American Journal of Pharmaceutical Education **2013; 77 (9) Article 186.**

SPECIAL ARTICLES

Educational Testing and Validity of Conclusions in the Scholarship of Teaching and Learning

Michael J. Peeters, PharmD, MEd,[a] Svetlana A. Beltyukova, PhD,[b] and Beth A. Martin, PhD, MS[c]

[a]College of Pharmacy and Pharmaceutical Sciences, University of Toledo, Toledo, Ohio
[b]Judith Herb College of Education, University of Toledo, Toledo, Ohio
[c]School of Pharmacy, University of Wisconsin-Madison, Madison, Wisconsin

Submitted May 21, 2013; accepted July 27, 2013; published November 12, 2013.

Validity and its integral evidence of reliability are fundamentals for educational and psychological measurement, and standards of educational testing. Herein, we describe these standards of educational testing, along with their subtypes including internal consistency, inter-rater reliability, and inter-rater agreement. Next, related issues of measurement error and effect size are discussed. This article concludes with a call for future authors to improve reporting of psychometrics and practical significance with educational testing in the pharmacy education literature. By increasing the scientific rigor of educational research and reporting, the overall quality and meaningfulness of SoTL will be improved.

Keywords: psychometrics, educational testing, scholarship of teaching and learning, reliability, validity

INTRODUCTION

The rigor of education research, including research in medical education, has been under scrutiny for years.[1,2] On the technical side, issues raised include lack of examination of the psychometric properties of assessment instruments and/or insufficient reporting of validity and reliability.[3-5] On the applied side, researchers have frequently based their conclusions on significance without addressing the practical implications of their findings.[6] These issues appear even more pronounced in the pharmacy education literature. In a review of over 300 articles published in pharmacy and medical education journals using educational tests, Hoover and colleagues found that pharmacy education articles much more often lacked evidence of reliability (and consequently validity) than did medical education articles, while neither consistently reported validity evidence.[7] While not specifically evaluated in that study, few pharmacy education articles reported an effect size of their studied intervention (MJ Hoover, e-mail, April 17, 2013).

It is encouraging that diverse pharmacy education instructors have authored many of the reviewed articles, representing a scholarship of teaching and learning (SoTL). However, authors still need to actively pursue psychometric evaluation of their student-learning assessments and examine the practical significance of the results. Increasing the technical rigor of research and reporting effect sizes will increase the overall quality and meaningfulness of SoTL. While doing so can be challenging, it can be accomplished without formal training. Just as scientists would not conduct experiments without verifying that their instruments were properly calibrated and would not claim that an experiment worked without indicating the magnitude of the effect, a SoTL investigator should not presume an assessment instrument's reliability and validity but rather should seek evidence of both prior to attempting statistical analyses and interpret the results of those analyses from the perspective of educational significance (ie, effect size). This should be standard practice not only for standardized tests but also for other types of assessments of student knowledge and abilities, including performance-based assessments (eg, objective structured clinical examinations [OSCEs]) and traditional classroom assessments (eg, assessments with true/false, multiple-choice questions, case clinical notes, short-answer questions, and essay questions) used in SoTL.

This paper can be seen as an extension of a measurement series in *Medical Education*[8] for a SoTL audience, wherein it explicitly discusses the interrelatedness of psychometrics, statistics, and validity of conclusions. It is intended as a less-technical review of several established practices related to reporting educational test psychometrics and effect sizes, while also explaining how addressing both will contribute important evidence to the overall

Corresponding Author: Michael J. Peeters, PharmD, MEd, University of Toledo College of Pharmacy and Pharmaceutical Sciences, 3000 Arlington Avenue, MS 1013, Toledo, OH 43614. Tel: 419-383-1946. Fax: 419-383-1950. E-mail: michael.peeters@utoledo.edu

validity of data-based conclusions. Some of these practices involve statistical computations while others are based on logic. Following these practices should help SoTL investigators, who may not have formal training in psychometrics or statistics, to increase the rigor of their scholarship. We also offer a brief overview of some major advanced psychometric models that can be used to obtain further validity evidence. It is beyond the scope and focus of this paper to show how to create and administer assessments or how to calculate most statistics. We hope that the level of language, ideas, and examples herein will be relevant to the diverse readership. Examples from published studies, mainly in pharmacy education, are provided to illustrate some of the ways in which SoTL researchers could report findings.

