OSCE. In the seminal, authoritative text *Generalizability Theory*, Brennan (2010) asserts that "Generalizability Theory is particularly well suited to evaluating assessments that are based on ratings of human performance." (pg. 117). That is, analyzing a performance-based assessment may most easily demonstrate G-Theory initially, before progressing to other applications. The second sub-study demonstrated how G-Theory could calculate the composite reliability of an entire course's final grades, through combining reliabilities from multiple examinations into one composite reliability for course-grades. This second sub-study is a very common teaching scenario and the subsequent D-Studies are a highlight from it. The third sub-study demonstrated how G-Theory can integrate short quizzes (i.e., brief assessments of students' learning) with a longer exam, resulting in a different example of composite reliability for course-grades. In some coursework in some institutions, use of quizzes is increasing either at the beginning of class meetings during a flipped-lecture or team-based learning that require outside preparation for class, or even to stimulate active-learning during lecture.

Sub-Study One (OSCE)

Within a third-year pharmacy practice skills lab-course, 97 PharmD students completed this OSCE. Of 1358 stations attempted by students, 1259 stations (92%) were passed on first attempt.

Research Question 1a. *What is the reliability of an Objective Structured Clinical Examination of 3rd-year Doctor of Pharmacy students at the University of Toledo?*

Student's scores were modeled as stations nested within three fixed weekly occasions ($p^{\bullet} \times s^{\circ}$). Reliability (G-coefficient) for the entire three weeks was 0.74.

Research Question 1b. *For an Objective Structured Clinical Examination of 3rd-year Doctor of Pharmacy students at the University of Toledo, what are the relative contributions of occasions and stations to examination score variance?*

The variance components among sources that were identified in this design are in Table 4. Of the observed total-score variance, students accounted for only 5%-28% in each of the weeks, while stations were 11%-16%, and student-station interaction (which includes undefined variation and residual) was very large at 56%-81%. [Note: Within G-Theory, the highest order interaction also will contain the residual error. (Brennan, 2010)]

Table 4

Variation sources (and percentage) from a G-Study of third-year PharmD students over three weeks of testing within an objective structured clinical examination of pharmacy practice

G-Study Facet	Week1	Week2	Week3
student	0.15 (17%)	0.06 (5%)	0.2 (28%)
station	0.1 (11%)	0.16 (14%)	0.12 (16%)
student x station (and residual error)	0.64 (71%)	0.94 (81%)	0.4 (56%)

Research Question 1c. *What would be the optimal number of stations for each of three weeks of an Objective Structured Clinical Examination of 3rd-year Doctor of Pharmacy students at the University of Toledo?*

With this variance, D-Studies provided g-coefficient estimates for stations each week and total-scores are in Table 4. Across the columns, reliability increased with more stations each week. As well, the three weeks combine into the Total Score row, where reliability shows improvement with multiple weeks of testing. Using the G-Study variation, Table 5 shows that only an OSCE with seven stations in each of three weeks (i.e., 21 stations) would meet a threshold reliability greater than 80%. The third week had stronger reliability estimates than either the first or second weeks. Alternatively, the second week showed lowest reliability; improving stations especially in week 2 should also help improve reliability overall and use fewer stations. For some reason(s), the stations that week did not statistically discriminate similar to other weeks. To look more

closely at week2, the station topics, circumstances surrounding administering those OSCE stations, as well as other curricular activities (e.g., exams in other classes) should be focuses.

Table 5

Reliability (G- coefficients) for the number of stations each week of an objective structured clinical examination of pharmacy practice for third-year PharmD students

	Stations per 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***

**** Greater than threshold of 0.80***

Meanwhile, Figure 7 shows reliability improvement as a function of increasing the number of stations each week. To identify "optimal" will depend on educational context. With the resources and constraints at the University of Toledo (UT), optimal would be to increase the number of stations as much as possible. UT's PharmD program has notable constraints on the scheduling (e.g., a 3-hour lab for about 50 students once each week along with a lecture for all 100 students), faculty (about seven faculty each session), and facilities (a practice lab with three classrooms and seven small rooms for private counseling). That said, Table 4 noted misalignment and poorer reliability of the week 2 stations. If these can be improved, as well as technical issues during implementation of activities that week, reliability should be vastly improved. As a

consequence of this analysis, UT faculty increased the number of stations as much as possible *and* revisited content mapping within an attempt to better align week2 OSCE stations.

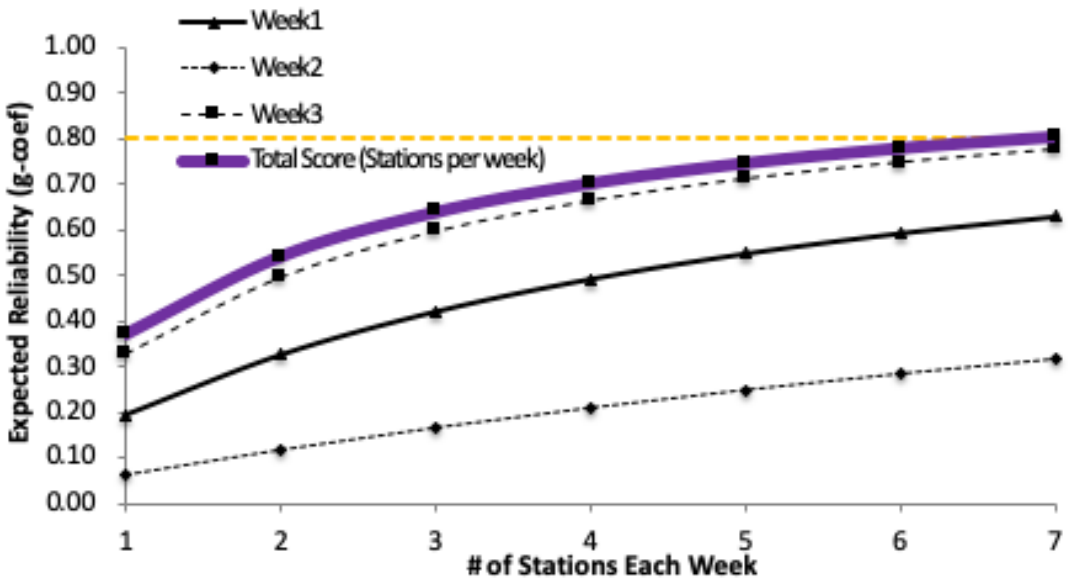

Figure 7. Estimated reliabilities (via G-coefficients) for number of stations within three weeks of testing of third-year PharmD students with an objective structured clinical examination of pharmacy practice

Sub-Study Two (Course Exams)

Within a pharmaceutics course, 101 first-year PharmD students took two midterms and one final-exam.

