Farewell to the Step 2 Clinical Skills Exam: New Opportunities, Obligations, and Next Steps
Rachel Yudkowsky, MD, MHPE, and Karen Szauter, MD

Abstract
The unexpected discontinuation of the United States Medical Licensing Examination Step 2 Clinical Skills (CS) exam in January 2021 carries both risks and opportunities for medical education in the United States. Step 2 CS had far-reaching effects on medical school curricula and school-based clinical skills assessments. Absent the need to prepare students for this high-stakes exam, will the rigor of foundational clinical skills instruction and assessment remain a priority at medical schools? In this article, the authors consider the potential losses and gains from the elimination of Step 2 CS and explore opportunities to expand local summative assessments beyond the narrow bounds of Step 2 CS.

The responsibility for implementing a rigorous and credible summative assessment of clinical skills that are critical for patient safety as medical students transition to residency now lies squarely with medical schools. Robust human simulation (standardized patient) programs, including regional and virtual simulation consortia, can provide infrastructure and expertise for innovative and creative local assessments to meet this need. Novel applications of human simulation and traditional formative assessment methods, such as workplace-based assessments and virtual patients, can contribute to defensible summative decisions about medical students’ clinical skills. The need to establish validity evidence for decisions based on these novel assessment methods comprises a timely and relevant focus for medical education research.

The abrupt suspension of the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills (CS) exam in March 2020 due to the COVID-19 pandemic was an important safety measure for medical students, simulated patients, and staff at testing centers.¹ The January 2021 announcement of the decision to fully discontinue this licensing requirement² represents a notable change in the way medical students will be assessed at the national level. This change necessitates a thoughtful dialog within the medical education community to ensure that the focus on, and quality of, teaching and assessing clinical skills does not diminish.

Step 2 CS was introduced in 2004 as a licensure requirement for students graduating from U.S. medical schools amid concerns that graduating students lacked key clinical skills. The exam was based on direct interactions with standardized (simulated) patients (SPs); it assessed students’ ability to gather relevant patient data through a medical interview and physical exam and to organize that information into a patient note, including evidence of clinical reasoning (i.e., differential diagnosis and plan). The exam scoring rubric separated the data gathering and interpretation content of the encounter from the communication and interpersonal skills students displayed while interacting with the SP.³ Step 2 CS was closely studied, and modifications to the encounter challenges and scoring processes evolved over time.⁴

The introduction of this new licensure requirement prompted medical schools to refocus attention on the instruction and assessment of students’ clinical skills and spurred the development and refinement of clinical testing at the medical school level.⁵ Ecker and colleagues⁶ reviewed the beneficial consequences of Step 2 CS in detail and offered suggestions for its enhancement. Although concerns were raised about the exam, especially regarding the cost, stress, and inconvenience to students, the assessment of students’ clinical skills through direct observation of their encounters with an SP contributed a critical component of licensure that was not otherwise attainable through written multiple-choice exams.⁷

How should medical schools respond to the elimination of Step 2 CS? As we ponder next steps, it is important to consider the potential consequences of the exam’s cancelation, both negative and positive, as well as the resulting responsibilities that now devolve to our schools.

What Has Been Lost?
The elimination of Step 2 CS results in both losses and opportunities. The primary loss is of a rigorous national-level standardized screening process to determine minimal competency in graduating medical students’ key clinical skills, including data gathering, data interpretation, and communication. Local, medical school-based assessments are highly vulnerable to validity threats.⁸ Construct underrepresentation can result from common practices such as using inappropriate assessment methods (e.g., written tests to assess physical exam skills) or insufficient and unsystematic sampling (e.g., drawing inferences from too few SP encounters). Construct irrelevant variance can be due to challenges such as poorly written cases and checklists, inadequate standardization of SP case portrayal, insufficient rater training, and rater biases and errors. In addition, local pass/fail standards are variable and may be adjusted post hoc. Medical schools and faculty are understandably reluctant to downcheck or fail their own students.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Rachel Yudkowsky, University of Illinois at Chicago College of Medicine, 986 CMET MC 591, 808 South Wood St., Chicago, IL 60612; telephone: (312) 996-3598; email: rachely@uic.edu.

First published online June 15, 2021
doi: 10.1097/ACM.0000000000004209
Copyright © 2021 by the Association of American Medical Colleges
leading (among other reasons) to persistent failure to fail. On the other hand, the Step 2 CS failure rate in 2018–2019 was 5% for first-time students from U.S. MD-granting medical schools, 15% for students from DO-granting schools, and 23% for international medical graduates. The inability to screen out those students who lack basic clinical skills comprises a threat to patient safety during residency and beyond.