VALIDITY

By its traditional definition, validity refers to the degree to which a test accurately and meaningfully measures what it is supposed to measure. The seminal work, *Test Validity and the Ethics of Assessment* reminds us that validity also refers to the appropriateness of inferences or conclusions from assessment data and emphasizes that it is an ethical responsibility of researchers to provide evidence in support of their inferences.[9] The more convincing the evidence, the stronger the validity of inferences and the higher the degree to which researchers' interpretations of assessment results are justifiable and defensible. With the focus on the nature of evidence underlying validity of inferences, the unitary validity framework presented in this text forms the basis of current testing standards.[10]

Differing from an older framework comprised of 3 separate types of validity (ie, content, criterion, and construct), Messick argues that "different kinds of inferences . . . require different kinds of evidence, not different kinds of validity"[9] and presents the current standards according to which researchers should think of validity as 1 unifying concept instead of several separate types of validity.[9] Further, researchers are advised to consider reliability as evidence of validity and not as a separate statistic. Approaching validity as 1 holistic concept allows researchers to focus on the evidence they need to collect to be confident in the quality of their assessment instrument. This evidence typically involves reliability or stability of the instrument, discussion of the content relevance of the items, evidence that the items form a stable linear measure that is able to differentiate more-able from less-able persons in a way that is meaningful and consistent with the theory. It is also sometimes necessary to establish that an assessment produces results comparable to some other well-known instrument or functions the same way with different subgroups; that is, researchers must frequently consider multiple sources of validity evidence to be able to argue with confidence that their assessment instrument generates meaningful data from which justifiable conclusions can be drawn. No single validity source can provide such evidence. Taking an evidence-seeking approach to validity also implies that validity is contextual and that gathering such evidence is a process wherein researchers seek their own evidence each time they use an assessment based on this proposed purpose, use, and interpretation. For this reason, researchers should not solely rely on validity evidence reported by others. As overwhelming as this may sound, it is a doable task that does not necessarily require advanced psychometric training. Validity is a matter of degree and researchers can always find ways to gather validity evidence at the level of their own expertise and may seek help from a psychometrician when needed. Much of validity evidence comes in the form of words and logical arguments, while some (eg, reliability) may involve statistical applications and even advanced psychometric analyses.

For example, every researcher should be able to provide evidence of content relevance and content coverage, and the process of gathering this evidence should start prior to administering an educational test. As part of the process, the researcher should operationally define the knowledge, skills, or abilities that are being measured. This does not require psychometric expertise and is deeply grounded in content expertise of the investigator(s) or other subject matter experts. To illustrate, let us examine an assessment of a physician's ability to practice medicine. To determine if the assessment we are considering is the right one to use, we need to reflect on how we define this ability and then draw on existing evidence and theories. Alternatively, we could match test items to a blueprint created after surveying numerous other content experts in the practice of medicine. If we define a physician's ability to practice medicine as the ability to apply knowledge, concepts, and principles, and to demonstrate fundamental patient-centered skills, the United States Medical Licensing Examination Step 3 would be a test of choice[11]; however, this examination would not be appropriate to use in nurse or pharmacist licensing because validity is contextual. For reporting purposes, an operational definition of what is being measured should be explicitly presented, given that inferences about the construct evaluated are based on this definition. The operational definition also becomes the driving factor in determining whether the right questions are included and whether they are constructed in such a way that they will elicit the needed information. Later in the research process, all of this knowledge about content relevance and coverage becomes evidence that is used to argue the validity of test-score interpretations and inferences about the construct

after the statistical analyses are completed. In the example above, the term "validity of an examination" was deliberately avoided; instead, the focus was on "validity of inferences" or "validity of conclusions" from the data generated by an examination in this sample of participants.

Gathering validity evidence does not stop at the item review and instrument selection stages. Although evidence of content coverage and relevance contributes important information about the construct being measured and might influence the nature of inferences, it cannot be used in direct support of inferences from the scores.[9] Therefore, while it is important to collect content-related evidence, investigators also need to seek other evidence after participants have completed the test, focusing on internal structure evidence, including reliability. Collection of this evidence involves investigating the extent to which survey instrument items function well together to measure the underlying construct in a meaningful way, and for this task, researchers typically consider several different options, such as computing reliability indices, conducting an item analysis, using factor analysis, using generalizability theory, and applying item response theory. If problems such as low internal consistency or inter-rater reliability, lack of the meaning of a variable, poor item fit, multidimensionality, lack of variance explained, or inconsistent item functioning across different subgroups of respondents are discovered, interpretations of the results and inferences about the construct should not be attempted. Instead, the investigator should go back to the drawing board, revisit the theory behind the construct, reexamine content relevance and coverage, and start the process again until the content and structure-related evidence points to good psychometric properties of an assessment instrument.

