Validity Evidence for the Emergency Medicine Standardized Letter of Evaluation

Paul Kukulski, MD
James Ahn, MD, MHPE

ABSTRACT

Background The standardized letter of evaluation (SLOE) is the application component that program directors value most when evaluating candidates to interview and rank for emergency medicine (EM) residency. Given its successful implementation, other specialties, including otolaryngology, dermatology, and orthopedics, have adopted similar SLOEs of their own, and more specialties are considering creating one. Unfortunately, for such a significant assessment tool, no study to date has comprehensively examined the validity evidence for the EM SLOE.

Objective We summarized the published evidence for validity for the EM SLOE using Messick's framework for validity evidence.

Methods A scoping review of the validity evidence of the EM SLOE was performed in 2020. A scoping review was chosen to identify gaps and future directions, and because the heterogeneity of the literature makes a systematic review difficult. Included articles were assigned to an aspect of Messick's framework and determined to provide evidence for or against validity.

Results There have been 22 articles published relating to validity evidence for the EM SLOE. There is evidence for content validity; however, there is a lack of evidence for internal structure, relation to other variables, and consequences. Additionally, the literature regarding response process demonstrates evidence against validity.

Conclusions Overall, there is little published evidence in support of validity for the EM SLOE. Stakeholders need to consider changing the ranking system, improving standardization of clerkships, and further studying relation to other variables to improve validity. This will be important across GME as more specialties adopt a standardized letter.

Introduction

The standardized letter of evaluation (SLOE) was developed by a Council of Emergency Medicine Program Directors (CORD) task force in 1995 for use in medical students' applications to emergency medicine (EM) residency.[1] In the 23 years since its inception, the SLOE has become the most important piece of information that program directors use to determine which candidates they will select to interview and how they will rank students for the Match.[2–4] The SLOE consists of the following (see online supplementary data for an example SLOE):

1. Grade (honors, high pass, pass, fail, with some institutions choosing to select only pass/fail)
2. "Global ranking" in which writers are instructed to rate the student against all other EM bound rotators, placing them in the top 10%, top third, middle third, or bottom third
3. Predicted placement on the institution's match list, again from top 10% to top, middle, and bottom third
4. Qualities necessary for success in EM ranked against peers
5. Narrative portion

An early study comparing the SLOE to the narrative letter of recommendation (NLOR) was favorable, indicating that the SLOE was significantly more user friendly, as it demonstrated a decrease in both writing and reviewing time, as well as being easier to interpret with high interrater reliability.[5] Other specialties, including otolaryngology, dermatology, and orthopedics, have adopted an SLOE as well. Due to these factors, a recent commentary in *Academic Medicine* highlighted these advantages of the SLOE over a NLOR and suggested that the SLOE be adopted by all specialties for use during the residency application process.[6] Across specialties, program directors cite letters of recommendation as highly important, ranking them the second most important factor for interview invites, only after failed USMLE Step 1 attempts.[7] Thus, increased use of the SLOE across specialties will have a significant effect on the transition from undergraduate to graduate medical education.

While there are demonstrated benefits of the SLOE over the NLOR, there has not been a comprehensive study of the validity evidence of the SLOE. Messick defines validity as the "inductive summary of both the

DOI: http://dx.doi.org/10.4300/JGME-D-20-01110.1

Editor's Note: The online version of this article contains an example of the standardized letter of evaluation.

existing evidence for and the potential consequences of score interpretations and use."[8] Providing evidence for the validity of an assessment tool is therefore necessary for the meaningful use of the tool. Here we present a scoping review of the published validity evidence of the EM SLOE, using Messick's framework for construct validity.[8] A scoping review was chosen to identify gaps and future directions, and because the heterogeneity of the literature makes a systematic review difficult.

Methods

A scoping review of the validity evidence of the EM SLOE was performed. Methods were developed following previously published guidance for conducting scoping reviews.[9]

In 2020, PubMed, Medline, Google Scholar, Web of Science Core Collection, and Embase were searched for "(sloe OR slor) emergency medicine" and all variations of the phrase "standard/standardized letter of recommendation/evaluation." Inclusion criteria included any studies in which the EM SLOE was the subject of study. Citations were then assessed as to whether the study question was related to validity and were excluded if not; abstracts were also excluded. The initial search was conducted by a single author (P.K.) erring on the side of inclusivity. Included citations were reviewed separately for exclusion criteria by both authors. Any disagreements were resolved by discussion.

