Establishing the Validity of Licensing Examination Scores

John R. Boulet, PhD, FSSH

The validation of the scores on licensure/certification examinations or, more appropriately, the pass/fail decisions that are made based on these scores, can be difficult and costly. Nevertheless, evidence must be procured to support any inferences we make based on the assessment scores.[1,2] Without this evidence, various stakeholders, including the examinees, could question the value of the assessment. More important, at least for the health professions, utilizing high-stakes licensure assessments with questionable validity could result in false positive decisions and, ultimately, allow for the provision of care by poorly qualified practitioners.

There are several frameworks that can be referenced to help categorize validity evidence. The argument-based model proposed by Kane is widely accepted and cited.[3,4] The interpretation of a test score rests on a series of assertions and assumptions that support that interpretation. Validity evidence is broken down into 4 categories: *scoring*, *generalization*, *extrapolation*, and *decision/interpretation*. For *scoring*, we are most concerned with evidence that the test was administered properly. For *generalization*, the argument requires evidence that the observations were appropriately sampled from the universe of test items. For *extrapolation*, evidence is needed to support claims that the observations represented by the test score are relevant to the measurement construct of interest. To support the *decision/interpretation* component, evidence to substantiate the theoretical framework necessary for score interpretation or evidence in support of any decision rules (eg, pass/fail status) is needed. Ideally, evidence for all components of the validity argument is desirable.

While the validation of test scores, and associated inferences we make based on the test scores, is an ongoing process, some evidence is more compelling than others. Based on Kane's framework, there is evidence to support the validity of the Comprehensive Osteopathic Medical Licensure Examination (COMLEX-USA) Level 2–Cognitive Examination (CE).[5] It is a computer-based assessment conducted under standardized conditions (*scoring*). The test forms are constructed using a systematic processes, and the sources of measurement error have been identified (*generalization*). Finally, the standards are established through the implementation of a defensible procedure (*decision*). While this evidence helps support the use of Level 2-CE scores as part of the licensure examination sequence for osteopathic physicians, it is far from complete.

In this issue of the *Journal of Graduate Medical Education*, Hudson and colleagues provide data indicating that performance on the Comprehensive Osteopathic Medical Licensure Examination (COMLEX-USA) Level 2–Cognitive Examination (CE) is related to performance on the Comprehensive Osteopathic Medical Achievement Test (COMAT).[6] This finding adds to the existing literature documenting the associations between COMLEX performance and both school-based and residency performance measures.[7–9] Approximately 24% to 46% of the variance of Level 2-CE scores can be explained by the COMAT scores. Similarly, the COMAT clinical subject scores explained 68% of the variance among Level 2-CE scores. Criterion-related evidence is central to the *extrapolation* stage of the validity argument. Hudson et al have provided some evidence to suggest that performance on assessments measuring like constructs are related. However, one could argue that this constitutes "weak" validity evidence. Examinees who test well tend to do well on tests, especially tests administered in the exact same format and with the same (high) stakes. More importantly, COMAT and Level 2-CE are constructed using the same development process and have overlapping content domains. Finally, even though students who had taken Level 2-CE prior to COMAT were excluded from the study, the 2 examinations are generally taken in the same time period. With this in mind, it is not surprising that the scores are related. This argument is not meant to discourage efforts to establish "criterion- or concurrent-related" validity. If the scores for COMAT and Level 2-CE were not related, this would represent a major threat to the validity of one, or both, assessments. Nonetheless, quantifying relationships of Level 2-CE or COMAT scores with other outcome measures is likely to yield more compelling validity evidence.

DOI: http://dx.doi.org/10.4300/JGME-D-19-00611.1

The study by Hudson and colleagues was narrowly focused on COMAT and Level 2-CE and, arguably, the validity of the Level 2-CE scores. When reviewing these types of investigations, one should be asking, "What other sources of validity evidence are needed"? In terms of the osteopathic licensing sequence, one would expect that Level 1, Level 2-CE, and Level 3 scores would be related, more so for examinations taken closer together. Likewise, based on content coverage, Level 1 scores should be, and have been shown to be, related to United States Medical Licensing Examination (USMLE) Step 1 scores.[10] To the extent that the clinical skills examination (Level 2-PE) measures different skills and abilities than the selected response examinations, the associations between the scores from this assessment and the scores from other licensure examinations should be weak. A more comprehensive look at the associations between the licensing examination scores, including any subscores, is certainly warranted. Based on the *extrapolation* argument, there can also be a number of factors that interfere with the assessment of proficiencies of interest. For Level 2-CE, and all other licensing examinations, timing could certainly be an issue. To the extent that the ability to respond quickly is not part of the construct of interest (ie, clinical skills and problem-solving techniques), inferences based on the scores may be error-prone. Testing agencies need to conduct a thorough exploration of potential sources of construct-irrelevant variance. The introduction of any extraneous, uncontrolled variables (eg, poorly constructed examination questions, insecure test questions, examinee testwiseness, examinee guessing, item bias) that affect assessment outcomes can compromise the legitimacy of the decisions made based on examination results.

