Stepping Back: Re-evaluating the Use of the Numeric Score in USMLE Examinations

Paul George¹ · Sally Santen² · Maya Hammoud³ · Susan Skochelak⁴

Published online: 8 January 2020

© International Association of Medical Science Educators 2020

Abstract

There are increasing concerns from medical educators about students' over-emphasis on preparing for a high-stakes licensing examination during medical school, especially the US Medical Licensing Examination (USMLE) Step 1. Residency program directors' use of the numeric score (otherwise known as the three-digit score) on Step 1 to screen and select applicants drive these concerns. Since the USMLE was not designed as a residency selection tool, the use of numeric scores for this purpose is often referred to as a secondary and unintended use of the USMLE score. Educators and students are concerned about USMLE's potentially negative influence on curricular innovation and the role of high-stakes examinations in student and trainee well-being. Changing the score reporting of the examinations from a numeric score to pass/fail has been suggested by some. This commentary first reviews the primary use and secondary uses of the USMLE scores. We then focus on the advantages and disadvantages of the currently reported numeric score using Messick's conceptualization of construct validity as our framework. Finally, we propose a path forward to design a comprehensive, more holistic review of residency candidates.

Commentary

There are increasing concerns from US medical educators about students' over-emphasis on preparing for a high-stakes licensing examination during medical school, especially the US Medical Licensing Examination (USMLE) Step 1, typically taken after foundational curricula, and USMLE Step 2 clinical knowledge (CK), typically taken during the 4th year. These examinations are scored using a three-digit score typically in a range from 140 to 260, with higher scores representing a stronger examination performance [1]. Each examination has a specific cut-off score that is used for the pass mark. The primary goal is to pass the examination.

However, three-digit numeric score is used by program directors for residency selection. Student emphasis has its primary basis in how residency program directors use USMLE numeric scores for screening and selecting applicants. Since the USMLE was designed as a tool for licensing and not as a residency selection tool, the use of scores for this purpose is often referred to as a secondary and unintended use of the USMLE numeric score. Medical students who desire to compete successfully for residency programs may not be focused on merely passing USMLE examinations, but rather on obtaining a maximum score. As a result, other important issues related to the current use of USMLE scores have developed. These include USMLE's potentially negative influence on curricular innovation and the role of high-stakes examinations in student and trainee well-being [2]. While we focus on the relation of the USMLE examinations to allopathic medical students in this commentary, our arguments extend to osteopathic students (and the use of COMLEX examinations).

As mentioned previously, examination performance reports on the computer-based testing components of USMLE is reported on a numeric score scale. The pass/fail standards for all USMLE Steps are reviewed every three to four years by a USMLE governance committee made up of representatives from medical school faculty, state board members, public members, and trainees (residents/fellows). While some state boards review the USMLE numeric score in the context of

✉ Paul George
Paul_George@brown.edu

¹ Warren Alpert Medical School of Brown University, 222 Richmond Street, Providence, RI 02912, USA

² Virginia Commonwealth University School of Medicine, 1201 East Marshal Street, Box 980565, Richmond, VA 23298, USA

³ University of Michigan Medical School, 1540 E Hospital Dr, SPC 4276, Ann Arbor, MI 48109-4276, USA

⁴ American Medical Association, 330 N. Wabash-43rd Floor, Chicago, IL 60611-5885, USA

licensure or post-licensure decisions, an applicant's pass/fail performance is the primary information state boards require.

Over the past decade, with movements of medical school curricula to pass/fail grading systems and with increased competition leading to many applications for some residency positions, the emphasis on maximizing USMLE numeric scores has grown, driven by stakeholders, including students and graduate medical education (GME) program directors. Most believe that an over-emphasis on any one or two assessments or examinations is inconsistent with the broad and complex competencies that predict performance in residency. Despite this, the continued use of USMLE numeric scores in residency screening and selection speaks to the perceived lack of other meaningful assessments to inform the undergraduate medical education (UME) to GME transition. Residency program directors often cite conflicts of interest or lack of trust about medical school evaluations given the desire for a school to have all their students match in the best possible programs. Clinical clerkship grades are often inflated or measure constructs unrelated to actual performance [3]. Due to the absence of other assessments that better inform the UME to GME transition, it seems likely that use of USMLE numeric scores will continue. Alternatively, if current residency screening and selection processes do not include USMLE numeric scores, it seems possible that decisions about applicants may be made based on measures that are less reliable or perhaps biased and unfair.

