RESEARCH

Educational Testing Validity and Reliability in Pharmacy and Medical Education Literature

Matthew J. Hoover, PharmD,[a,b,c*] Rose Jung, PharmD, MPH,[c] David M. Jacobs, PharmD,[c,d*] and Michael J. Peeters, PharmD, MEd[c]

[a]College of Pharmacy, Northeast Ohio Medical University, Rootstown, Ohio
[b]Cleveland Clinic Marymount Hospital, Garfield Heights, Ohio
[c]University of Toledo College of Pharmacy and Pharmaceutical Sciences, Toledo, Ohio
[d]University of Buffalo School of Pharmacy and Pharmaceutical Sciences, Buffalo, New York

Submitted May 21, 2013; accepted July 29, 2013; published December 16, 2013.

Objectives. To evaluate and compare the reliability and validity of educational testing reported in pharmacy education journals to medical education literature.

Methods. Descriptions of validity evidence sources (content, construct, criterion, and reliability) were extracted from articles that reported educational testing of learners' knowledge, skills, and/or abilities. Using educational testing, the findings of 108 pharmacy education articles were compared to the findings of 198 medical education articles.

Results. For pharmacy educational testing, 14 articles (13%) reported more than 1 validity evidence source while 83 articles (77%) reported 1 validity evidence source and 11 articles (10%) did not have evidence. Among validity evidence sources, content validity was reported most frequently. Compared with pharmacy education literature, more medical education articles reported both validity and reliability (59%; $p<0.001$).

Conclusion. While there were more scholarship of teaching and learning (SoTL) articles in pharmacy education compared to medical education, validity, and reliability reporting were limited in the pharmacy education literature.

Keywords: validity evidence, educational testing, pharmacy education, medical education

INTRODUCTION

Pharmacy educators use a wide variety of evaluation methods to ascertain whether students achieved specific learning objectives. When developing and evaluating the effectiveness of a doctor of pharmacy (PharmD) curriculum, educators must consider the standards for validity and reliability of educational testing.[1] Standardized tests, such as the Pharmacy College Admission Test and North American Pharmacist Licensure Examination, are used as bookends to assess students' pharmacy-related knowledge and infer competence for licensure.[2,3] Predictive evidence exists for these student performances.[4-7] Educational testing throughout a PharmD program should provide valid and reliable assessment of students' abilities.

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
*Drs. Hoover and Jacobs were first- and second-year residents during the time this study was conducted.

When reporting evaluation methods used in the educational research of health professions, it is essential to consider evidence for validity and reliability. The authors were not aware of any literature reviews assessing the extent of validity and reliability reporting associated with evaluation methods in the pharmacy education literature. The objectives for this study were to characterize reliability and validity with educational testing reported in pharmacy education journals, and compare these with medical education literature reporting.

METHODS

We evaluated validity and reliability reporting in articles that focused on educational testing of learner knowledge, skills, or abilities. To describe levels of reliability and validity reporting associated with pharmacy education literature, articles published in pharmacy education journals were reviewed and the findings were compared to medical education articles. Journals reviewed within pharmacy education were *American Journal of Pharmaceutical Education (AJPE)*, *Currents in Pharmacy*

Teaching and Learning, Pharmacy Education, Annals of Pharmacotherapy, and *American Journal of Health-System Pharmacy*. Journals reviewed within medical education were *Medical Education*, *Academic Medicine*, *Medical Teacher*, *Teaching and Learning in Medicine*, and *Journal of Graduate Medical Education*. Using purposive sampling, we included these journals because they were deemed most likely to include a good cross-section of educational testing.

Within each journal, the table of contents was reviewed for each issue from 2009 to 2012, and 2 reviewers independently identified articles that used educational testing. If an abstract suggested use of educational testing, the reviewer examined the article's full text to ultimately determine eligibility. Examples of educational testing methods included multiple-choice questions, true-false questions, long-answer case notes, and performance-based assessments such as objective structured clinical examinations or clerkship outcomes assessments. Educational testing methods could have been present in the form of examinations, other course work, or even as periodic assessments outside of coursework such as end-of-year or preclinical practice experience examinations. Among the included studies in this analysis, participants were pharmacy and medical learners (students or residents). We included all articles published in the pharmacy education or medical education journals listed above between January 2009 and December 2011; all study designs and countries of origin were reviewed. Articles were excluded if only learner attitudes or opinions were assessed.