RELIABILITY

Reliability — reproducibility or consistency of assessment scores within a sample of participants — is a major initial quantitative "quality index" of the assessment data as well as essential evidence toward the overall validity. However, as with any other validity evidence, its use alone is not sufficient for arguing overall validity. Reliability alone has limited usefulness because it indicates only that an assessment's items measure something consistently. Reliability does not indicate *what* knowledge, skills and/or abilities are being measured. Thus, along with reliability, other validity evidence is crucial before any validity conclusions can be made.

While seeking reliability evidence, researchers should select the most appropriate type and explicitly identify this type while reporting a SoTL study. Two common types of reliability include internal consistency reliability and inter-rater reliability. The first is particularly relevant for many SoTL investigators. The internal consistency reliability index shows the extent to which patterns of responses are consistent across items on a single test occasion, whereas the inter-rater reliability index indicates consistency across raters and is reported when judges or scorers are involved. That said, it is also possible and sometimes necessary to report more than 1 reliability index (eg, when using multiple educational tests). In many instances, reliability is specific to each use of an assessment and can change when the assessment is used with another group or when an assessment is even slightly modified. Thus, it is prudent to report reliability for every occasion within any scholarship.[12]

Reliability can be relative or absolute.[13,14] When repeated measurements are used, relative reliability refers to a consistency or reproducibility of rankings of scores in a sample, whereas absolute reliability refers to the degree to which individual scores are reproduced (eg, when judges agree on 1 specific score). A relative reliability index can be helpful when student performances are ranked in relation to 1 another, whereas absolute reliability could be helpful when the focus is on specific scores from an evaluation. Of the 2, absolute reliability is often omitted, even though it is considered by some to be of greater clinical[15] and educational[16-18] value. Standard error of measurement (SEM) is a recommended absolute reliability index to report, whereas internal consistency reliability would be a choice as relative reliability index. That said, both absolute and relative reliability should be assessed when appropriate.

Internal Consistency Reliability

Internal consistency reliability is the relative reliability of items within an assessment. It is most easily obtained for a quiz or examination wherein all the items are on the same scale and in the same format (eg, multiple-choice, short-answer, or longer-answer clinical cases). Cronbach alpha is typically calculated when the questions on a test are on a rating scale, while a Kuder-Richardson formula 20 (KR-20) is applied to dichotomous (yes/no) or multiple-choice testing with only 1 correct answer. Commonly accepted ranges for internal consistency reliability are widely available, and often a coefficient >0.7 is sufficient for a classroom test used in SoTL (ie, some random inconsistency is not only expected but also allowed because no measurement can ever be perfect). However, life-changing decisions necessitate higher reproducibility of test scores, and internal consistency should be >0.80 or >0.90 on high or very high-stakes assessments.[19] In an example of reporting this relative internal consistency reliability type, the authors not only reported Cronbach's

alpha for each of their 4 assessments but also evaluated their indices against their minimum *a priori* Cronbach alpha level.[20]

When internal consistency reliability is low, there are several ways to improve it. The most common strategies include: increasing the number of response options/distractors (eg, 5 possible answers instead of 3), and using a greater number of related questions on a test, including items that range in difficulty so that the test would not discourage low-ability students but would still be challenging enough for high-ability students. Notable sources suggest that the number of response options should be limited[21] and that reliability may actually decrease if those extra distractors are not plausible. Some of the above strategies require only content expertise while others rely on the researcher's knowledge of computing item-discrimination and item-difficulty indices. A detailed discussion of these item-analysis indices is beyond the scope of this paper but can be found in other sources.[22] Researchers are strongly encouraged to evaluate and, if necessary, improve internal consistency reliability of their tests to improve precision of their assessments and ultimately influence the accuracy of conclusions and decisions based on the results.