Research Question 2a. *What is the composite course-level reliability for multiple examinations in a 1st-year Doctor of Pharmacy students in a didactic basic-science course at the University of Toledo?*

In this basic-science PharmD course, 101 first-year students took two midterms and 1 final exam. The first midterm exam was 50 multiple-choice questions (MCQs), had a mean of 40.8/50 (81%; standard deviation=4.4), and KR20 of 0.69 for these students.

The second midterm examination was 43 MCQs, had an average of 36.4 (standard deviation=3.5), and KR20 of 0.65 for these students. The final examination was 67 MCQs, had an average of 55.3 (standard deviation=4.3), and KR20 of 0.67 for these students. Using G-Theory to model items nested within occasions/exams (p x i : o), the composite reliability (G-coefficient) of course-grade, with all examinations combined, was 0.71.

Research Question 2b. *What are the relative contributions to variance in scores from examination occasions and examination items for 1st-year Doctor of Pharmacy students in a didactic basic-science course at the University of Toledo?*

Results of variance components are in Table 6. There did not appear to be variation from occasions alone (i.e., one examination was more difficult for everyone) and little variance (1%) came from some students finding one entire examination more difficult than the others. However, 21% of variation came from variation in some items being easier or more difficult, over the multiple occasions. However, the vast majority of variation (76%) came from the interaction of students with items nested in the examination occasions; that is, some students found certain items easier or more difficult. It is noteworthy that only 2% of variation was attributed to absolute differences in students' knowledge.

Table 6

Variation sources from G-Study of first-year PharmD students in a basic-science course

	Examination	
G-Study Facet	**Variance**	**Percent**
Student	0.003	2
Occasion	0.000	0
Item : Occasion	0.034	21
Student x Occasion	0.001	1
Student x Item : Occasion (and residual error)	0.113	76

Because the exams being combined, two potentially "easier" midterms (i.e., tested over content in past few weeks and not cumulative for the entire semester like the final exam) along with the comprehensive final exam (equally-weighted in final grade), another G-Theory analysis was done of the midterms only. The variance components for these midterms was very similar to the entire (s=2%, o=0%, i:o=19%, sXo=1%, sXi:o=78%), suggesting that these were not different within this G-Theory perspective.

Research Question 2c. *How does the number of examination occasions impact reliability compared to increasing the number of examination items for 1st-year Doctor of Pharmacy students in a didactic basic-science course at the University of Toledo?*

How reliability would change as a function of number of occasions and number of items per occasion is shown in Table 7.

Table 7

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***

*** Meets "acceptable" threshold of 0.80**

As the number of exams increased, fewer questions were needed per exam, and fewer questions over all exams combined. For a course-level grade reliability (g-coefficient) of 0.80, options in Table 6 include four exams of 80MCQs (320 items), five exams of 60MCQs (300 items), or six exams of 40MCQs (240 items). Overall, more occasions lent to fewer numbers of items overall. The G-coefficients from Table 7 are illustrated in Figure 8. A declining impact from occasions can be compared to the decline with increasing numbers of items.

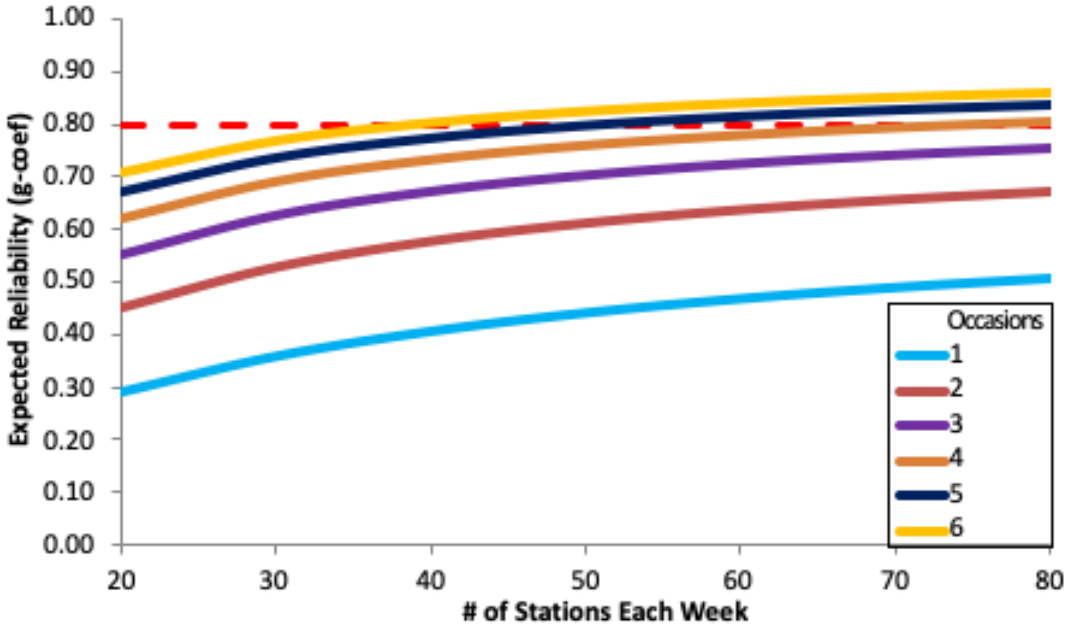

Figure 8. Estimated reliabilities (via G-coefficients) for one through six testing occasions in a first-year PharmD basic-science course

An "optimal" mixture of examination items and occasions will exceed acceptable reliability of >0.8, although will vary by educational setting. Trade-offs can be weighed by an instructor—that is, do they have longer periods of class-time (for which a longer exam could be realized), or can an instructor carve out more smaller blocks for more shorter examinations. As noted above, many shorter examinations can be "more efficient" and save in the number of items that need to be created. Noting the difficulty with writing good items, this may be the more prudent scenario for most instructors.