As a result, and to improve a stressed residency selection system which does not meet stakeholders' needs, there is an active debate focused on USMLE numeric score reporting. While many options are theoretically possible, this debate generally involves a numeric score versus pass/fail reporting. In the remainder of this commentary, we will outline some of the current drivers, pros, and cons of the use of the USMLE examinations and the numeric (three-digit) score as it pertains to use these scores for secondary uses.

As assessments of medical knowledge and skills prior to entering residency, USMLE Step 1 and Step 2 CK have strengths. Medical school faculty physicians and practicing physicians develop the exam which measures constructs relevant to medical education, specifically many of the Accreditation Council on Graduate Medical Education (ACGME) competencies. The rigorous test development process and the reliability of the exam, (which provides structural evidence within Messick's conceptualization of construct validity) [4] of the examinations to classify examinees as passing or failing have made USMLE a standard among medical licensing examinations worldwide. For the purpose of licensing, the pass-fail nature of the examination is fit-for-purpose of testing medical knowledge for licensing purposes, as it is used along with other metrics required for licensing. There is also evidence, albeit limited, that incremental performance on medical licensing examinations, such as USMLE, predicts

outcomes valued by GME directors (USMLE specifically specialty board certification performance but not residency performance), state medical boards (USMLE predicts adverse board actions but not actual practice performance), and the public [5, 6].

Proponents of scored USMLE numeric scoring also contend that, for US allopathic graduates and international medical graduates (IMG) training at disparate institutions, there exists no other assessment that allows a standard comparison among physicians who are entering supervised practice. Scored USMLE performance is felt by many to provide a single comparable metric across applicants, however limited. From the perspective of an established medical school, whose graduates have a favorable or historic reputation, this may not be important. Students attending newer medical schools or schools with less reputation, particularly international schools, often favor USMLE numeric scoring as a way to be considered in the residency selection process [7].

The disadvantages to numerically scored USMLE results are notable as well. Framed within Messick's conceptualization of construct validity, the USMLE exams may not meet a validity argument for secondary uses. More US medical schools are opting for an abbreviated foundational curriculum, with earlier and more extensive clinical training. In addition, some schools are moving Step 1 to after the clerkship year, where traditionally medical students take Step 2 CK [8]. The question becomes whether two high stakes examinations are needed at the same testing point—after the clerkship year and if so, should those two examinations be better integrated. The construct tested on Step 1 may not be as relevant to the ultimate role of the physician and inconsistent with content evidence within Messick's framework. For example, the level of biochemical pathways tested on USMLE Step 1 do not apply to the vast majority of physicians' clinical practice. Similarly, critical skills such as communicating with colleagues, nurses, and consultants are not measured.

The explicit and hidden costs for the licensing examinations, including Step 1 cannot be ignored. There is a registration cost associated with the examinations; there are additional, non-insignificant costs for third-party review material, which students use to study for Step 1. There is also the time medical students take away from medical school curriculum to study for the examination. Most importantly, and as mentioned previously, the examination is being used as a convenient measure to screen candidates for residency programs. And, because it is used for these screens, the stakes of doing well on Step 1 have increased significantly, creating the potential for anxiety, depression, and burnout in medical students. Demographic group performance differences exist in USMLE as they do in many other standardized examinations. Thus, any overemphasis on a numeric score as necessary metric for residency selection could limit the desired diversity goals of a residency training program [9].

