Are United States Medical Licensing Exam Step 1 and 2 Scores Valid Measures for Postgraduate Medical Residency Selection Decisions?

William C. McGaghie, PhD, Elaine R. Cohen, and Diane B. Wayne, MD

Abstract

Purpose

United States Medical Licensing Examination (USMLE) scores are frequently used by residency program directors when evaluating applicants. The objectives of this report are to study the chain of reasoning and evidence that underlies the use of USMLE Step 1 and 2 scores for postgraduate medical resident selection decisions and to evaluate the validity argument about the utility of USMLE scores for this purpose.

Method

This is a research synthesis using the critical review approach. The study first describes the chain of reasoning that underlies a validity argument about using

test scores for a specific purpose. It continues by summarizing correlations of USMLE Step 1 and 2 scores and reliable measures of clinical skill acquisition drawn from nine studies involving 393 medical learners from 2005 to 2010. The integrity of the validity argument about using USMLE Step 1 and 2 scores for postgraduate residency selection decisions is tested.

Results

The research synthesis shows that USMLE Step 1 and 2 scores are not correlated with reliable measures of medical students', residents', and fellows' clinical skill acquisition.

Conclusions

The validity argument about using USMLE Step 1 and 2 scores for postgraduate residency selection decisions is neither structured, coherent, nor evidence based. The USMLE score validity argument breaks down on grounds of extrapolation and decision/interpretation because the scores are not associated with measures of clinical skill acquisition among advanced medical students, residents, and subspecialty fellows. Continued use of USMLE Step 1 and 2 scores for postgraduate medical residency selection decisions is discouraged.

"There is no such thing as a valid test!" assert Clauser and colleagues.[1] These scholars teach that validity is not a property of tests or examinations. Instead, validity is about the accuracy of decisions made from test scores for a focused reason. This rationale comes from advances in test score interpretation

and use based chiefly on the work of Michael Kane.[2–4] Kane presents a framework for test score interpretation that uses an argument-based approach to validity. According to this framework, an argument about the validity of a test score must be structured, coherent, and evidence based. The argument should progress from a test's origins to its administration, scoring, and interpretation. The argument-based approach involves a cascaded chain of reasoning and evidence that leads to claims about test score validity for a specific purpose, in a particular context, with a singular population.

After test design and development, the chain begins with *scoring*, evidence that the test was administered properly and that scores were derived and recorded accurately. The second component, *generalization*, involves evidence about score reliability including item or case sampling, test length, and score precision. The third component, *extrapolation*, requires "evidence that the observations represented by the test score are relevant to the target proficiency or construct measured by the test."[1] Finally, "the

decision/interpretation component of the argument requires evidence in support of any theoretical framework required for score interpretation or evidence in support of decision rules."[1] An argument about the validity of a test score interpretation depends on logically consistent evidence for each of the four components and the integrity of the overall chain of reasoning.

The three-step United States Medical Licensing Examination (USMLE) is a key feature of medical personnel evaluation in North America. The purpose of the USMLE, expressed in the 2010 *Bulletin of Information*, is to provide "individual medical licensing authorities ('state medical boards') … a common evaluation system for applicants for medical licensure."[5] However, since the 1993 inception of the exam, O'Donnell and colleagues[6] acknowledge that USMLE "board scores are often used for *nonlicensure*-related purposes [including] evaluation of examinees' levels of academic achievement, the evaluation of educational programs, and the selection of examinees into residency programs." There are interpretive risks involved in

Dr. McGaghie is Jacob R. Suker, MD, Professor of Medical Education in the Augusta Webster, MD, Office of Medical Education and Faculty Development, and professor of preventive medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

Ms. Cohen is research assistant, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

Dr. Wayne is residency program director and vice chair for education, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois.