Also threatened by the elimination of Step 2 CS are the positive byproducts of a national high-stakes assessment. Step 2 CS motivated curriculum deans, faculty, and students to devote time, effort, and resources to clinical skills development, as evidenced by the growth in doctoring courses in the past 20 years. Assessment drives the curriculum; will medical schools now infer that it is no longer important to rigorously teach and assess history-taking, physical exam, and communication skills? Will students conclude that it is no longer necessary to devote time and effort to learning these skills? Medical schools have committed substantial resources to human simulation (SP) programs to support doctoring courses and provide a practice Step 2 CS exam for students. Once available, SP programs frequently provide additional curricular programming, including activities that promote the acquisition of advanced skills such as difficult conversations, clinical reasoning, high-value care, and telehealth. Absent the need to prepare for a national high-stakes clinical skills exam, will budget support for these programs and their attendant curricular benefits continue?

What Might Be Gained?

The cancelation of Step 2 CS clearly reduces the financial burden on medical students and the stress of preparing for, traveling to, and participating in the exam. There may be curricular benefits to the exam’s elimination as well. Many medical schools patterned their graduation competency clinical skills exams on Step 2 CS to familiarize their students with the USMLE format. Eliminating the need to practice for a high-stakes licensing exam affords the opportunity for local clinical skills assessments to grow beyond the narrow scope and format of Step 2 CS, which focused only on a narrow subset of the clinical skills needed by interns: the ability to communicate with a patient and to gather and interpret data toward making an initial diagnosis of a new patient. SP encounters can be leveraged to assess other types of patient interactions and entrustable professional activities such as obtaining informed consent, transitioning care to another provider, and recognizing a patient requiring urgent care. Medical schools can experiment with different types of postencounter notes and postencounter challenges such as allowing students to access resources used during clinical practice. Other formats can be explored, including group, longitudinal, and interprofessional encounters. Any of these options would serve to broaden the exam and enhance its content validity. Step 2 CS provided only a summative pass/fail decision; local assessments can provide much more detailed and informative feedback to guide future learning.

Meeting the Need

With the loss of a national-level exam, the responsibility for high-quality summative assessment of clinical skills must revert to medical schools. Patients, their families, and the licensing bodies sworn to protect them must be assured that graduating students can function safely and effectively to provide clinical care under appropriate supervision. Residency programs need to know that interns are prepared for their responsibilities and which interns need additional coaching and supervision (previously signaled by a history of having failed Step 2 CS, especially if repeatedly). Students want to know that they are well prepared to assume a higher level of independent practice. Finally, medical school faculty, administrators, and accrediting bodies need to confirm that their curricula are indeed producing safe and capable physicians.

In meeting these needs, it is critical to identify assessment objectives that can be satisfied only through an encounter with a patient and those that can be accomplished through other means. For example, when assessing data gathering and interpretation in the context of a new patient, some clinical reasoning tasks (e.g., knowing what history and physical exam information to obtain, developing a differential diagnosis based on the patient, and identifying the appropriate diagnostic tests to be done immediately) might be assessed through well-constructed written tests, Key Feature exams, or virtual patient programs. Patient encounters are uniquely suited to elicit and assess behaviors such as (1) whether the student is able to perform a physical exam and identify physical exam findings; (2) whether the student is sufficiently proficient to appreciate and interpret salient history and physical exam findings during the course of the encounter (i.e., go beyond rote questioning to obtain the specific data needed to support a working differential diagnosis); (3) whether the student has the marginal capacity to relate to the patient effectively and empathically while obtaining the history and physical exam, rather than focusing their attention on thinking of what question to ask next; and (4) whether the student is able to sensitively modify their communication and interpersonal behavior to accommodate and respond to diverse patient communication styles, personalities, emotions, and needs.

The first of these goals can be accomplished only through a live, face-to-face, hands-on patient encounter. The second could potentially be accomplished through a sophisticated virtual patient program, although these require much study before being deployed for high-stakes decisions. The third and fourth require a live patient encounter but likely could be accomplished remotely. Focusing on the unique affordances of live SP encounters and designing assessment tools that target these affordances will allow this resource-intensive assessment modality to be leveraged most efficiently and effectively.

Future Directions

The loss of Step 2 CS generates opportunities as well as challenges for medical schools and faculty. There now exists a unique impetus to broaden the scope of clinical skills exams and expand the toolbox of available assessment methods. It is imperative that we do so in a way that promotes the validity of high-stakes decisions and confidence in their results. This demands, and is an opportunity for, a concerted program of educational research. Table 1 identifies some next steps and related research agendas.
The USMLE has committed to “partner with the medical education and medical board community to better develop innovative ways to assess the breadth of clinical skills in medicine.” The urgent need for medical schools to assess students’ clinical skills rigorously and affordably should spur local and national efforts to study and improve the validity and utility of a variety of assessment modalities; these could be used in conjunction with face-to-face human simulation to enrich understanding of students’ development. Workplace-based assessments could be leveraged to provide information about students’ physical exam skills, communication skills, and professionalism, and screen-based simulations such as virtual patients could provide valuable insights about students’ clinical reasoning skills. The use of telesimulation (i.e., SP encounters conducted remotely) and simulated telehealth encounters has exploded during the COVID-19 pandemic; these could provide an additional source of ratings and feedback for students from the crucial perspective of the patient and enable more frequent and rigorous assessment.