Sub-Study Three (Course Quizzes)

One-hundred students took seven quizzes and one exam during a clinical pharmacy course.

Research Question 3a. *In a didactic course of 2nd-year Doctor of Pharmacy students at Drake University, do quizzes add to the reliability of an exam?*

In a clinical-science PharmD course, 100 second-year PharmD students completed seven quizzes and one exam over the entire module. Quizzes had a total of 50MCQs (with 5-9MCQs on each quiz), with most KR20's less than or equal to 0.54. These KR-20 reliability coefficients are in Table 8. The exam had 32MCQs (KR20 reliability=0.67). The G-study model used had items nested within eight fixed assessments ($p^{\bullet} \times i^{\circ}$). Using quiz & exam weightings of 18% & 82% as in the course syllabus, the reliability of the composite grade reliability was 0.73; this was improved over the exam alone.

Table 8

Reliability (via KR20) for seven quizzes and an exam during a second-year PharmD clinical-science module

	Quiz 1	**Quiz 2**	**Quiz 3**	**Quiz 4**	**Quiz 5**	**Quiz 6**	**Quiz 7**	**Exam**
Number of Items	9	9	9	5	5	5	8	32
Mean (SD)	7.3 (1.2)	7.2 (1.1)	7.0 (1.6)	4.1 (1.0)	4.0 (1.1)	4.2 (1.0)	6.5 (1.3)	26.6 (3.4)
Reliability (KR20)	0.26	0.00	0.54	0.30	0.34	0.30	0.51	0.67

Research Question 3b. *What are the contributions to variance of scores for quiz-items and exam-items in a didactic course of 2nd-year Doctor of Pharmacy students at Drake University?*

The variance components among sources that were identified in this design are in Table 9. Of the observed total-score variance, students accounted for 0%-62% within

each of the assessments, while items were 1%-21%, and student-item interaction (which includes undefined variation and residual) was very large at 31%-79%.

Table 9

Raw (and percentage) variance components from a G-study of seven quizzes and exam during a second-year PharmD clinical-science module

G-study Facet	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Quiz 6	Quiz 7	Exam
Student	0.004 (3%)	0.0 (0%)	0.017 (9%)	0.011 (7%)	0.014 (9%)	0.011 (8%)	0.014 (9%)	0.007 (5%)
Item	0.038 (20%)	0.035 (21%)	0.032 (18%)	0.013 (9%)	0.013 (8%)	0.002 (2%)	0.029 (19%)	0.025 (18%)
Student x item (and residual)	0.112 (73%)	0.128 (79%)	0.128 (73%)	0.129 (84%)	0.136 (84%)	0.124 (90%)	0.110 (72%)	0.111 (78%)

The quizzes were preparation for the examination. Similar to the second G-Theory analysis for Research Question 2b, variance components for the quizzes were conducted. These results were also very similar as noted in Table 10. Once again, this demonstrated consistency in the quizzes being added to the exam.

Table 10

Raw (and percentage) variance components from a G-study of seven quizzes (excluding exam) during a second-year PharmD clinical-science module

G-study Facet	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Quiz 6	Quiz 7
Student	0.004 (3%)	0.0 (0%)	0.017 (9%)	0.011 (7%)	0.014 (9%)	0.011 (8%)	0.014 (9%)
Item	0.038 (25%)	0.035 (21%)	0.032 (18%)	0.013 (9%)	0.013 (8%)	0.002 (2%)	0.029 (19%)
Student x item (and residual)	0.112 (73%)	0.014 (79%)	0.128 (73%)	0.129 (84%)	0.136 (84%)	0.124 (90%)	0.110 (72%)

Research Question 3c. *How does the number and weighting of items from quizzes impact reliability beyond items from the exam in a didactic course of 2nd-year Doctor of Pharmacy students at Drake University?*

In this cohort of students, reliability increased with larger numbers of items. Additionally, a greater number of quizzes increased reliability. These trends can be seen in the D-Studies of Table 11.

Table 11

Estimated reliabilities (via G- coefficients) for number of items on each quiz for a second-year PharmD clinical-science module

	Number of Items on Each Quiz									
	1	2	3	4	5	6	7	8	9	10
Quiz1	0.04	0.07	0.10	0.13	0.16	0.19	0.21	0.24	0.26	0.28
Quiz2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Quiz3	0.13	0.23	0.31	0.37	0.43	0.47	0.51	0.54	0.57	0.60
Quiz4	0.15	0.26	0.35	0.42	0.47	0.52	0.56	0.59	0.62	0.64
Quiz5	0.18	0.31	0.40	0.47	0.53	0.57	0.61	0.64	0.67	0.69
Quiz6	0.15	0.25	0.34	0.41	0.46	0.51	0.54	0.58	0.61	0.63
Quiz7	0.13	0.23	0.30	0.37	0.42	0.47	0.50	0.54	0.57	0.59
Total Quiz Score	0.26	0.41	0.51	0.58	0.63	0.68	0.71	0.73	0.76	0.78

As noted in Table 12, using quiz & exam weightings of 18% & 82% as defined in the course syllabus, reliability of the entire module was 0.73. Doubling the quiz-weight to 36%, increased the reliability of module grade to 0.77. Reliability of 0.80 was achieved with equal-item-weights for quizzes and the exam (i.e., equal item-weight for 50 quiz-items and 32 exam-items = 82 assessment-items).

Table 12

Estimated reliabilities (via G-coefficients) for composite course-grades as a function of different quiz-item weights, for a second-year PharmD clinical-science module

Weighting of Quizzes	Estimated Reliability
18%	0.73
36%	0.77
61%	0.80

This is also illustrated in Figure 9.