Correspondence should be addressed to Dr. McGaghie, Augusta Webster, MD, Office of Medical Education and Faculty Development, Northwestern University Feinberg School of Medicine, 1-003 Ward Building, 303 East Chicago Avenue, Chicago, IL 60611-3008; telephone: (312) 503-0174; fax: (312) 503-0840; e-mail: wcmc@northwestern.edu.

using scores from a test like the USMLE for purposes beyond its pass/fail licensure intent. O'Donnell and colleagues[6] caution, "If the USMLE is to be used for nonlicensure-related decisions, it is important to be able to interpret correctly the scores away from the pass/fail point."

The *Standards for Educational and Psychological Testing*,[7] published by the American Educational Research Association, the American Psychological Association, and the National Council for Measurement in Education, are the "gold standard" regarding the use of test scores for key personnel decisions. These standards have been endorsed by the American Board of Medical Specialties and the National Board of Medical Examiners (NBME). The standards assert that "appropriate use and sound interpretation of test scores … are the responsibility of the test user." Specifically, Standard 1.3 states,

> If validity for some common or likely interpretation [e.g., postgraduate residency selection] has not been investigated, or if the interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be cautioned about making unsupported interpretations.[7]

Despite the assertion that USMLE scores are to be used only for licensure decisions, the Federation of State Medical Boards of the United States and the NBME allow USMLE Part 1 and 2 scores to be used for another nonvalidated purpose—residency application via the Electronic Residency Application Service (ERAS). The 2010 *Bulletin of Information*[5] tells prospective residents, "If you use ERAS, you may request electronic transmittal of your USMLE transcript to residency programs that participate in ERAS."

Residency program directors routinely use USMLE scores in the applicant selection process, despite its licensure intent. To illustrate, Green and colleagues[8] recently reported the results of a national program directors survey on selection criteria for postgraduate residencies. Across all medical specialties, program directors ranked USMLE Step 1 and Step 2 scores second and fifth, respectively, in importance for resident selection. These findings are confirmed by a 2008 survey of residency program directors conducted by the National Resident Matching Program. In this large

sample of almost 2,000 program directors, USMLE Step 1 score was the factor most commonly used when selecting candidates to interview.[9] This assumes a validity argument can be made that links USMLE Step 1 and 2 scores with variables that matter in residency education.[10] Such correlations have been demonstrated in studies involving supervisors' ratings of resident performance as outcome measures, although coefficients are modest and USMLE scores are overinterpreted.[11–13] Research also shows that subjective clinical ratings of trainee performance frequently yield unreliable data that are subject to many sources of bias.[14] A recent systematic review covering the medical education literature from 1955 to 2004 demonstrates that research to verify the presumptive correlation of USMLE scores (or their predecessors) and objective measures of medical trainees' clinical skills has not yet been reported.[15]

Given this history, is there strong validity evidence about using USMLE Step 1 and 2 scores for postgraduate residency selection beyond their licensure intent? Is the validity argument for residency selection structured, coherent, and evidence based?

The objectives of this report are to (1) study the chain of reasoning and evidence that underlies the use of USMLE Step 1 and 2 scores for postgraduate medical resident selection, and (2) evaluate the validity argument about the utility of USMLE scores for resident selection.

Method

This is a research synthesis using the "critical review" approach advocated by Norman and Eva.[16,17] These scholars argue that research reviews should be deliberately selective and critical, not exhaustive. This study extracts and summarizes (1) USMLE Step 1 and 2 scores and (2) reliable clinical performance data drawn from nine research reports published by Northwestern University investigators from 2005 to 2010. These were the only studies found in a search conducted during spring 2010 that assess the correlation between USMLE Step 1 and 2 scores and objective, reliable clinical performance evaluations. Our search strategy covered three literature databases (MEDLINE, Web of Knowledge,

PsychINFO) and employed search terms and concepts (e.g., medical education, residency training, clinical skills, USMLE) and their Boolean combinations. We searched from 1990 to April 2010. We also reviewed reference lists of all selected manuscripts to identify additional reports. The intent was to perform a detailed and thorough search of peer-reviewed publications that have been judged for academic quality to assess the correlation between USMLE scores and clinical performance of advanced medical students and postgraduate trainees.