As mentioned previously, the Step examinations are at best an imperfect measure for predicting physician success in residency, again raising concerns about Messick's conceptualization of construct validity, this time in regard to external evidence, in which the exam does not have clear predictive qualities for a physician's future practice [10]. A recent study concludes, "Although [the] United States Medical Licensing Examination (USMLE) Step 1 is often used as a comparative factor, most studies do not demonstrate its predictive value for resident performance, except in the case of test failure" [10]. Finally, while the USMLE Step exams were never meant to be a high-stakes measure of an applicant's suitability, for residency—they are. Thus, the consequences of not passing this exam (i.e., not matching into a residency program or not being able to practice as a physician) do not meet the consequential evidence in Messick's framework.

In summary, we believe the pass-fail scoring is fit-for-purpose as a part of licensure, but the numeric score is not-fit-for-purpose and may not have adequate validity evidence for residency selection. Noting that the Step examinations do not hold true to Messick's conceptualization of construct validity, we recommend other methodologies to screen candidates for residency positions. This, however, will not be a topic of widespread agreement. We suggest that residency program directors, in conjunction with leaders from UME and other key stakeholders, design a comprehensive, more holistic review of residency candidates, which would include passing Step 1, Step 2 CK, and Step 2 clinical skills but would not be the sole measures of selection for interviews. This holistic review should involve a newly designed, rigorous assessment of a candidate's suitability for a GME program involving variables, such as medical school academic performance, volunteer activities, and research participation that meet Messick's unified theory of validity. As such, we are heartened that this conversation is happening and key stakeholders are part of the conversation. National organizations involved in medical education met at an invitational meeting around many of the issues outlined in this commentary in spring of 2019 [11]. We believe this is a first key step in developing a comprehensive plan to address the issues surrounding the USMLE Step examinations, while at the same time preserving the critical advantages of using this examination as part of a larger process in ensuring the safe, effective care of patients.

Compliance with Ethical Standards

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Approval Not applicable.

Informed Consent Not applicable.

References

1. National Resident Matching Program, Data Release, and Research Committee: Results of the 2018 NRMP program director survey. National Resident Matching Program, Washington, DC. 2018. <https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf>.
2. Moynahan KF. The current use of United States medical licensing examination step 1 scores: holistic admissions and student well-being are in the balance. *Acad Med*. 2018;93(7):963–5.
3. Zaidi NLB, Kreiter CD, Castaneda PR, Schiller JH, Yang J, Grum CM, et al. Generalizability of competency assessment scores across and within clerkships: how students, assessors, and clerkships matter. *Acad Med*. 2018;93(8):1212–7.
4. Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Meas Issues Pract*. 1995;14(4):5–8. <https://doi.org/10.1111/j.1745-3992.1995.tb00881.x>.
5. 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):1157–62.
6. Cuddy MM, Young A, Gelman A, Swanson DB, Johnson DA, Dillon GF, et al. Between USMLE performance and disciplinary action in practice: a validity study of score inferences from a licensure examination. *Acad Med*. 2017;92(12):1780–5.
7. Lewis CE, Hiatt JR, Wilkerson L, Tillou A, Parker NH, Hines OJ. Numerical versus pass/fail scoring on the USMLE: what do medical students and residents want and why? *J Grad Med Educ*. 2011;3(1):59–66.
8. Jurich D, Daniel M, Paniagua M, Fleming A, Harnik V, Pock A, et al. Moving the United States Medical Licensing Examination Step 1 after core clerkships: an outcomes analysis. *Acad Med*. 2019;94:371–7.
9. Rubright JD, Jodoin M, Barone MA. Examining demographics, prior academic performance, and the United States Medical Licensing Examination Scores. *Acad Med*. 2018. DOI <https://doi.org/10.1097/ACM.00000000000002366>.
10. Hartman ND, Lefebvre CW, Manthey DE. A narrative review of the evidence supporting factors used by residency program directors to select applicants for interviews. *J Grad Med Educ*. 2019;11:268–73.
11. <https://www.usmle.org/inCus/>. Accessed July 9, 2019.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.