Reviewers independently extracted reliability and validity evidence dichotomously (yes/no) for each included study. We used the same definitions for sources of reliability and validity evidence that were used in a prior medical education review.[8] Evidence sources for reliability included test-retest reliability such as reporting a correlation coefficient of scores from tests taken twice over a period of time by learners, a coefficient for internal consistency such as the Cronbach alpha, and inter-rater reliability such as intraclass correlation. Validity evidence required descriptions and intended purpose of the evaluation method and statistical or psychometric testing of learners from at least 1 evidence source among content, construct, or criterion validity. Evidence of content validity was defined as the degree to which an assessment instrument (test or rubric) accurately represented the skills or characteristics it was designed to measure. Reviewers determined evidence of content validity based on description of a study's assessment instrument; ie, whether the content source and assessment format were descriptive enough that a reader could comprehend and replicate this assessment. Evidence of construct validity was defined as the degree to which an instrument's internal structure measured the theoretical construct it was intended to measure. Reviewers sought this evidence from each study's potential use of factor analysis, Rasch analysis, or item analysis. Criterion validity was the degree to which an instrument produced the same result as another accepted or proven external measure or outcome. The reviewers deemed that criterion validity was present if the studied assessment correlated with another external assessment source such as board examination scores or a critical thinking assessment.

For completeness in validity and reliability reporting, an overall rating was designated to articles based on the presence of validity or reliability descriptions. Studies that reported reliability and at least 1 other evidence for validity were considered completely reported. Articles that contained either validity or reliability evidence were considered partially complete while articles without any validity or reliability descriptions were deemed as absent.

The 2 reviewers had excellent agreement on data extraction (K 0.978; 96% positive agreement, 99% negative agreement) with discrepancies resolved by consensus (discussion) between reviewers.

Not all education research is scholarship of teaching and learning (SoTL).[9] Recognizing the importance of SoTL in faculty development of instructors' pedagogical expertise, we also sought to more closely examine SoTL investigators' use of psychometric testing. Using a definition of SoTL authorship, articles that appeared to be reported by a classroom instructor within their course were designated as SoTL articles.[9] These instructors would be less likely to have formal psychometric training before becoming a faculty member, but may have participated in subsequent faculty development training with psychometrics. These SoTL articles were compared between pharmacy and medical education journals using the same article rating noted above.

Another subgroup analysis compared reporting of validity and reliability evidence among *AJPE* article categories (ie, Instructional Design and Assessment, Teachers' Topics, and Research categories). We questioned whether psychometric reporting would be more rigorous in the *Journal's* Research category as opposed to its other (mainly SoTL) categories.

Continuous variables were summarized as median values and ranges while categorical variables were summarized as frequencies and percentages. Comparisons between groups were performed using the chi-square test for categorical data and nonparametric Mann-Whitney U test for continuous data. A P value of less than 0.05 was considered significant. Statistical tests were conducted using SAS, version 9.2 (SAS Institute, Cary, NC).

Table 1. Comparison of Article Ratings by Journal Type

Validity and Reliability Reporting[a]	Medical Education (n=198)	Pharmacy Education (n=108)	*P*
Complete	117 (59)	14 (13)	<0.001[b]
Partially complete	73 (37)	83 (77)	
Absent	8 (4)	11 (10)	

[a] Complete=both validity and reliability described; partially complete=either validity or reliability described; absent=neither validity nor reliability described.
[b] As determined by chi-square test.

RESULTS

Of 2,372 possible articles initially searched, only 306 articles actually used educational testing (198 medical education articles and 108 pharmacy education articles). For the time period, we did not identify any educational testing use in *Annals of Pharmacotherapy* or the *American Journal of Health-Systems Pharmacy*, though examples of education testing use were found in all other searched journals. For extracted studies, we did not find any difference between journal type (pharmacy education vs medical education) and year of article publication (2009, 2010, or 2011; $p=0.30$).