The research synthesis of the nine reports involves data from 393 medical students and residents across the five-year time span. The majority of participants were enrolled in Northwestern undergraduate and postgraduate training programs. However, nephrology fellows from three metropolitan Chicago programs also participated. The performance data concern clinical skill acquisition by third-year medical students, internal medicine residents, emergency medicine residents, and nephrology fellows. The skills include cardiac auscultation, central venous catheter (CVC) insertion, advanced cardiac life support (ACLS), communication with patients, thoracentesis, and temporary hemodialysis catheter (THDC) insertion. This study is a variation on the theme of secondary data analysis, synthesis, and presentation promoted by research methods scholars.[18]

We extracted and tabulated correlations between USMLE Step 1 and 2 scores and reliable clinical performance scores from the nine research reports. Spearman rho correlations were calculated in each study to evaluate the association of USMLE Step 1 and 2 scores with reliable measures of student, resident, or subspecialty fellow acquisition of key clinical skills. Correlations are reported from the actual data and also corrected for attenuation (unreliability). Reliability coefficients (KR-21, alpha, and kappa) are data quality estimates ranging from 0.00 to 1.00. Reliability values above 0.80 are considered acceptable for research and evaluation. Measures of clinical skills include an audiovisual evaluation of cardiac auscultation[19] and observational checklist evaluations of CVC insertion, ACLS, communication with patients, thoracentesis, and THDC insertion.

Results

Measures of clinical skills were diverse. Cardiac auscultation skills were assessed by the trainee's ability to perform a physical exam and formulate a clinical diagnosis based on findings. ACLS skills were evaluated by participants' team leadership and communication in addition to medical knowledge and patient care regarding basic and advanced patient resuscitation. Communication skills were measured by 14 physician attributes rated by patients. Three skills were predominantly technical (CVC insertion, THDC insertion, thoracentesis). However, these procedural assessments also included components such as history taking, medical decision making, and patient communication.

A summary of the correlations of USMLE Step 1 and 2 scores with reliable measures of clinical skill acquisition among

medical student, resident, and fellow participants from the nine studies is presented in Table 1.[20–28] For USMLE Step 1, the correlations range from −0.05 to 0.29 (median = 0.02); none are statistically significant. For USMLE Step 2, the correlations range from −0.16 to 0.24 (median = 0.18); one is statistically significant, yet accounts for a meager proportion of the variation among the scores ($0.23^2 = 5\%$). When correlations are corrected for attenuation, they range from −0.06 to 0.33 (median = 0.03) for USMLE Step 1. For Step 2, the corrected correlations range from −0.03 to 0.27 (median = 0.22).

Discussion

USMLEs are carefully crafted measures of acquired medical knowledge that are administered and scored under standardized conditions. These characteristics fulfill the *scoring* link in

the validity argument chain. There is evidence[5] that USMLE Step 1 and Step 2 scores are highly reliable to satisfy the *generalization* link in the validity argument chain. However, USMLE Step 1 and Step 2 scores fall short on grounds of *extrapolation* because they lack association with measures of clinical skills that matter among advanced medical students, residents, and subspecialty fellows. The validity argument also breaks down in terms of *decision/interpretation* because the absence of an empirical link between USMLE scores and measured clinical skill acquisition shows that the examination scores do not have clinical correlates. By contrast, there is much evidence from medical education that multiple-choice test scores are correlated strongly with other multiple-choice test scores.[29,30] In this context, high correlations among scores are due to common measurement methods

Table 1

Correlation of United States Medical Licensing Exam (USMLE) Step 1 and 2 Scores With Reliable Measures of Clinical Skills Among Medical Students, Internal Medicine and Emergency Medicine Residents, and Nephrology Fellows (Studies Published 2005–2010)