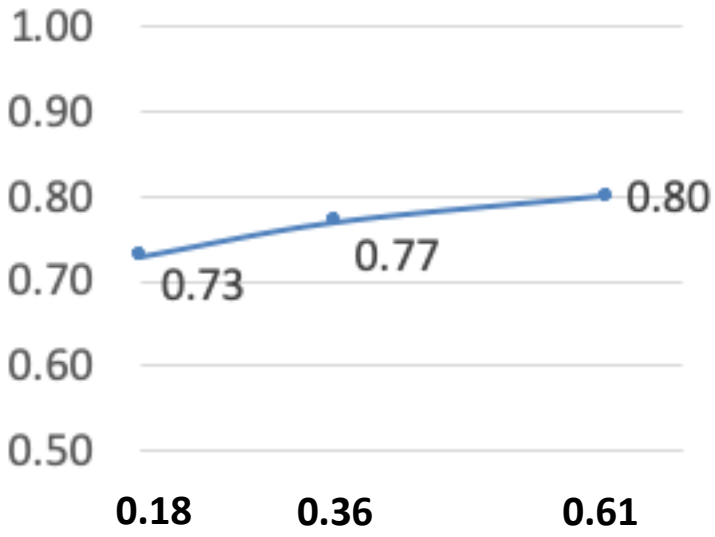

Figure 9. Expected reliability as a function of item-weight given to quizzes, in a second-year PharmD clinical-science module

Once again, the “optimal” weighting of quizzes will vary by institution. While an equal weighting of quiz and exam items provided the best reliability, many instructors may shy away from weighting quizzes and exams similarly. However, looking closer at Table 8, Quiz 2 was had a KR-20 of 0.0 and so was misaligned with the other quizzes and examination. Closer investigation of this quiz as well as the circumstances surround this quiz’s administration will be reviewed further. With that improvement, the doubling of quiz weight may suffice. This is worth further analysis after next iteration.

Chapter Five

Conclusions, Implications, and Recommendations

Summary of the Study

This dissertation demonstrates G-Theory for three separate sub-studies. As reported in Chapter One, each learning assessment was from a different category of Miller's Pyramid; each was a different use and so needs generalization evidence of its own (within Kane's Framework for Validation). At its core, this validation process is data-driven.

The first sub-study demonstrated how G-Theory could calculate, along with optimize, the reliability of a performance-based assessment of skills within an OSCE. In his seminal book *Generalizability Theory*, Brennan (2010) asserted that "Generalizability Theory is particularly well suited to evaluating assessments that are based on ratings of human performance" (pg. 117). That is, analyzing a performance-based assessment may most easily demonstrate G-Theory initially, before progressing to other applications.

The second sub-study demonstrated how G-Theory could calculate the composite reliability of an entire course's final grades, through combining reliabilities from multiple exams into one composite reliability for course-grades. This second sub-study is a very common teaching scenario and the subsequent D-Studies were a highlight from it. How many exams of how many items should yield a course-grade that is acceptably-reliable (α=0.8)?

The third sub-study demonstrated how G-Theory could integrate short quizzes (i.e., brief assessments of students' learning) with a longer exam, resulting in a different example of composite reliability for course-grades. The use of quizzes has increased In

some coursework in some institutions, with a quiz at the beginning of class meetings over required outside preparation material (e.g., for flipped-classroom or team-based learning), or even to stimulate active-learning during a more conventional lecture. While test-scores from many exams are not often sufficiently reliable ($\alpha \geq 0.8$) on their own, could quizzes augment the reliability of the course's final-grade?

Summary of Findings. This investigation had three sub-studies. Each was a demonstration of G-Theory in pharmacy education for a different application. First, a performance-based OSCE that assessed PharmD students' pharmacy-practice skills was analyzed. This sub-study characterized a reliability for this complex learning assessment, with more concurrent sources of measurement error than with one written exam. As opposed to internal consistency with a standard written exam with score variance coming from only *students* and exam *items*, score variation for this OSCE also came from a further source, *stations* (which accounted for 11-16% of variation in these OSCE scores). Thus, this sub-study demonstrated that increasing the number of stations in this OSCE could helpfully improve the reliability for this high-stakes learning assessment.

In the second sub-study, multiple exams within a pharmaceutics course were combined into an improved course-level reliability for final course/letter grades. Thus, this second sub-study demonstrated that multiple similar exams could be added towards a reliability for the course-grade. As sources of variation, *students*, number of *items* per exam, number of exam *occasions*, as well as the interactions of these sources impacted variation of course-grades. For an acceptable course-level reliability, trade-offs could be better understood, as varying the number of items and number of exam occasions could be planned for a course instructor's unique needs and scheduling (i.e., fewer longer

exams, more frequent shorter exams, or somewhere in-between). Playing the trade-offs, it appeared that a common situation of administering multiple shorter exams can be rigorous and sound overall despite the limited reliability from each occasion on its own. Guidance for the number of exams and items per exam were suggested from D-studies.

In the third sub-study, multiple quizzes were combined with an exam to show how reliability of the course-grade could be improved by integrating these multiple learning assessments. Along with exam items, quiz items also explained substantial variation in final course-grades. That is, using numerous short quizzes throughout a module could improve the rigor in its assessment of learning, with a notable caveat. It was also shown that if this potential benefit of quizzes was to be exploited to improve reliability of the course grade, quizzes would need a substantial weight in comparison to the weight of exams. Thus, increasing the percentage weight given to the quizzes could improve the reliability of this course's final grades. Once again, D-studies provided guidance for the number of quiz items, as well as suggesting the influence by manipulating the weighting of quizzes versus the exam (in the syllabus-defined calculation of final course-grades).

Conclusions

Context Matters. Each institution will have different needs, lending to revising learning assessments to those needs. Each institution will have a different mixture of resources, whether scheduling, more limited faculty/personnel, a greater ratio of staff to faculty, more limited financial resources, or sharing resources (e.g., physical facilities) with another health-profession at that institution. Noting these contextual factors (needs, resources) should underscore that characterizing the reliability of these unique learning

assessments at every institution, as well as exploring future modifications to a learning assessment. Validation is vital for high-stakes testing and should be a key aspect for every institution involved in pharmacy education (Peeters & Cor, 2019). Validation is local, data-driven, and evidence based.

The various learning assessments and combinations of these learning assessments should be constructed from sound understanding of testing theory, as opposed to historical precedent, which is pejoratively termed "rear-end validity" (Smith, 2006). Only because a learning assessment has been used for a certain purpose for the last 20 years, does not mean that purpose is valid. It may never have been valid, and time alone is not validity evidence for use of scores from a learning assessment. Instead, an argument for use of scores from a learning assessment for a specific purpose should be empirically confirmed through validation. While there are four levels of inferential evidence to support validation in Kane's Framework for Validation, initial evidence for generalization is vital. A crucial evidence for generalization is reliability.