Standard Error of Measurement

Test scores are not perfect representations of student knowledge, skills, or abilities, and SEM is capable of capturing this inherent imprecision in scores. Assessing the extent of measurement error is important for the validity of inferences from the scores. SEM is an *absolute* reliability index that can be used when seeking reliability evidence, as it shows the extent to which individual scores would be repeated if students were tested again and again. As such, it should not be confused with the standard error of the mean (also commonly and confusingly abbreviated as SEM), which refers to groups rather than individuals and shows the extent to which a sample/group mean score would be reproducible if the test were administered again and again to different same-sized samples.

When the focus is on individual student scores rather than group means, SEM can be quite helpful. The following formula may facilitate understanding of the concept of SEM,[17,18] and can employ either an internal consistency or inter-rater reliability coefficient.

$$SEM = SD \times \sqrt{(1\text{-reliability})}$$

SD is the standard deviation of test scores among examinees, whereas reliability is a test's KR-20, Cronbach's alpha, intraclass correlation (ICC), or Cohen's kappa coefficient

As suggested by the formula, SEM is directly related to a test's reliability, uses that test's standard deviation (SD), and is reported in the units of each specific test. Once computed, SEM can be used to develop confidence intervals for interpretation of test scores and thus represents important validity evidence. By convention, ± 1 SEM around a test score yields a 68% confidence interval around that test score, and ± 2 SEM yields a 95% confidence interval. The latter is the most typical in medical research and may be the most applicable for SoTL research as well. The choice of interval depends on the desired level of precision, with greater confidence being expected in high-stakes testing situations, wherein SEM has most often been used. We could not find an example of reporting SEM with internal consistency reliability in the pharmacy education literature. An example from medical education can be found in an evaluation of SEM in borderline (pass/fail) scores of medical students within a medical school's integrated assessment program.[23]

Inter-rater Reliability

As the name suggests, inter-rater reliability would be the choice if an assessment involves multiple raters or judges. Investigators typically need to consider how consistent the judges are and how much agreement they demonstrate when, for example, observing a skill performance. Judge consistency, or inter-rater consistency, reflects the extent to which multiple raters are in consensus about which examinees are more knowledgeable, able, or skilled, and which are less so. Therefore, high inter-rater consistency means that judges produce similar rankings of examinees' scores. As such, this type of reliability would be an example of *relative* reliability. Judge agreement, or inter-rater agreement, on the other hand, represents *absolute* reliability,[15] showing the extent to which raters give the same (and accurate) rating to the same examinee's skill performance.

Depending on the data, there are different statistics that can be computed for determining inter-rater consistency. The most known indices include an ICC for continuous data, and a Cohen kappa for categorical data. In the pharmacy education literature, there is an example of using an ICC to determine the degree of agreement between the analytical checklist scores obtained in 2 different conditions (real-time and video).[24] The ICC was 0.951, which the authors interpreted as high agreement (values of less than 0.4 indicated poor agreement; between 0.4 and 0.8, fair to good agreement; and greater than 0.8, excellent agreement). An example of Cohen kappa can be found in an analysis of letter grades from the 2008-2009 academic year using 2 independent faculty evaluations representing categorical-level data.[25] In this paper,

the researchers reported a Cohen kappa of 0.98 as evidence of inter-rater reliability.

There are several ways to address problems with inter-rater reliability. The most common strategies include increasing the number of observations by each rater, providing raters with detailed scoring descriptions in a rubric, and using a larger scoring scale. Increasing observations and increasing the scoring scale are similar to common strategies for improving internal consistency reliability of an assessment. A similar warning applies to the use of a larger scoring scale: while a larger scoring scale can be more reliable, raters may be less likely to agree with 1 another. Consequently, they may be imprecise in their evaluation of a examinee's performance, which would result in low inter-rater agreement (and absolute reliability). An example of seeking inter-rater consistency and inter-rater agreement is provided in Table 1, which also illustrates an impact of the length of the scoring scale using ICC and percent agreement. Including both consistency and agreement is not common in clinical research.[13,15,26] However, given that doing so is highly recommended, SoTL investigators should attempt to report both when student performances were judged. The pharmacy education literature includes a study in which both inter-rater consistency and agreement were reported. When scoring a PharmD program's admission essays, a balance between the 2 indices was sought, the result being a minimal SEM and an acceptably high ICC.[27] While use of ICC alone favored 1 rubric, including SEM showed a more complete picture, resulting in more confident evaluations with less measurement error and alterations in rubric design toward a more holistic scoring.