Trainees	No.	% U.S. medical school graduates	Clinical skills exam	Reliability	Correlations		Correlations corrected for attenuation	
					USMLE 1	USMLE 2	USMLE 1	USMLE 2
Medical students	117	N/A	Cardiac auscultation[20]	0.85 KR-21*	−0.04	N/A	−0.05	N/A
Residents								
Internal medicine		97						
	90		Central venous catheter insertion[21,22]	0.93 kappa	0.08	−0.16	0.11	−0.25
	79		Advanced cardiac life support scenarios[23,24]	0.82 kappa	0.01	0.23†	0.02	0.31
	47		Advanced cardiac life support patient outcomes[25]	0.83 kappa	−0.04	−0.02	−0.05	−0.03
	30		Communication[26]	0.98 alpha	0.03	0.07	0.04	0.09
	40		Thoracentesis[27]	0.94 kappa	−0.05	0.18	−0.06	0.22
Emergency medicine		100						
	12		Central venous catheter insertion[22]	0.93 kappa	0.21	0.24	0.25	0.27
Nephrology fellows		56						
	18		Temporary hemodialysis catheter insertion[28]	0.83 kappa	0.29	0.22	0.33	0.25
Total	393‡							

* Kuder–Richardson 21 reliability coefficient.
† $P < .05$.
‡ Unique participants. Forty residents engaged in multiple studies.

rather than a link to a consistent trait like clinical competence.[29]

Use of USMLE Step 1 and 2 scores for postgraduate resident selection is a decision rule that is not evidence based unless the target outcome is another multiple-choice test. In this case, the validity argument makes sense only if the purpose of resident selection is to choose trainees who achieve high USMLE Step 1 and Step 2 scores and high scores on multiple-choice specialty board examinations. Measures used for resident selection that do not capture "real world" skills needed for clinical practice contribute little to the chain of validation reasoning whose end point is measured clinical competence.[31]

This research synthesis demonstrates that USMLE Step 1 and 2 scores are not correlated with reliable measures of students', residents', and fellows' clinical skill acquisition—cardiac auscultation, CVC insertion, ACLS, communication with patients, thoracentesis, and THDC insertion. These are competencies and skills that matter on clinical and professional grounds. Studying these correlations at multiple levels—students, junior and senior postgraduate trainees, and subspecialty fellows—shows that USMLE scores do not correlate with clinical skills near the time of the examinations or during subsequent clinical training.

The argument that USMLE Step 1 and 2 scores are valid predictors of clinical performance that matters is not sustained by the evidence presented here. Links in the validity argument involving extrapolation and decision/interpretation are not supported by these data, and the integrity of the chain of reasoning is broken. This idea is not new. Scholars have pointed out for at least 20 years that USMLE Step 1 and 2 scores, and their predecessors, are not designed for use in postgraduate resident selection and are not linked with clinical performance.[32,33]

The results of this data synthesis are consistent, but we acknowledge that the number of studies reviewed is small and primarily from trainees at one institution. Also, we reviewed a wide range of skills among medical trainees, yet it is impossible to assess all physician skills for correlations with USMLE Step 1 and 2 scores.

What are alternative approaches to sort and select physicians for competitive postgraduate residency positions? Are there measures that are better linked with clinical competence acquisition? Several studies have been reported that hold promise for improved measurement policy and practice about selecting medical learners. The measures include the University of Michigan's Postgraduate Orientation Assessment,[34] an OSCE for incoming residents; the Israeli MOR (a Hebrew acronym for "selection for medicine"), a simulation-based assessment center for evaluating the personal and interpersonal qualities of medical school candidates[35]; and the multiple mini-interview developed at McMaster University to evaluate medical candidates at undergraduate and postgraduate levels.[36] Each of these measurement procedures relies on practical evaluations of candidates' technical, professional, and interpersonal skills rather than measures of acquired knowledge. Strengths of these studies include assessment of skills needed for actual patient care and use of assessment measures that yield reliable data. This approach measures competence rather than intelligence[37] and is designed to select doctors who will provide high-quality patient care rather than achieve high multiple-choice test scores. Further study is needed to link assessment strategies such these with enhanced residency selection procedures and subsequent trainee performance. It is also necessary to develop and adopt these alternative approaches on a larger scale for availability to program directors nationally.