Steps to Generalizability Theory. Central to reliability of learning assessments is quantifying, understanding, and ultimately best-controlling measurement errors to best support decision-making for scores from a learning assessment. The first step is to quantify reliability. Demonstration with G-Theory supports a framework for reliability, especially notable with complex performance-based assessments such as an OSCE with its multiple sources of score variation. With these variation sources identified, next is to quantify the relative contributions from each (i.e., variance components). Most powerful is what follows this (Streiner, Norman, & Cairney, 2015). D-studies from G-Theory allow extrapolation to "what if" scenarios. Through playing the trade-offs, local test-

developers can better explore and understand what alterations in the amount of different variation sources can have on a learning assessment's reliability. This with data from one specific learning assessment at one institution and so helping to provide the most efficient use of their resources.

For an OSCE, this could mean analysis for altering the number of stations, using multiple occasions to extend the number of stations, or altering the number of raters in each station. For a set of written exams in a course, this could mean analysis for altering the number of items on each exam and/or the number of exams in a course. For a didactic course using quizzes, this could mean analyzing trade-offs between the number of quiz items as well as how different weights of quizzes versus an exam towards the overall course-grade could affect the reliability of overall grades in that course.

Context Specificity Matters. Furthermore, some questions (or performance-based tasks) are more difficult than others for some but not all students. That is, different students' performances on different items can vary; some students may find certain types of question (or performance-based tasks) easier while other students may find certain types of questions (or performance-based tasks) more difficult. Within G-Theory, *context specificity* is the interaction of students with items or tasks (van der Vleuten, 2014). Many times (some experts would say "predictably so") there is profound variance in the interaction of student with items (or stations/tasks in a performance-based assessment), often with more variance than the student or item (or station) facets alone (van der Vleuten, 2014).

This notable issue was found in each sub-study of this dissertation. In the first sub-study of an OSCE, context specificity accounted for a whopping 56-81%. In the

second sub-study, students alone showed only a 2% variance overall; however, their variance from context specificity of items (i.e., context specificity as often used elsewhere) was 76%. Moreover, the facet of exam occasions did not show any contribution to score variance (which is a positive finding itself, suggesting that these exams were similarly difficult for everyone): although, 1% of variance came from differences in context specificity of these entire exams (i.e., some students found Exam1 more difficult while other students were just the opposite and found Exam2 more difficult). This 1% contribution illustrates, in this sample, the contribution of a "bad day for taking a test" along with some students finding one exam easier while other students perform better on a different exam. In the third sub-study, the contribution of context specificity to learning assessments, similar to the other sub-studies, was a substantial 31-79%. Thus, the contribution of context specificity cannot nor should not be overstated.

That said, context specificity was combined with residual error in these sub-studies. While G-Theory reports in some disciplines do not report residual error (or leave it to the reader to know that it is part of the highest-level interaction of all facets), others do report it specifically. Within this dissertation Chapter 4, it is specifically reported; tables of these interactions include the descriptor *residual error*. These interactions include error, and so any implications from variance of these components is also based on prior literature.

Generalizability Theory for Validation in Pharmacy Education. While validation evidence is crucial, it is under-reported in the pharmacy education literature. For pharmacy education, G-Theory appears to be a promising tool to help generate validation evidence, as well as inform validity for future uses of that learning assessment.

As shown, it is flexible for specifics of a learning assessment; specifics which will likely differ among pharmacy education institutions. These sub-studies demonstrated that G-Theory was a viable tool to parse variance from multiple sources in the complexity of skills-based assessment, as well as to integrate test-scores from multiple exams towards reliability of the course-grade, or from multiple quizzes with an exam towards reliability of the course-grade. In all scenarios, G-Theory can help to optimize a learning assessment's reliability, rigor, and fairness.

Over all three sub-studies, relative contributions to variance in scores from multiple sources (students, stations, weeks/occasions, and items) were identified. While commonly assumed by various instructors in pharmacy education, students' ability was not the only source of variance for any of these assessments.

Summary. Principles and implications can be drawn from these pharmacy education applications and especially how they are informed by applications in other areas of health-professions education and higher education more generally. While a number of these follow, it must first be repeated that each institution's context and details of each learning assessment can differ; thus, validation evidence should be generated by each institution for their own learning assessments. Validation from one institution should not be assumed for another institution—validation is and should be a local activity. Enough differs among colleges/schools of pharmacy with needs and resources, that reliability (and validity ultimately) could differ with various educational situations.

Recommended Best Practices for Validation. Drawn from these sub-studies as well as literature on validation, a number of best practices for validation were identified.

First and foremost, validation is institution-specific (Peeters & Martin, 2017). Specific variables that affect scores on a learning assessment can vary among institutions because the needs, intentions, and resources of each institution will vary. Thus, specifics of each institution's context (e.g., their assessments' intentions, needs, resources, and staffing) should direct their own validation of a learning assessment (including its design). For instance, didactic courses may have multiple assessed projects along with exams.

Secondly, paying specific attention to sampling of items is essential in designing a learning assessment. There are profound implications to validity with both covering the needed test content, but also for reliability. As multiple independent sampling has shown, having more independent data sources (e.g., items, OSCE stations, multiple exams, or multiple quizzes) can improve rigor for the entire learning assessment (Hansen, 2016; van der Vleuten, 2005).

Thirdly, context specificity (also known as content specificity or case specificity) is a notable and common scourge for assessment in education. To overcome context specificity, sampling widely and multiple times from a content domain are needed. Scores from a learning assessment have sources of variation other than students; thus, to reliably and fairly differentiate among students, multiple independent sampling is needed. This means multiple related exams towards a course grade (and even more exams if they are shorter in length), using related quizzes to help improve insufficient reliability from

an exam, as well as evaluating multiple related tasks within a performance-based assessment.

Applying these principles for validation to learning assessments beyond those studied herein, more related assessments should help rigor as well. As a prominent specialist in medical education, van der Vleuten (2005) notes "any method can be sufficiently reliable, provided sampling is appropriate across conditions of measurement" (p. 312). Thus, in a course with multiple exams, introducing an assignment (or multiple assignments) aligned with course objectives, could be helpful. As multiple related assessments provide more confidence and slightly different information; breaking the assignment down into pieces that require different related skills (e.g., introduction, methods, results/interpretation) could be advantageous. Each independent (though related) piece could be scored separately and combined together into a more reliable course grade—a grade more reliable than if only one overall project grade was provided at the end of a course.