There was a significant difference in complete, partially complete, and absence of validity and reliability descriptions among articles published in pharmacy and medical educational literature ($p<0.001$). Compared with medical education literature, pharmacy education literature appeared to have less complete reporting (59% vs 13%, respectively) and more partially complete reporting (37% vs 77%, respectively) (Table 1). There was also low absent reporting (4% vs 10%) in either medical and pharmacy education literature.

Table 2 shows a comparison between journal types for reliability and validity evidence sources. Evidence for content validity was reported the most in both journal types. However, many pharmacy education articles lacked reliability evidence. Neither construct nor criterion validity were reported often in either journal type.

Sixty-one of 198 medical education articles and 82 of 108 pharmacy education articles were categorized as SoTL (31% vs 76%; $p<0.001$). Table 3 shows a breakdown of those articles for reporting psychometric descriptions. Most of the SoTL articles in pharmacy education came from *AJPE's* Instructional Design and Assessment category. In the *AJPE* subgroup, 98 articles with educational testing were reported. Table 4 shows a breakdown of those articles by *AJPE*'s categories. No difference in reporting validity or reliability was seen among *AJPE*'s categories ($p=0.06$).

DISCUSSION

Pharmacy education authors overall were diligent about describing the content validity of their educational testing and we could visualize the assessment being used. However, reliability was reported less frequently in the pharmacy education literature than in the medical education literature. When we compared journal types, the number of articles with educational testing published each year was not different and did not seem to be a factor.

Pharmacy education journals had more SoTL articles than medical education journals, which we found encouraging. However, validity and reliability descriptions in those pharmacy education articles were reported less, with the lack of reliability reporting being most notable. The *AJPE* subgroup illustrated that validity and reliability reporting were similar among *AJPE* categories. Research category articles were not better reported than Instructional Design and Assessment or Teachers' Topics category articles. The reporting of educational testing psychometrics appears to need improvement across all *AJPE* categories.

Our study did have limitations. Some studies that used educational testing could have been overlooked with the search strategy we used. However, we wanted to focus our efforts in evaluating the pharmacy and medical education articles in journals most widely viewed by educators. We identified a large number of articles in both pharmacy

Table 2. Comparison of Reporting of Validity Evidence Sources by Journal Type

Validity Evidence Source	Medical Education (n=198)	Pharmacy Education (n=108)	*P*
Internal consistency	100 (51)	10 (9)	<0.001[a]
Inter-rater reliability	74 (37)	10 (9)	<0.001[a]
Content validity	185 (93)	90 (85)	0.016[a]
Construct validity	28 (14)	9 (8)	0.14[a]
Criterion validity	22 (11)	6 (6)	0.11[a]

[a] As determined by chi-square test.

Table 3. Comparison of Scholarship of Teaching and Learning Articles

Scholarship of Teaching & Learning	Medical Education (n=198), No. (%)	Pharmacy Education (n=108), No. (%)	*P*
Articles	61 (31)	82 (76)	<0.001[b]
Validity and reliability reporting[a]			
Complete	24 (39)	7 (9)	<0.001[b]
Partially complete	31 (51)	67 (81)	
Absent	6 (10)	8 (10)	

[a] Complete=both validity and reliability described; partially complete=either validity or reliability described; absent=neither validity nor reliability described.
[b] As determined by chi-square test.

and medical education literature; a few more studies would not have changed the conclusion. We did have strong inter-rater reliability in our searching. Also, we originally conducted a pilot study in searching using database keywords. This approach resulted in a low number of articles. We modified this search strategy to include abstracts of selected pharmacy and medical education journals and were able to identify a more substantial sample from the literature. Every test has inherent validity and reliability properties (whether excellent, acceptable, or poor) and absence of reporting does not necessarily imply that authors did not assess these properties before reporting. When acceptable, these properties may simply have been omitted from final reporting. We categorized SoTL based on each article's description suggesting that the authors were instructors evaluating their own classroom activities. This definition had been suggested previously.[9] We may have errored in categorizing a few articles as SoTL as we could not conclusively determine this for every study. Despite these limitations, our study is a useful reflection of recent validity and reliability reporting in the medical and pharmacy literature.