ADVANCED PSYCHOMETRIC MODELS

Complexity of research projects often calls for more advanced approaches to gathering validity evidence Table 2. As previously mentioned, some of the most popular approaches include generalizability theory, factor analysis, and item response theory.

Table 1. Examples of Inter-rater Reliability and Inter-rater Agreement*

Example 1

	Ratings for Each Student on a 9-Point Rating Scale					
Raters	**Student 1**	**Student 2**	**Student 3**	**Student 4**	**Student 5**	**Student 6**
Judge 1	5	6	7	8	9	8
Judge 2	6	7	8	9	9	9

Note: Inter-rater reliability as estimated by intraclass correlation (ICC) is 0.977. It is high because ratings given to each student by the 2 judges were similar (eg, 5 and 6 for Student 1, or 9 and 9 for Student 5). There was high consistency with student rankings between the judges; that is, Student 5 was at the top of both judges' distributions, followed by Student 4 and Student 6, then by Student 3, and Student 2. Both judges were in consensus that Student 1 was at the bottom. However, there was also variability between the judges in the ratings themselves. Percentage of agreement between judges was quite low, as the judges agreed on only 1 of 6 ratings (17%; for Student 5).

Example 2

	Ratings for Each Student on a 4-Point Rating Scale					
Raters	**Student 7**	**Student 8**	**Student 9**	**Student 10**	**Student 11**	**Student 12**
Judge 1	1	2	3	4	3	2
Judge 2	2	2	3	4	4	2

Note: Inter-rater reliability as estimated by intraclass correlation is 0.915. It is still high because ratings given to each student by the 2 judges remained similar (eg, 1 and 2 for Student 7, or 2 and 2 for Student 8 and Student 12). While consistency among judges remained quite high, the ICC here was lower than in Example 1 because there is less variability in scores by using a smaller 4-point scoring scale. However, in contrast to Example1, percentage of agreement between judges was considerably higher, as judges agreed on 4 of 6 ratings (67% for Student 8, Student 9, Student 10 and Student 12).

* Fewer rating-scale categories (ie, 4 instead of 9) resulted in more agreement among raters, making the standard error of measurement (SEM) smaller. The scale, however, should still discriminate. Although a 2-point scale may not discriminate into enough groups, it should, theoretically, yield a good agreement. Therefore, a balance of agreement using SEM and discrimination using ICC is needed. Support for the use of shorter scales also comes from cognitive psychology, which has shown that most people have trouble discriminating beyond 7 categories in many sensory domains; raters can differ considerably when more than 7 categories are used. Additionally, some argue that scales with an odd number of categories (ie, 5,7,9) have a problematic middle value and should be avoided and give preference to scales that have the same number of positive and negative response options (ie, balanced scales). Yet other investigators insist that the number of scale categories is an empirical question and will use a Rasch rating scale analysis to verify scale functionality. Researchers, however, do agree that longer scales (eg, 10-point scales) typically yield high measurement error.

Table 2. Methods of Collecting Validity Evidence and Enhancing the Quality of Scholarship of Teaching and Learning (SoTL)

Validity Evidence for SoTL	How to Evaluate and Report Validity Evidence in SoTL
Overall validity	Provide a clear definition of the knowledge, skills and/or ability this test is supposed to measure. Consider as many sources of validity evidence as possible: Evidence based on test content (more below) Evidence based on response processes Was a pilot test of the assessment done? Evidence based on internal structure of test (includes reliability; more below) Evidence based on relations to other variables What other evidence do you have that these learners improved their learning? Evidence based on consequences of testing Are there larger implications for pharmacy education or societal policy from these results? In manuscript's discussion, refer to evidence to argue validity of interpretations of test scores and inferences about the construct.
Evidence based on test content and relevance	Describe the intended learner population for whom the instrument was developed and report how representative your sample is of that population. Reference literature surrounding that content. Provide descriptive statistics of scores (ie, mean/median, range/deviation/interquartile range, percent pass). Give credibility/qualifications of test developer(s): how are they qualified for this test's content? Describe how the test items were generated (eg, existing items from another instrument, test plan or blueprint, include expert reviewers qualifications, if used).
Evidence of internal structure	Include scoring rubric information for the test, if applicable. Report indices of reliability. Consider factor analysis if a test has multiple sections or domains. Consider Rasch analysis when seeking comprehensive validation of a test and looking for proper ways to obtain person ability measures.
Reliability	Describe the sample that actually took the test. Provide a measure of internal consistency (eg, Cronbach's alpha, Kuder-Richardson formula 20 [KR-20]). If raters are used, report the extent to which students' scoring patterns are consistent between raters (eg, Cohen's kappa, intraclass correlation) and the level of agreement between raters. Consider generalizability theory analysis if a test has multiple variables and multiple sources of measurement variation (eg, multiple variables of stations, raters and students in an objective structured clinical examination (OSCE). Generalizability theory can provide 1 overall process reliability coefficient for the entire test.
Agreement	Report percentage of agreement or standard error of measurement (SEM) to establish the extent of imprecision in educational test scores.
Effect size	How large is the effect between sample groups? Report the effect size (eg, Cohen's d with pre/post or between-test-groups comparisons, R^2 as coefficient of determination).