### Table 1

<table>
<thead>
<tr>
<th>What is needed</th>
<th>Next steps</th>
<th>Related research topics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>To replace the loss of a standardized national exam</strong></td>
<td>• Mandate by the LCME and/or state licensing bodies</td>
<td>• What do various stakeholders think about such a mandate?</td>
</tr>
<tr>
<td></td>
<td>• Faculty and staff development in assessment design</td>
<td>• Which best practices and quality assurance processes are transferable from human simulation to other performance assessment methods?</td>
</tr>
<tr>
<td></td>
<td>• Identification of robust quality assurance practices to identify and mitigate threats to validity</td>
<td>• What are the benefits and challenges of assessment consortia?</td>
</tr>
<tr>
<td></td>
<td>• Development of national standards of best practice</td>
<td></td>
</tr>
</tbody>
</table>

| **To leverage the opportunity for increased flexibility in clinical skills assessment** | • Continued adaptation of telesimulation, WBAs, VPs, and other methods to be used in high-stakes assessments | • What is the validity evidence for these assessment methods? |
| | • What are the main threats to their validity? How can these threats be mitigated? | • What are the educational and personal consequences of new assessment methods? |
| | • What are the main threats to their validity? How can these threats be mitigated? | • How can available VP data be mined to provide additional insights about performance? |
| | • What are the educational and personal consequences of new assessment formats? | |
| | • What are the key characteristics, validity threats, and validity evidence for new formats? | |
| | • What are the educational and personal consequences of new assessment content? | |
| | • How should data be aggregated across domains and formats toward high-stakes decisions? | |
| | • What are the educational and personal consequences of including new assessment content? | |
| | • What design principles optimize transfer of learning from the simulation lab to clinical settings? | |
| | • What are the essential features of systems to monitor longitudinal skills development? | |
| | • What types of assessment feedback best support the development and maintenance of skills? | |

Abbreviations: USMLE, United States Medical Licensing Examination; LCME, Liaison Committee on Medical Education; WBA, workplace-based assessment; VP, virtual patient; SP, standardized patient; EPA, entrustable professional activity.
of students who are located far from the main campus of the medical school. These modalities are currently used primarily for low-stakes formative feedback and skills development. Effective use of these and other promising methods for the purpose of high-stakes assessment demands a rigorous analysis of their strengths and limitations, similar to the intensive study of human simulation conducted when SPs were first introduced. Because no single assessment method is perfect, a robust understanding of how to qualitatively and quantitatively aggregate longitudinal and episodic information from different modalities toward a final high-stakes summative decision is vital. Here too, it is essential to differentiate between aggregating data for the purpose of low-to-medium-stakes formative assessments (e.g., semiannual evaluations) and the evidentiary requirements of a final, high-stakes summative decision that a student is indeed ready to function safely as an intern.

Far from becoming redundant within such an assessment system, the expertise offered by mature SP programs becomes more valuable than ever. Rigorous, repeated, and systematic assessment through direct observation of students’ clinical skills in a controlled environment comprises a highly reliable source of information about what students can (and cannot) do in a clinical setting. However, conducting local assessments of clinical skills with the rigor to support high-stakes decisions is a challenging task. Educators need robust skills in writing SP cases, blueprinting exams, standardizing SP portrayals, training SP and/or faculty raters, maintaining quality assurance standards, monitoring and minimizing threats to exam validity, and providing actionable reports to students and faculty. Standards of Best Practice for human simulation, such as those promulgated by the Association of Standardized Patient Educators, are key. The development of regional consortia, such as the Mid-Atlantic Consortium and the California Consortium for the Assessment of Clinical Competence, may assist in leveraging expertise and resources across medical schools. The use of telesimulation can enable trained SPs and SP training expertise to be shared across schools while increasing the diversity of available SPs. The use of remote observers and raters, based in consortium partner schools, may facilitate the development of rater expertise and help mitigate local pressures contributing to the “failure to fail.”

The demise of Step 2 CS carries both great risks and exciting opportunities for medical education in the United States. A rigorous summative assessment of clinical skills is critical for patient safety as medical students transition to residency. A strong mandate from the Liaison Committee on Medical Education and licensing bodies would affirm the value of clinical skills curricula and confirm that the responsibility for a credible summative assessment of clinical skills now lies squarely with medical schools. Robust human simulation (SP) programs, including regional and virtual consortia, can provide infrastructure and expertise for innovative and creative local assessments to meet this need. Novel applications of human simulation and traditional formative assessment methods can contribute to defensible summative decisions. The need to establish validity evidence for decisions based on these novel assessment methods creates a powerful opportunity for medical education research and scholarship, and we look forward to seeing the results of such studies presented in these pages.

Funding/Support: None reported.
Other disclosures: None reported.
Ethical approval: Reported as not applicable.

R. Yudkowsky is professor and director of graduate studies, Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, Illinois; ORCID: https://orcid.org/0000-0002-2145-7982.

K. Szauter is assistant dean of educational affairs, University of Texas Medical Branch, Galveston, Texas; ORCID: https://orcid.org/0000-0002-2064-3535.

References