Implications

Reliability is a critical component of validity, especially for high-stakes testing. Evidence for validation of learning assessments is part of accreditation standards in pharmacy education (Accreditation Council for Pharmacy Education, 2014). The sub-studies within this dissertation illustrate a number of generalizable findings.

First, each specific institution has their own unique needs and available resources. Learning from others' data is not the same as validation with your own institution's data. General trends in D-Studies can be generalizable among different institutions. Although. specific numbers of any D-Study can be questionable to generalize beyond one

institution's context. For instance, one institution may provide substantial training of their item writers, while another institution takes a more hands-off approach to development of item-writers. Thus, items created by educators at each institution may differ in quality, and the reliability of exams at those institutions may be notably different, although this can even differ between individual educators at any single institution. In another instance with a performance-based assessment, an institution may be able to provide multiple raters for each station although not have scheduling or physical space to allow a large number of stations; meanwhile, another institution may have fewer available raters and more available space—which could facilitate more stations. If reliability is suboptimal, either scenario could be better than doing nothing. That said, content specificity often shows a general preference for more stations over more raters per station if an institution is able (van der Vleuten, 1996). G-Theory can enable this decision-making as it both computes the specific reliability for an administration, as well as helps explore and identify how this assessment could better be optimized for a specific institution's needs and available resources.

Second, context specificity (also called content specificity or case specificity; Eva, Neville, & Norman, 1998) has been a common finding and notorious, influential scourge of any learning assessment. Eva (2003) aptly concludes "context specificity is a profoundly general phenomenon" (pg.588). Furthermore, another prominent medical education specialist, Van der Vleuten (2014), advances that "today, context specificity is almost a platitude" (pg.234). Context specificity can be defined "by the observation that an individual's performance on a particular problem or in a particular situation is only a

weak predictive of the same individual's performance on a different problem or in a different situation" (Eva, 2003, pg.587).

As a case in point, OSCEs are the gold-standard for performance-based assessment *because* of their multiple situations within a single assessment (Eva, 2003). It is a finding and issue that is generalized to "virtually all setting in areas of medical education including problem-solving, clinical skills performance, professionalism, communication, team performance and leadership" (van der Vleuten, 2014). So, overcoming context specificity appears centrally-involved in creating rigorous, reliable performance-based assessments, and specifically, performance-based assessments necessitating multiple tasks.

However, many assessments in pharmacy education do not reflect this issue of context specificity. They appear to assume all variations is from students alone. While it can depend as to just how many tasks are needed, it seems prudent that reliability be computed to confirm a learning assessment as acceptably reliable (i.e., this is one use of G-Theory).

Third, a student's ability is not the only source of variance. It is one of multiple sources of variance in scores, and so a student's ability should not be construed as the only or even largest source of variation (as it appears it is often assumed unfortunately). Recall that reliability describes the consistency of association for final scores with students' ability. Thus, other non-student sources of variation (e.g., stations, raters, weeks/occasions, and items) can lower reliability. However, improved sampling of these other non-student sources can improve reliability. Improved sampling can help a learning

assessment become better at statistically-discriminating within a certain cohort of students.

Fourth, internal consistency from multiple exams can be combined towards a course-level reliability for course grades. Recent computer-based software and its reports has enabled educators to more easily examine the internal consistency and item-analysis for their testing; though this internal consistency (undoubtedly) will differ for different exams. However, these multiple exam internal consistencies can be combined. Reliability at a course-level can be analyzed. Many times, high-stakes decisions with progression of PharmD students may be based on patterns of their course-grades; understanding reliability of these course-grades may be more helpful than looking at individual learning assessments.

Fifth, quizzes, with many occasions yet fewer questions on each occasion, can also augment towards an acceptable course-level reliability. However, the many occasions with multiple quizzes could also introduce further sources of construct-irrelevant variance. That is, if the quizzes are not aligned with one another, they may be trying to measure something different than important course content; there are many variables with multiple quizzes that potentially may or may not add to a course-grade. That said, learning assessments at any institution may differ in specific variance from multiple context sources (e.g., teacher, students, educational environment), and so validation should be undertaken with data from those specifics at each institution. A pedagogy that may use quizzes extensively (e.g., team-based learning, flipped classrooms), may benefit from both noting if and how quizzes align with the exam (noted

in the variance component from each quiz), and quantifying how the weight of quizzes could help the reliability of the course exam(s).

Sixth, and most often related to performance-based assessments, a further implication is the commonly-erred focus by many on "inter-rater reliability" and rater training. True, this can be one focus (if needed), but should not be the only focus (and may not even be necessarily needed in some cases; Serres & Peeters, 2012). In fact, Newble and colleagues (1980) showed in a controlled experiment that rater training can have a minimal effect for most raters. Van der Vleuten and colleagues (1989) confirmed and extended this in a randomized controlled trial. Therein, the overall effect of rater training was minor or possibly detrimental. Rater training was most helpful for inexperienced non-experts, while there was minimal effect among experts (although it also appeared to make some types of errors worse among those same experts). While rater training can use substantial resources to undertake, its benefit may be little to none.

Recommendations for Future Research

Reliability is central to fairness in testing (AERA et al, 2014). It is vital that reliability be analyzed for fairness of learning assessments in pharmacy education. Exams with high-stakes should be prioritized, but ideally reliability should be analyzed and examined for all testing. That said, principles from higher-stakes testing can be applied to learning assessments with lower-stakes. However, specifics of a learning assessment can differ among institutions—different students, different learning environments, as well as different numbers of questions, stations, items, and raters. As a result, each pharmacy education institution should generate validation evidence for learning assessments within their unique context.

In various situations (such as OSCEs), the impact of more occasions, beyond providing more questions, has received little research attention. Might early occasions have a formative effect on later exam occasions?

A single well-designed performance-based OSCE is not the only method that should be used within a program of assessment. Yes, that OSCE, especially if used for high-stakes decision-making, should be sufficiently rigorous (reliable, and also driven by content that is valid). However, a program of assessment should use many other assessments toward the complex sundry of integrated competencies needed from a pharmacy education graduate. No single learning assessment method can do this alone (van der Vleuten, 2005). A portfolio may be more appropriate to integrate data from multiple sources (Heeneman, Pool, Schuwirth, van der Vleuten & Driessen, 2015). That said, pharmacy education should strive to improve the quality and rigor of learning assessments—with variability coming from sources beyond what has most often bee described in the pharmacy education literature. Test-scores for fewer, more rigorous learning assessments can set the stage for a more rigorous program of assessment.