Messick's framework for validity includes the following aspects: Content, Response Process, Internal Structure, Relation to Other Variables, and Consequences.[8] The study question in each article was reviewed by each author and placed into 1 of the 5 categories that seemed the best fit. There were no disagreements.

To determine whether a study provided evidence for or against each aspect of validity, each author again independently assessed the results and conclusions of the study. Any disagreement between the authors was resolved with a discussion.

Results

The initial search terms returned 212 citations. After application of the inclusion and exclusion criteria, 22 articles were included in our review. The majority of studies assessed a single question with a dichotomous outcome. One study with multiple questions was determined to have "mixed" evidence. There is no published literature examining the evidence for content validity.

Response Process

Fourteen studies have been published about the SLOE that could be categorized as representing evidence for response process, which makes this the most studied aspect of the SLOE.[5,10–22] Three of the 14 studies provided evidence for validity and 11 of the 14 provided evidence against validity of the SLOE.

In favor of the SLOE, a study discovered that the interrater reliability was 0.97, in contrast to NLORs that had an interrater reliability of 0.78.[5] The second study looked at gender bias in the narrative portion of the SLOE at one institution and found that the narrative was "relatively free of gender bias."[10] The third, published in 2019, again looked at gender differences in the narrative portion and determined that there was no difference in word type frequency.[11]

Eleven studies provided evidence against response process validity.[12–22] Six studies have shown that authors do not adhere to the ranking guidelines and that ranking inflation is rampant on the SLOE.[12–17] One review found that "nearly all" applicants were ranked near the top and that only 2% of letters used the bottom rankings.[12] Another study demonstrated that students were ranked in the "top 10%" 40% of the time, 83% of students were "above the level of their peers," and more than 95% of SLOEs ranked the students in the "top third" compared to their peers in the "qualifications for EM" section.[13] Similarly, a survey of SLOE writers found that only 39% admitted to using the full scale to rank applicants.[14] However, the most recent study in this area does show improvement from these 3 earlier studies, demonstrating a more even distribution between the categories of top 10% and top, middle, and bottom third.[15] Even with the demonstrated improvement, writers still exhibited a reluctance to use the full scale as students were still ranked in a top-heavy fashion.[15] Additionally, 68% of SLOE writers do not follow the given SLOE instructions, and 67% of writers were not formally instructed on how to fill out a SLOE.[16]

Another study examining grading differences found wide grading practice variability between clerkships.[18] The percentage of students who received an honors grade at a specific clerkship varied between from 1% to 87%, some schools used 3-point grade scales while others used 5-point scales, and some schools were graded as pass/fail.[18] The grade is included on the SLOE.

Furthermore, studies have shown that variables specific to the letter writer can affect the SLOE. Literature demonstrated higher ratings being given to students by less experienced writers and by writers who have known the student for a longer period of time.[19] Similarly, student scores were consistently

higher on a letter written by their home institution compared to those written after visiting clerkships.[20] Moreover, while the 2 studies described above state that there is no gender effect in the SLOE, 2 other studies do testify to this as a phenomenon.[21,22] A study found that it was significantly more likely for a student to receive the highest possible ranking if the student was female and the writer was female; no other differences existed for any other gender pairing.[21] Finally, female students were found to have statistically significant higher scores than male students on the SLOE.[22]

The majority of studies regarding response process provide robust evidence against validity. Additionally, studies regarding gender differences provide conflicting conclusions. This aspect of validity has been studied the most, and while the evidence against validity is discouraging, the most recent and largest study does show a significant improvement in an even distribution of rankings, along the top 10, and top, middle, and bottom third, versus older studies.