Conclusions

The USMLE Step 1 and 2 examinations and their scores are designed to contribute to medical licensure decisions. Use of these scores for other purposes, especially postgraduate residency selection, is not grounded in a validity argument that is structured, coherent, and evidence based. Continued use of USMLE Step 1 and 2 scores for postgraduate medical residency selection decisions is discouraged.

References

1 Clauser BE, Margolis MJ, Swanson DB. Issues of validity and reliability for assessments in medical education. In: Holmboe ES, Hawkins RE, eds. Practical Guide to the Evaluation of Clinical Competence. Philadelphia, Pa: Mosby Elsevier; 2008.

2 Kane MT. Validation. In: Brennan RL, ed. Educational Measurement. 4th ed. Westport, Conn: American Council on Education and Praeger Publishers; 2006.

3 Kane MT. An argument-based approach to validity. Psychol Bull. 1992;112:527–535.

4 Kane MT. Content-related validity evidence in test development. In: Downing SM, Haladyna TM, eds. Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates; 2006:131–153.

5 United States Medical Licensing Examination 2010. Bulletin of Information. Philadelphia, Pa: Federation of State Medical Boards of the United States and the National Board of Medical Examiners; 2009.

6 O'Donnell MJ, Obenshain SS, Erdmann JB. Background essential to the proper use of the results of Step One and Step Two of the USMLE. Acad Med. 1993;68:734–739. http://journals.lww.com/academicmedicine/Abstract/1993/10000/Background_essential_to_the_proper_use_of_results.2.aspx. Accessed September 15, 2010.

7 American Educational Research Association; American Psychological Association; National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 1999.

8 Green M, Jones P, Thomas JX Jr. Selection criteria for residency: Results of a national program directors survey. Acad Med. 2009;84:362–367. http://journals.lww.com/academicmedicine/Fulltext/2009/03000/Selection_Criteria_for_Residency__Results_of_a.24.aspx. Accessed September 15, 2010.

9 National Resident Matching Program: Results of the 2008 Program Director Survey. http://www.nrmp.org/data/programresultsbyspecialty.pdf. Accessed September 15, 2010.

10 Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37:830–837.

11 Fine PL, Hayward RA. Do the criteria of resident selection committees predict residents' performances? Acad Med. 1995;70:834–838. http://journals.lww.com/academicmedicine/Abstract/1995/09000/Do_the_Criteria_of_Resident_Selection_Committees.27.aspx. Accessed September 15, 2010.

12 Paolo AM, Bonaminio GA. Measuring outcomes of undergraduate medical education: Residency directors' ratings of first year residents. Acad Med. 2003;78:90–95. http://journals.lww.com/academicmedicine/ Fulltext/2003/01000/Measuring_Outcomes_ of_Undergraduate_Medical.17.aspx. Accessed September 15, 2010.

13 Durning SJ, Pangaro LN, Lawrence LL, et al. The feasibility, reliability, and validity of a program director's (supervisor's) evaluation form for medical school graduates. Acad Med. 2005;80:964–968. http://journals.lww. com/academicmedicine/Fulltext/2005/10000/ The_Feasibility,_Reliability,_and_Validity_ of_a.18.aspx. Accessed September 15, 2010.

14 Williams RG, Klamen DA, McGaghie WC. Cognitive, social, and environmental sources of bias in clinical competence ratings. Teach Learn Med. 2003;15:270–292.

15 Hamdy H, Prasad K, Anderson MB, et al. BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. Med Teach. 2006;28:103–116.

16 Norman G, Eva KW. Quantitative Research Methods in Medical Education. Edinburgh, UK: Association for the Study of Medical Education; 2008.

17 Eva KW. On the limits of systematicity. Med Educ. 2008;42:852–853.

18 Shadish WR, Cook TD, Campbell DT. Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston, Mass: Houghton Mifflin; 2002.

19 The MIAMI Group. UMedic User Manual. Miami, Fla: Gordon Center for Research in Medical Education, University of Miami Miller School of Medicine; 2007.