Some of the most compelling validity evidence, especially for licensure examinations, rests with establishing a link between the scores and "real-world" outcomes of interest.[11,12] While licensing examinations aren't necessarily developed with the expressed purpose of predicting future outcomes, there should be some relationship between the knowledge and skills measured and future performance in practice. Unfortunately, the most persuasive evidence is also the most difficult to secure. Since individuals who do not pass the licensing examinations do not practice, the population for any predictive validity study is homogenized. Also, many of the outcome measures (eg, patient data, disciplinary actions) can be difficult to obtain. Finally, there is always a problem of attribution. To the extent that the practice of medicine is team-based, individual practitioner outcomes, which can also be dependent on a host of environmental factors (eg, hospital size) and the characteristics of the patient population (eg, disease burden), become less useful as indicators of ability. As such, the relationships between licensing examination scores and practice outcomes may be attenuated. However, from a validity standpoint, if the ultimate purpose of COMLEX-USA is to establish competency for initial licensure, and the content of the examinations reflects what physicians should know and be able to do, positive relationships between the scores and "real-world" outcomes are to be expected.

The study by Hudson et al does provides some data to support the validity of Level 2-CE and COMAT scores. Since osteopathic and allopathic graduates can now match into the same residency programs,[13] program directors need to be confident that both COMLEX-USA and the USMLE measure the knowledge, skills, and abilities needed for quality patient care. The National Board of Osteopathic Medical Examiners is encouraged to continue its efforts to collect evidence to support the use of COMLEX-USA scores and the associated medical school graduation and state licensure decisions that are based on these scores.

References

1. Boulet J, van Zanten M. Ensuring high-quality patient care: the role of accreditation, licensure, specialty certification and revalidation in medicine. *Med Educ*. 2014;48(1):75–86. doi:10.1111/medu.12286.
2. Boulet JR, McKinley DW. Criteria for a good assessment. In: McGaghie WC, ed. *International Best Practices for Evaluation in the Health Professions*. London: Radcliffe Publishing, Inc.; 2013:19–43.
3. Kane M. Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Lang Test*. 2012;29(1):3–17. doi:10.1177/0265532211417210.
4. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560–575. doi:10.1111/medu.12678.
5. Gimpel JR, Horber D, Sandella JM, Knebl JA, Thornburg JE. Evidence-based redesign of the COMLEX-USA series. *J Am Osteopath Assoc*. 2017;117(4):253–261. doi:10.7556/jaoa.2017.043.
6. Hudson KM, Tsai T-H, Finch C, Dickerman JL, Liu S, Shen L. A validity study of the COMLEX-USA Level 2-CE and COMAT clinical subjects: concurrent and predictive evidence. *J Grad Med Educ*. 2019;11(5):521–526.
7. Li F, Gimpel JR, Arenson E, Song H, Bates BP, Ludwin F. Relationship between COMLEX-USA scores and performance on the American osteopathic board of

emergency medicine Part I certifying examination. *J Am Osteopath Assoc*. 2014;114(4):260–266. doi:10.7556/jaoa.2014.051.

8. Hudson KM, Feinberg G, Hempstead L, Zipp C, Gimpel JR, Wang Y. Association between performance on COMLEX-USA and the American College of Osteopathic Family Physicians In-Service Examination. *J Grad Med Educ*. 2018;10(5):543–547. doi:10.4300/JGME-D-17-00997.1.
9. Li F, Kalinowski KE, Song H, Bates BP. Relationships between the comprehensive osteopathic medical achievement test (COMAT) subject examinations and the COMLEX-USA Level 2–Cognitive Evaluation. *J Am Osteopath Assoc*. 2014;114(9):714–721. doi:10.7556/jaoa.2014.140.
10. Sandella JM, Gimpel JR, Smith LL, Boulet JR. The use of COMLEX-USA and USMLE for residency applicant selection. *J Grad Med Educ*. 2016;8(3):358–363. doi:10.4300/JGME-D-15-00246.1.
11. Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Acad Med*. 2014;89(8):1157–1162. doi:10.1097/ACM.0000000000000310.
12. Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA*. 2007;298(9):993–1001. doi:10.1001/jama.298.9.993.
13. Buser BR. A single graduate medical education accreditation system: ensuring quality training for physicians and improved health care for the public. *J Am Osteopath Assoc*. 2014;114(4):231–232. doi:10.7556/jaoa.2014.063.

John R. Boulet, PhD, FSSH, is Vice President, Research and Data Resources, Foundation for Advancement of International Medical Education and Research, Educational Commission for Foreign Medical Graduates.

Corresponding author: John R. Boulet, PhD, FSSH, Educational Commission for Foreign Medical Graduates, 3624 Market Street, Philadelphia, PA 19104, 251.823.2227, jboulet@faimer.org