Internal Structure

There is one study published relating to the internal structure of the SLOE. A 2001 study correlated the rank of "guaranteed match" (the highest possible ranking prior to SLOE revision in 2002) with other parts of the SLOE.[23] The authors demonstrated that the guaranteed match ranking was correlated with the honors grade, a ranking of "outstanding" on differential diagnosis, a ranking of "outstanding" on work ethic, and a ranking of "outstanding" on the global assessment, all as one would expect, providing some evidence for internal structure.[23] However, guaranteed match also correlated with the author's position, as well as if the author and student had a relationship outside of the emergency department.[23] This single study from 2002 provides very little overall evidence either way for internal structure, demonstrating that this aspect of validity of the SLOE needs further study.[24]

Relation to Other Variables

Four studies have been published regarding the SLOE's relation to other variables.[2,24–26] The first study compared rankings on the SLOR (this study was undertaken prior to the instrument's name was changed to SLOE) to a ranking of residents' "final success" upon graduation, with "final success" being defined after the faculty ranked each graduating resident against all previous residents at one institution.[24] The SLOR was not strongly correlated with this measure of success in residency.[24] The next study examined whether the SLOE category "predicted rank on the match list" correlated with the actual match list and found that the assessment accurately predicted the final rank order 26% of the time.[25] The authors found that the students' positions on the SLOE were overestimated 66% of the time and underestimated 8% of the time. A later study showed that the global assessment portion of the SLOE was positively correlated with the final rank list, with a Spearman's correlation of 0.332.[2] Finally, the most recent article compared the individual's SLOE to their performance as a graduating resident; institutions grouped the residents into thirds based on a score created from the numerical values on their Accreditation Council for Graduate Medical Education Milestone assessments. The authors found that the residents' "final ability" was correlated with the SLOE's global assessment as well as the SLOE's ranking of competitiveness.[26] In summary, there is minimal study regarding relation to variables, making it hard to draw conclusions in either direction. While the results from the 3 studies are mixed, the 2 most recent studies are trending in the correct direction for validity.

Consequences

Two articles have been published regarding the consequences of the SLOE.[3,4] Both are surveys of EM program directors which found that the SLOE is the most important piece of data when choosing who to interview and, subsequently, rank.[3,4] These studies provide evidence that consequences of the SLOE are high; however, no studies have been performed looking at how the high-stakes nature of the SLOE may affect letter writers or how it may affect students' behavior during a clerkship. While we can predict with some degree of certainty that the consequences to the SLOE are very high, studies are necessary to uncover its exact relation to the validity of the SLOE. Currently, it is not possible to conclude how the high consequences of the SLOE affect its validity.

See the TABLE for a summary of the evidence for validity of the EM SLOE.

Discussion

Overall, we found that the evidence for validity for the EM SLOE is lacking. While the SLOE has good evidence for content validity owing to its creation process, there is not strong evidence for any other aspect of validity.

We believe the development process for the SLOE provides evidence for content validity. CORD initially convened a task force in 1995 to create the SLOE after concerns that usual NLORs were not adequate.[1] The task force was comprised of a representative

TABLE
Literature Summary

Author, Year	Participants	Aims	Results	Evidence for Validity?
Content				
Keim et al, 1999	SLOE Task Force	Describe the creation process of the EM SLOE	▪ Task force convened in 1995, consensus development process with EM education experts ▪ Pilot first year, edits made after survey of program directors	Yes
Response Process				
Girzadas et al, 1998	20 SLORs and 20 NLORs submitted to one program	Compare SLOR to NLOR in EM applications	▪ Interrater reliability was 0.97 for the SLOR, compared to 0.78 for the NLOR ▪ Average time to interpret a SLOR was 16 seconds vs 90 seconds for an NLOR	Yes
Harwood et al, 2000	432 SLORs submitted to one program	Assess grade and rank distribution on the SLOE	SLOR authors did not use the full scale ▪ Grades: 55% honors, 36% pass, 9% pass ▪ Global assessment: 37% outstanding, 49% excellent, 14% very good or good ▪ Match: 23% guaranteed, 50% very likely, 27% likely and possible	No
Girzadas et al, 2004	835 SLORs submitted to one program	Assess for gender bias on rankings on the SLOE	▪ A female author writing a letter for a female applicant was highly associated with giving the highest Match rank on the SLOR ▪ No other gender combination was significant	No
Love et al, 2013	602 SLORs submitted to 3 different programs	Assess grade and rank distribution on the SLOE	Showed ranking inflation ▪ On global assessment, 40% of students were top 10% ▪ 95% of students were in the top third compared to peers for the qualifications for EM section	No
Beskind et al, 2014	1253 SLORs submitted to 3 different programs	Determine whether characteristics of the letter writer affected rankings on the SLOE	▪ Less experienced writers were more likely to give a higher ranking ▪ The length of time an author knew the applicant was associated with high rankings	No
Hegarty et al, 2014	320 of 695 (46%) CORD members	Survey SLOE authors on their practices regarding filling out SLOEs	▪ 67% of SLOE writers did not receive instruction in how to fill out a SLOE ▪ 68% of SLOE writers state they do not follow the instructions on certain questions	No
Grall et al, 2014	1457 SLORs submitted to 3 different programs	Assess grade and rank distribution on the SLOE	Showed ranking inflation ▪ For 4-point scale variables, 91% were ranked as the top 2 options ▪ For 3-point scale ratings, 94.6% were ranked as the top 2 options ▪ Less than 2% of SLOEs were ranked in the bottom third	No
Li et al, 2017	237 first-rotation SLOEs of applicants invited to interview at one program	Assess the narrative portion of the SLOE for gender bias	▪ Examined 237 SLOEs and found that the narrative portion was "relatively free of gender bias"	Yes