There is a continuum between individual learning assessments, and those used in aggregate within programmatic assessment. Thus, programmatic assessment can be seen as an extension from the findings of this dissertation. It may be helpful to evaluate an entire professional program (e.g., all four years of a PharmD program) to identify if and when reliability is sufficient to make decisions regarding progression in program, as well as which learning assessments are truly helpful towards the program's reliability. Might a portfolio be helpful to employ and triangulate assessment information from multiple sources (e.g., Heeneman, Pool, Schuwirth, van der Vleuten & Driessen, 2015)? Within a

G-Theory framework, might independent learning assessments from multiple sources in a program be integrated in a portfolio, such that review of portfolios could summarize and built towards a better composite reliability for the whole program? As a consequence, the programmatic decisions should consistent (reliable) and based on rigorous evidence.

Reflexivity

Of note for this dissertation is that the dissertation author was the instructor of record for the OSCE. This is notable because he also served as one of the twenty faculty raters and may have introduced bias into scores for some students. However, given the multiple raters for each student, the underlying theory of multiple sampling, and the overall reliability of scores from the assessment, this bias should be minimal.

Take-Home Message

If scores from a learning assessment are to matter, then context should matter, validation should matter, and reliability should matter.

References

Accreditation Council for Pharmacy Education. (2015). Accreditation Standards and Key Elements for the Professional Program in Pharmacy Leading to the Doctor of Pharmacy Degree ("Standards 2016"). Retrieved from: https://www.acpe-accredit.org/pdf/Standards2016FINAL.pdf.

Al-Jarallah, K.F., Moussa, M.A.A., Shehab, D., & N.Abdella. (2005). Use of interation cards to evaluate clinical performance. *Medical Teacher, 27*(4)*,* 369-374.

Allalouf, A., Klapfer, G., & Fronton, M. (2008). Comparing vertical and horizontal scoring of open-ended questionnaires. *Practical Assessment, Research & Evaluation*, *13*(8), 2.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, L.W., & Krathwohl, D.R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY: Longman.

Atilgan, H. (2013). Sample size for estimation of g and phi coefficients in generalizability theory. *Egitem Arastirmalari-Eurasian Jornal of Educational Research, 51,* 215-228.

Atkinson, R.K., Derry, S.J., Renkl, A., & Wortham, D. (2000). *Review of Educational Research, 70*(2), 181-214.

Bergman, E., de Feijter, J., Frambach, J., Godefrooij, M., Slootweg, I., Stalmeijer, R., & van der Zwet, J. (2012). AM Last Page: A guide to research paradigms relevant to medical education. *Academic Medicine, 87*(4), 545.

Berliner, D.C. (2002). Educational Research: The hardest science of all. *Educational Researcher*, *31*(8), 18-20.

Bloch, R., & Norman, G. (2012). Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. *Medical Teacher, 34*(11)*,* 960-992.

Boorsbom, D, Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Boursicot, K.A.M., Roberts, T.E., & Burdick, W.P. (2014). Structured assessments of clinical competence. In T Swanwick (Ed). *Understanding Medical Education: evidence, theory and practice* (pp. 293-304). Chichester, UK: Wiley Blackwell.

Brannick, M.T., Erol-Korkmaz, H.T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education*, *45*(12), 1181-1189.

Brennan, R.L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21.

Briesch, A.M., Chafouleas, S.M., & Johnson, A. (2016). Use of generalizability theory within K–12 school-based assessment: A critical review and analysis of the empirical literature. *Applied Measurement in Education*, *29*(2), 83-107.

Cook, D.A., Brydges, R., Ginsburg, S., Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education, 49*(6)*,* 560-575.

Cor M.K., & Peeters, M.J. (2015). Using generalizability theory for reliable learning assessments in pharmacy education. *Currents in Pharmacy Teaching and Learning, 7*(3)*,* 332-341.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: a key to unlock professional assessment. *Medical Education, 36*, 972-978.

Crossley, J., Russell, J., Jolly, B., Ricketts, C., Roberts, C., Schuwirth, L., & Norcini, J. (2007). 'I'm pickin' up good regressions': the governance of generalisability analyses. *Medical Education, 41*, 926-934.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109-117.

DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education*, *34*(5-6), 576-591.

Driessen, E., Van Der Vleuten, C., Schuwirth, L., Van Tartwijk, J., & Vermunt, J.D.H.M. (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. *Medical Education*, *39*(2), 214-220.

Eva, K. W., Neville, A. J., & Norman, G. R. (1998). Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. *Academic Medicine 73*(Suppl 10), S1-S5..

Eva, K.W. (2003). On the generality of specificity. *Medical Education*, *37*(7), 587-588.

Fielding, D.W., & Regehr, G. (2017). A call for an integrated program of assessment. *American Journal of Pharmaceutical Education, 81*(4), article 77.

Gao, X., Shavelson, R.J., & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science: promises and problems. *Applied Measurement in Education, 7*(4), 323-342.

Haertl, E.H. (2006). Reliability. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed, pp. 65-110). Portsmouth NH: American Council on Education.

Hambleton, R.K., & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47.

Hanson, M.D., Kulasegaram, K.M., Coombs, D.L., & Herold, J. (2012). Admissions files review: applying the multiple independent sampling (MIS) methodology. *Academic Medicine, 87*(10), 1335-1340.

Hanson, M.D., Woods, N.N., Martimianakis, M. A., Rasasingham, R., & Kulasegaram, K. (2016). Multiple independent sampling within medical school admission interviewing: an "intermediate approach". *Perspectives on Medical Education*, *5*(5), 292-299.

Heeneman, S., Oudkerk Pool, A., Schuwirth, L.W., van der Vleuten, C.P., & Driessen, E.W. (2015). The impact of programmatic assessment on student learning: theory versus practice. *Medical Education*, *49*(5), 487-498.

Hendrickson, A., & Yin, P. (2010). Generalizability Theory. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 115–122). New York, NY: Routledge.