Generalizability Theory

If a SoTL researcher is using a test that is similar to an OSCE, in which multiple raters are scoring items while nested in different stations, the reliability indices discussed above would not be sufficient, even if more than 1 is reported. While researchers could evaluate the internal consistency of the items at each station or the inter-rater reliability among 2 or more raters at each station, those reliability indices would quantify only the measurement error specific to the items or the raters at that 1 station but would fall short of capturing the reliability of all items and raters across all the stations. Generalizability theory

(G-theory) is a model that could be considered, as it simultaneously models variation in data from multiple sources (eg, stations, raters, and items) and provides 1 overall (combined) process reliability index for an *entire* multi-station assessment. In other words, this index includes both internal consistency and inter-rater reliability information. A further advantage of using G-theory is that subsequent decision-study analyses could identify which changes in the number of items, raters, and/or stations would improve the process reliability most beneficially for future assessments. As such, a distinct benefit with G-theory is that it can help optimize the number of raters and stations needed for a reliable test, which can be an important consideration given the limited curricular resources (ie, faculty, space, finances) available to pharmacy colleges and schools for assessment of student learning.

The G-theory model is widely accepted in medical education research and OSCEs are a noteworthy application of G-theory.[28] Using G-theory, important findings from OSCE-based evaluations have revealed that a global rating scale for each station (instead of a detailed checklist) was at least as *reliable* as a checklist over the multiple OSCE stations.[29] Regarding *validity* evidence, they also demonstrated that, among medical students, residents and physicians, global rating scales were able to detect changes in growing clinical expertise that checklists could not capture.[30] An example of using G-theory in pharmacy education can be found in research that analyzed interview data from pharmacy resident candidates.[31] Each candidate was interviewed by 8 interviewers, with 1 or 2 interviewers nested within their own interview session (or station); subsequent decision studies showed that placing one interviewer into multiple, separate stations was much more reliable than simply adding more interviewers to an existing station (or panel).

Factor Analysis

Factor analysis is another approach wherein investigators analyze correlations among test (or instrument) items. These correlations can be used as the internal structure evidence to support validity of conclusions. Within this analysis, inter-item correlations are uncovered, and items are grouped to represent different meaningful factors (or domains).[32] The most commonly reported statistics for factor analysis include inter-item correlations, eigenvalues, explained variance, and reliability indices for all factors and the entire assessment. In 1 use of factor analysis with data from 7 institutions, the researchers developed a professionalism assessment tool and examined its internal structure.[33] Although the 33-item instrument was generally used to assess an overall level of student pharmacist professionalism, 5 more-specific factors were unveiled through exploratory factor analysis, with subsequent confirmation within another larger cohort of students. This allowed the authors to better understand the specific dimensions of professionalism captured by their instrument as well as to assess the internal consistency reliability for each identified domain using a KR-20 statistic.