20 Butter J, McGaghie WC, Cohen ER, Kaye M, Wayne DB. Simulation-based mastery learning improves cardiac auscultation skills in medical students. J Gen Intern Med. 2010; 25:780–785.

21 Barsuk JH, McGaghie WC, Cohen ER, Balachandran JS, Wayne DB. Use of simulation-based mastery learning to improve the quality of central venous catheter placement in a medical intensive care unit. J Hosp Med. 2009;4:397–403.

22 Barsuk JH, McGaghie WC, Cohen ER, O'Leary KS, Wayne DB. Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. Crit Care Med. 2009;37:2697–2701.

23 Wayne DB, Butter J, Siddall VJ, et al. Simulation-based training of internal medicine residents in advanced cardiac life support protocols: A randomized trial. Teach Learn Med. 2005;17:210–216.

24 Wayne DB, Butter J, Siddall VJ, et al. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. J Gen Intern Med. 2006;21:251–256.

25 Wayne DB, Didwania A, Cohen ER, Schroedl C, McGaghie WC. Improving the quality of cardiac arrest medical team responses at an academic teaching hospital. Am J Respir Crit Care Med. 2010;181:A1453.

26 Makoul G, Krupat E, Chang C-H. Measuring patient views of physician communication skills: Development and testing of the Communication Assessment Tool. Patient Educ Couns. 2007;67:333–342.

27 Wayne DB, Barsuk JH, O'Leary KS, Fudala MJ, McGaghie WC. Mastery learning of thoracentesis skills by internal medicine residents using simulation technology and deliberate practice. J Hosp Med. 2008;3: 48–54.

28 Barsuk JH, Ahya SN, Cohen ER, McGaghie WC, Wayne DB. Mastery learning of temporary hemodialysis catheter insertion skills by nephrology fellows using simulation technology and deliberate practice. Am J Kidney Dis. 2009;54:70–76.

29 Forsythe GB, McGaghie WC, Friedman CP. Construct validity of medical clinical competence measures: A multitrait-multimethod matrix study using confirmatory factor analysis. Am Educ Res J. 1986;23:315–336.

30 Perez JA, Greer S. Correlation of United States Medical Licensing Examination and internal medicine in-training examination performance. Adv Health Sci Educ Theory Pract. 2009;14:753–758.

31 McMahon GT, Tallia AF. Perspective: Anticipating the challenges of reforming the United States Medical Licensing Examination. Acad Med. 2010;85:453–456. http://journals. lww.com/academicmedicine/Abstract/2010/ 03000/Perspective__Anticipating_the_ Challenges_of.18.aspx. Accessed September 15, 2010.

32 Nungester RJ, Dawson-Saunders B, Kelley PR, Volle RL. Score reporting on NBME examinations. Acad Med. 1990;65:723–729. http://journals.lww.com/academicmedicine/ Abstract/1990/12000/Score_reporting_on_ NBME_examinations.2.aspx. Accessed September 15, 2010.

33 Berner ES, Brooks CM, Erdmann JB. Use of the USMLE to select residents. Acad Med. 1993;68:753–759. http://journals.lww.com/ academicmedicine/Abstract/1993/10000/Use_ of_the_USMLE_to_select_residents.5.aspx. Accessed September 15, 2010.

34 Lypson ML, Frohna JG, Gruppen LD, Woolliscroft JO. Assessing residents' competencies at baseline: Identifying the gaps. Acad Med. 2004;79:564–570. http:// journals.lww.com/academicmedicine/Fulltext/ 2004/06000/Assessing_Residents__ Competencies_at_Baseline_.13.aspx. Accessed September 15, 2010.

35 Ziv A, Rubin O, Moshinsky A, et al. MOR: A simulation-based assessment centre for evaluating the personal and interpersonal qualities of medical school candidates. Med Educ. 2008;42:991–998.

36 Eva KW, Reiter HI, Trinh K, et al. Predictive validity of the multiple mini-interview for selecting medical trainees. Med Educ. 2009; 43:767–775.

37 McClelland DC. Testing for competence rather than for "intelligence." Am Psychol. 1973;28:1–14.