Hodges, B. (2003). OSCE! Variations on a theme by Harden. *Medical Education*, *37*(12), 1134-1140.

Hoover, M.J., Jung, R., Jacobs, D.M., & Peeters, M.J. (2013). Educational testing validity and reliability in pharmacy and medical education literature. *American Journal of Pharmaceutical Education, 77*(10)*,* article 213.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed, pp. 17-64). Portsmouth NH: American Council on Education.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Khan, K.Z., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The objective structured clinical examination (OSCE): AMEE guide. 81. Part I: an historical and theoretical perspective. *Medical Teacher, 35*(9), e1437-e1446.

Lane S., Raymond, M.R., & Haladyna, T.M. (2016). *Handbook of Test Development*. New York, NY: Taylor & Francis.

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, *65*(9), S63-7.

Moonen-van Loon, J.M.W., Overeem, K., Donkers, H.H.L.M., van der Vleuten, C.P.M., Driessen, E.W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Sciences Education, 18*(5)*,* 1087-1102.

Munoz, L.Q., O'Byrne, C., Pugsley, J., & Austin, Z. (2004). Reliability, validity and generalizability of an objective structured clinical examination (OSCE) for assessment of entry-to-practice in pharmacy. *Pharmacy Education, 5*.

Newble, D.I., Hoare, J., & Sheldrake, P.F. (1980). The selection and training of examiners for clinical exams. *Medical Education*, *14*(5), 345-349.

Newble, D. (2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education*, *38*(2), 199-203.

Norman, G. (2017). Generalization and the qualitative-quantitative debate. *Advances in Health Sciences Education: Theory and Practice, 22*(5), 1051-1055.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943.

Onwuegbuzie, A.J., & Leech, N.L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International Journal of Social Research Methodology, 8*(5), 375-387.

Peeters M.J., Beltyukova, S.A., & Martin, B.A. (2013a). Educational testing and validity of conclusions in the scholarship of teaching and learning. *American Journal of Pharmaceutical Education, 77*(9), article 186.

Peeters, M.J., & Cor, M.K. (2019). High stakes testing: Implications for assessment practice in pharmacy education. *Currents in Pharmacy Teaching & Learning, 11*(12) [online ahead of print]

Peeters, M.J., & Martin, B.A. (2017). Validation of learning assessments: A primer. *Currents in Pharmacy Teaching & Learning, 9*(5)*,* 925-933.

Peeters, M.J., Serres, M.L., & Gundrum, T.E. (2013b). Improving reliability of a residency interview process. *American Journal of Pharmaceutical Education, 77*(8)*,* article 168.

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Medical Teacher, 32*(10)*,* 802-811.

Peng, S.K.L.P.Z. (2007). Classical versus Generalizability Theory of measurement. *Examinations Measurement, 4*, 009.

Renkl, A. (2002). Worked-out examples: Instructional explanations support learning by self-explanations. *Learning & Instruction, 12*, 529-556.

Rios, J.A., Li, X., & Faulkner-Bond, M. (2012, October). *A review of methodological trends in generalizability theory.* Paper presented at the annual conference of the Northeastern Educational Research Association, Rocky Hill, CT.

Serres, M.L., & Peeters, M.J. (2012). Overcoming content specificity in admission interviews: the next generation? *American Journal of Pharmaceutical Education*, *76*(10), 207.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Sherbino, J., Kulasegaram K., Worster, A. & Norman G.R. (2013). The reliability of encounter cards to assess the CanMEDS roles. *Advances in Health Sciences Education, 18*(5)*,* 987-996.

Sireci, S.G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25*(3), 27-34.

Smith, B.H. (2006). Rear end validity: A caution. In Bootzin, R.R., & McKnight, P.E. (2006). *Strengthening Research Methodology: Psychological Measurement and Evaluation*. American Psychological Association, 233-247.

Streiner D.L., Norman, G.R., & Cairney J. (2015). *Health Measurement Scales*. (5th ed.). New York, NY: Oxford University Press.

Sturpe, D.A. (2010). Objective structured clinical examinations in Doctor of Pharmacy programs in the United States. *American Journal of Pharmaceutical Education, 74*(8), article 148.

Swanson, D.B., Clauser, B.E., & Case, S.M. (1999). Clinical skills assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security. *Advances in Health Sciences Education, 4*(1), 67-106.

Swanson, D.B., Norman, G.R., & Linn, R.L. (1995). Performance-based Assessment: lessons from the health professions. *Educational Researcher, 24*(5), 5-11, 35.

Swanson, D.B. & van der Vleuten, C.P.M. (2013). Assessment of clinical skills with standardized patients: state of the art revisited. *Teaching and Learning in Medicine, 25*(S1), S17-S25.

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation based assessment of clinical competence. *Advances in Health Sciences Education*, *23*(2), 323-338.

Tavakol, M., & Dennick, R. (2012). Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE guide no. 66. *Medical Teacher, 34*(3), e161-e175.

Ten Cate, O., & Regehr, G. (2019). The Power of Subjectivity in the Assessment of Medical Trainees. *Academic Medicine, 94*(3), 333-337.

Tweed, M., & Miola, J. (2001). Legal vulnerability of assessment tools. *Medical Teacher, 23*(3)*,* 312-4314.

Van Bavel, J.J., Mende-Siedleckia, P., Bradya, W.J., & Reinero, W. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Science, 23*, 6454–6459.

van der Vleuten, C.P.M. (1996). The assessment of professional competence: developments, research and practical implications. *Advances in Health Sciences Education*, *1*(1), 41-67.

van der Vleuten, C.P.M. (2014). When I say… context specificity. *Medical Education*, *48*(3), 234-235.

van der Vleuten, C.P.M., Luyk, S. V., Ballegooijen, A.V., & Swanson, D.B. (1989). Training and experience of examiners. *Medical Education*, *23*(3), 290-296.

van der Vleuten, C.P.M., & Schuwirth, L.W. (2005). Assessing professional competence: from methods to programmes. *Medical Education*, *39*(3), 309-317.

Worthen, B.R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan, 74*(6), 444-454.

Zibrowski, E.M., Myers, K., Norman, G., & Goldszmidt, M.A. (2011). Relying on others' reliability: challenges in clinical teaching assessment. *Teaching and Learning in Medicine, 23*(1), 21-27.