Rasch Analysis

Unlike factor analysis or G-theory, item response theory and specifically the Rasch model represent an advanced alternative to classical test theory. This model takes into consideration each student's ability and each item's difficulty, and examines how students interact with each item based on their abilities.[34] The Rasch model produces several useful statistics that provide rich evidence of validity. The most commonly reported statistics include: different reliability indices (Rasch reliability and separation), item fit (ie, how well the items function together as a unidimensional measure of some underlying construct), item-person map (ie, a visual "ruler" that allows researchers to determine qualitatively if the meaning of the measure matches the theory), scoring scale functioning (particularly useful with rating scales in that it shows empirically if all the categories in the scale are used by examinees consistently), and judging bias. The Rasch model generates estimates of item difficulty and person abilities and is sample-independent. As such, the Rasch model is used in high-stakes testing for physicians, chiropractors, and nurses, as well as in high school graduation examinations by numerous state departments of education. Reproaching the example from kappa above, that article is also an example of a Rasch analysis of a PharmD student presentation rubric and judge bias.[24] In that study, a Many-Facets Rasch model was used "to determine the rubric's reliability, quantify the contribution of evaluator harshness/leniency in scoring, and assess grading validity by comparing the current grading method with a criterion-referenced grading scheme." The researchers reported high rubric reliability ($r=0.98$) and recommended that several grades be adjusted to eliminate evaluator leniency, although they concluded that evaluator leniency appeared minimal. [24]

EFFECT SIZE

Once investigators are confident that they have assessed student performances accurately and meaningfully (ie, determined that their assessment instrument has good psychometric properties), they can conduct statistical analyses of the data and proceed with the interpretation of the results from the standpoint of both statistical and educational significance. The former involves reporting

a *p* value while the latter requires reporting an effect size.[6] Together with statistical significance, effect sizes provide powerful evidence of the validity of conclusions. For example, they may be particularly informative when significance is not found but a large change or difference is observed or when significance is found but the effect size is negligible. Without noting the size of the effect, the researchers may falsely conclude no effect of the intervention in the first situation and effect of the intervention in the second. In reality, however, the sample size may have been too small to reach statistical significance *or* the assessment tool may not have been sensitive enough to capture change or difference.

Two of the most common indices for effect size are Cohen's d and R^2. Cohen's d is used when researchers compare 2 groups using standard deviation units. With the help of 1 of numerous online calculators, it can be easily calculated once an investigator knows the means and standard deviations for each group. By convention, effect sizes of <0.2 are considered small; 0.5, medium; and 0.8, large. Since many meaningful educational interventions have had a medium-large effect size,[35] reporting an effect size value may assist readers in identifying interventions that might also be promising in their comparable educational settings. That said, effect size does not guarantee generalizability, although the likelihood that interventions with large effect sizes and proper experimental controls will replicate is high. An example of reporting an effect size can be seen in a study evaluating a prescribing error module.[36] In this study, the researchers compared 2 groups on 3 worksheets and noted a large effect of 0.85 associated with a significant difference on 1 of the worksheets.

The second index, R^2, also known as coefficient of determination, shows a percentage of variance shared by 2 or more variables and, as such, is easy to use and interpret. By convention, accounting for less than 2% is considered a small effect; up to 13%, a medium effect; and for 25% or more, a large effect. Researchers can also easily convert between a Cohen's d and R^2 using the following formula: $d=2r/sqrt(1-R^2)$.[37]

SUMMARY

This primer was written to encourage a more rigorous and scholarly approach to SoTL research that uses student-learning assessments. The key message is that researchers should aim to provide evidence to maximize the validity of their conclusions. This evidence should include both psychometrics of the instruments used and practical significance of the test results. By increasing the scientific rigor of educational research and reporting, the overall quality and meaningfulness of SoTL will be improved.

ACKNOWLEDGEMENT

The authors did not have grant support for this manuscript nor do they have any financial conflicts of interest to report. All authors substantially contributed to the writing and editing of this manuscript.