With a growing focus in higher education on student learning, educators are turning to literature for evidence-based teaching methods. They are searching for descriptions of teaching methods and evaluation of those methods. Researchers in this field must recognize the importance of validity and reliability reporting. A short series of articles in the *Journal of Graduate Medical Education* provides some guidance for teaching and learning investigations,[10-12] while a larger *Medical Education* series on assessment practices may give more complex, added insight.[13]

In describing current levels of reporting, we hope to increase awareness of the need for psychometric testing with assessment methods. We have also developed a primer on psychometrics for a pharmacy education readership provide guidance for future authors.[14] A similar followup study in a few years may help to determine if reporting practices have improved. Teaching programs, including resident teaching and learning curricula, may be another avenue for educating academicians of the need to address this important aspect of pharmacy education testing.

CONCLUSION

Most of the pharmacy education articles we reviewed completely or partially reported validity and reliability evidence of educational testing, but reporting was limited compared to medical education articles. While the larger quantity of pharmacy education articles of SoTL was encouraging, reporting of reliability associated with educational testing needs improvement. We encourage efforts to improve reporting of these standards for testing.

Table 4. Comparison of Reporting Validity and Reliability by *American Journal of Pharmaceutical Education* Category

	American Journal of Pharmaceutical Education Category (n=98), No. (%)		
	Research	Instructional Design and Assessment and Teachers' Topics Articles	*P*
Number of reports	23 (23)	75 (77)	<0.001[b]
Validity and reliability reporting[a]			
Complete	6 (26)	8 (11)	0.06[b]
Partially complete	16 (70)	60 (80)	
Absent	1 (4)	7 (9)	

[a] Complete=both validity and reliability described; partially complete=either validity or reliability described; absent=neither validity nor reliability described.
[b] As determined by chi-square test.

ACKNOWLEDGEMENTS

A poster of this study was presented at the 2012 ACCP Annual Meeting in Hollywood, Florida.

REFERENCES

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington DC. American Psychological Association; 1999.
2. Popovich NG, Grieshaber LD, Losey MM, Brown CH. An evaluation of PCAT examination based on academic performance. *Am J Pharm Educ.* 1977;41(2):128-132.
3. Lowenthal W, Wergin JF. Relationships among student pre-admission characteristics, NABPLEX scores, and academic performance during later years in pharmacy school. *Am J Pharm Educ.* 1979;43(1):7-11.
4. Kuncel NR, Crede M, Thomas LL, Klieger DM, Seiler SN, Woo SE. A meta-analysis of the validity of the pharmacy college admission test (PCAT) and grade predictors of pharmacy student performance. *Am J Pharm Educ.* 2005;69(3):Article 51.
5. Meagher DG, Lin A, Stellato CP. A predictive validity study of the pharmacy college admission test. *Am J Pharm Educ.* 2006;70(3): Article 53.
6. Cunny KA, Perri M. Historical perspective on undergraduate pharmacy student admissions - the PCAT. *Am J Pharm Educ.* 1990;54(1):1-6.
7. Newton DW, Boyle M, Catizone CA. The NAPLEX: evolution, purpose, scope, and educational implications. *Am J Pharm Educ.* 2008;72(2):Article 33.
8. 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.
9. Medina M, Hammer D, Rose R, et al. Demonstrating excellence in pharmacy teaching through scholarship. *Curr Pharm Teach Learn.* 2011;3(4):255-259.
10. Sullivan GM. Deconstructing quality in education research. *J Grad Med Educ.* 2011;3(2):121-124.
11. Sullivan GM. A primer on the validity of assessment instruments. *J Grad Med Educ.* 2011;3(2):119-120.
12. Sullivan GM. Writing education studies for publication. *J Grad Med Educ.* 2012;4(2):133-137.
13. Jolly B, Spencer J. The metric of medical education. *Med Educ.* 2002;36(9)798-799.
14. Peeters MJ, Beltyukova SA, Martin BA. Educational testing and validity of conclusions in the scholarship of teaching and learning. *Am J Pharm Educ.* 2013;77(9):Article 186.