TABLE
Literature Summary (Continued)

Author, Year	Participants	Aims	Results	Evidence for Validity?
Hall et al, 2017	1075 applications to one program consisting of grades from 236 different clerkships	Assess grade variability between different schools	▪ The percentage of students that receive an honors grade at a school ranges from 1%–87% ▪ Some schools are pass/fail ▪ Some schools use 3-point grade scales, some use 5 ▪ Some schools give grades, but not honors	No
Pelletier-Bui et al, 2018	99 respondents, survey sent to CORD and CDEM (clerkship directors in EM) listservs	Survey SLOE authors on their practices regarding filling out SLOEs	▪ 39% responded that they strictly adhere to the ranking guidelines	No
Jackson et al, 2019	6715 SLOEs for 3138 unique applicants accessed from the eSLOE database	Assess grade and rank distribution on the SLOE	Showed ranking inflation (although improved from the 2013 study) ▪ Global assessment: 18% top 10%, 37% top third, 35% middle third, 10% lower third ▪ Match rank list: 18% top 10%, 36% top third, 32% middle third, 12% lower third, 2% unlikely to rank	No
Boysen-Osborn et al, 2019	624 applicants to one program	Compare rankings on SLOEs written by a student's home institution to those written after a visiting rotation	▪ Authors created an overall composite score for a SLOE ▪ The composite score was better on SLOEs written by a home school than those obtained on a visiting clerkship	No
Miller et al, 2019	822 first rotation SLOEs submitted to one program 64% male and 36% female	Assess differences in word type frequency by gender on the narrative portion of the SLOE	▪ No significant difference in word type frequency by gender in the narrative portion	Yes
Andrusaitis et al, 2019	2092 SLOEs submitted to one program	Assess for gender bias in overall scores on the SLOE	▪ Females have better overall scores on the SLOE than males	No
Internal Structure				
Girzadas et al, 2001	411 SLORs submitted to one program	Find associations between a ranking of "guaranteed match" (the highest rank at the time) and other rankings on the SLOE and author variables	A ranking of "guaranteed match" was highly correlated with both ▪ An honors grade, an outstanding ranking on differential diagnosis, an outstanding ranking on work ethic, and an outstanding ranking on global assessment ▪ The authors position and having clinical contact outside the ED	Mixed
Relation to Other Variables				
Hayden et al, 2005	54 graduating residents from one program	Compare SLOE rankings to residents' "final success" upon graduation	▪ Ranked graduating residents into percentiles (against all previous residents) at one institution ▪ The SLOR was not strongly correlated with this measure of success	No

TABLE
Literature Summary (Continued)