REFERENCES

1. Shea JA, Arnold L, Mann KV. A RIME perspective on the quality and relevance of current and future medical education research. *Acad Med.* 2004;79(10):931-938.
2. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ.* 2002;36(9):800-804.
3. Ratanawongsa N, Thomas PA, Marinopoulos SS, et al. The reported validity and reliability of methods for evaluating continuing medical education: a systematic review. *Acad Med.* 2008;83(3):274-283.
4. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med.* 2005;20(12):1159-1164.
5. Sullivan GM. Deconstructing quality in education research. *J Grad Med Educ.* 2011;3(2):121-124.
6. Sullivan GM, Feinn R. Using effect size- or why the p-value is not enough. *J Grad Med Educ.* 2012;3(3):279-282.
7. Hoover MJ, Jacobs DM, Jung R, Peeters MJ. Validity and reliability with educational testing in the pharmacy and medical education literature. *Am J Pharm Educ.* 2013. In press.
8. Jolly B. The metric of medical education. *Med Educ.* 2002;36(9):798-799.
9. Messick S. Test validity and the ethics of assessment. *Am Psychol.* 1980;35(11):1012-1027.
10. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830-837.
11. Federation of State Medical Boards. Medical Licensing Examination. http://www.fsmb.org/m_usmlestep3.html. Accessed July 19, 2013.
12. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38(9):1006-1012.
13. deVet HCW, Mokkink LB, Terwee CB, Hoekstra OS, Knol DL. Clinicians are right not to like Cohen's kappa. *Br Med J.* 2013;346: f2125.
14. Baumgarter TA. Norm-referenced measurement: reliability. In: Safrit MJ, Wood TM, ed. *Measurement Concepts in Physical Education and Exercise Science.* Champaign, IL: Human Kinetics Publishers; 1989:45-72.
15. deVet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol.* 2006; 59(10):1033-1039.
16. Norcini J. What should we do about unreliable scores? *Med Educ.* 2000;34(7):501-502.
17. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP(UK) examinations. *BMC Med Educ.* 2010;10:40.
18. Harvill LM. An NCME Instructional manual on standard error of measurement. *Educ Meas Issues Pract.* 1991;10(2):33-41.
19. Penny JA, Gordon B. *Assessing Performance.* New York NY. The Guilford Press; 2009.
20. Alston GL, Love BL. Development of a reliable, valid annual skills mastery assessment examination. *Am J Pharm Educ.* 2010; 74(5):Article 80.

21. Case SM, Swanson DB. *Constructing Written Test Questions for the Basic and Clinical Sciences.* 3rd ed. Philadelphia, PA: National Board of Medical Examiners; 2002.
22. Helpful tips for creating reliable and valid classroom tests: evaluating the test [newsletter]. Madison, WI: The Learning Link, University of Wisconsin-Madison Teaching Academy; January 2004. http://testing.wisc.edu/LL01-041.pdf. Accessed July 19, 2013.
23. Hays R, Gupta TS, Veitch J. The practical value of the standard error of measurement in borderline pass/fail decisions. *Med Educ.* 2008; 42(8):810-815.
24. Sturpe DA, Huynh D, Haines ST. Scoring objective structured clinical examinations using video monitors or video recordings. *Am J Pharm Educ.* 2010;74(3):Article 44.
25. Peeters MJ, Sahloff EG, Stone GE. A standardized rubric to evaluate student presentations. *Am J Pharm Educ.* 2010;74(9): Article 171.
26. Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *J Clin Epidemiol.* 2011;64(1):96-106.
27. Peeters MJ, Schmude KA, Steinmiller C. Inter-Rater Reliability and false confidence in precision: using standard error of measurement within PharmD admission essay rubric development. *Curr Pharm Teach Learn.* 2013. In press.
28. Crossley J, Davies H, Humphris G, Jolly B. Generalizability: a key to unlock professional assessment. *Med Educ.* 2002; 36(10): 972-978.
29. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med.* 1998;73(9):993-997.
30. Hodges B, Regehr G, McNaughton N, Tiberius R, Hansen M. OSCE checklists do not capture increasing levels of expertise. *Acad Med.* 1999;74(10):1129-1134.
31. Peeters MJ, Serres M, Gundrum T. Reliability of a residency interview process: reducing the impact from content specificity. *Am J Pharm Educ.* 2013; 77(8):Article 168.
32. Wetzel AP. Factor analysis methods and validity evidence: a review of instrument development across the medical education continuum. *Acad Med.* 2012;87(8):1060-1069.
33. Kelley KA, Stanke LD, Rabi SM, Kuba SE, Janke KK. Cross-validation of an instrument for measuring professionalism behaviors. *Am J Pharm Educ.* 2011;75(9):Article 179.
34. Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Rheum.* 2007;57(8):1358-1362.
35. Norman G. Is experimental research passé. *Adv Health Sci Educ Theory Pract.* 2010;15(3):297-301.
36. Peeters MJ, Kamm GL, Beltyukova SA. A computer-based module for prescribing error instruction. *Am J Pharm Educ.* 2009; 73(6):Article 101.
37. Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. *J Exp Psychol Gen.* 2012; 141(1):2-18.