Author, Year	Participants	Aims	Results	Evidence for Validity?
Oyama et al, 2010	102 SLORs from 5 programs	Compare predicted Match list position on the SLOE to the actual Match list position	▪ 26% of SLOEs had a predicted match rank that matched the actual match rank ▪ 66% of the time the SLOE overestimated the rank position ▪ 8% of the time it underestimated the rank position	No
Breyer et al, 2012	127 applications to one program	Compare predicted Match list position on the SLOE to the actual Match list position	▪ Global assessment on the SLOE was positively correlated with final rank list for Match ▪ Spearman's correlation 0.332	Yes
Bhat et al, 2015	277 residents consisting of 3 graduating classes from 9 programs	Compare SLOE rankings to residents' "final ability" upon graduation	Faculty ranked residents' "final ability" upon graduation, which ▪ Correlated with the global assessment ▪ Correlated with ranking of competitiveness on the SLOE	Yes
Consequences				
Love et al, 2014	150 of 159 (94.3%) EM program directors	Survey EM program directors about their perspectives regarding the SLOE	▪ SLOE was ranked as the number one data point when deciding who to interview	No
Negaard et al, 2018	120 members of the CORD listserv	Survey EM program directors to describe EM residency selection criteria	▪ The visiting rotation SLOE was ranked as the number one data point when creating the final Match list ▪ The home rotation SLOE was third most important data point when creating the final Match list	No

Abbreviations: SLOE, Standardized Letter of Evaluation; EM, emergency medicine; SLOR, Standardized Letter of Recommendation; NLOR, Narrative Letter of Recommendation; CORD, Council of Residency Directors in Emergency Medicine; CDEM, Clerkship Directors in Emergency Medicine; ED, emergency department.

sample of CORD membership, consisting of program directors, assistant program directors, and clerkship directors. In 1999, Keim et al described the initial creation process and how the task force determined what to include on the form.[1] In 1996 and 1999, the SLOR was edited by the task force based on unpublished surveys that had been distributed to program directors throughout the country.[1] The task force was reconvened in 2011 to update and improve the SLOE. Changes were made after 2 published studies and one unpublished survey, which included a change to the name from the Standardized Letter of Recommendation to the Standardized Letter of Evaluation.[3,16] Additional categories were added to the "Qualifications for EM" section, including teamwork, ability to communicate a caring nature to patients, how much guidance an applicant would need in residency, and predicted success in residency. Further, CORD has shown that it can adapt quickly when necessary; the task force reconvened in 2020 to address SLOE issues related to changes due to the COVID-19 pandemic. This process provides continuing evidence for content validity, as the content of the SLOE changes to reflect the changing informational needs of program directors. We, therefore, conclude that the content of the SLOE should represent what the SLOE is intended for, and have evidence for content validity.

Response process has been the most studied, and the evidence overall currently argues against validity. Studies on the dermatology SLOR, otolaryngology SLOR, and orthopedic SLOR have all demonstrated similar rank inflation.[27–29] The overall theme emerging from the literature is that better rater training will improve adherence to ranking distribution; however, there may not be evidence to support this claim. Multiple studies do show that rater training can improve the quality of assessment reports and improve the ability of faculty to assess residents.[30,31] Nevertheless, studies also show that rater training has

no effect, even on standardized clinical examinations.[32,33] On the EM SLOE, adherence to the rating system has improved over the years, and the authors of the most recent study suggest that rater training is the reason for the improvement.[15] While an increased focus on rater training may have improved adherence to the rankings on the EM SLOE, the questionable effect of rater training in general and number of years the EM SLOE has existed leads us to believe that rater training is unlikely to yield further improvement to the SLOE's response process.

Concern about the consequences of the SLOE may limit adherence to the ranking scale despite any additional rater training. A survey presented at the 2016 CORD Academic Assembly shows that 40% of EM program directors do not match students ranked in the lower third.[34] Further, current instructions on the electronic SLOE (eSLOE) state that when choosing a comparative ranking, writers should consider only "candidates you have recommended in the last academic year" (see online supplementary data). If an institution writes a small number of SLOEs, this potentially creates a situation that creates an unfavorable designation for an otherwise competitive student. For example, an outstanding student who is slightly outperformed by a handful of others should technically be rated as "lower third" even though the writer knows the performance was outstanding. Based on the above survey data, the current SLOE asks writers to choose between adhering to the ranking scale or potentially consigning outstanding students to a lower likelihood of matching. Therefore, the consequences of a "lower third" ranking may dampen any positive effect that rater training may have on ranking scale adherence.

Thus, rather than continuing to study whether or not there is strict adherence to the ranking system or pushing for further rater training, we submit that a reconsideration of the current ranking system and instructions is necessary. Rather than using norm-based percentiles that create difficulties in compliance, criterion-based descriptors may help writers faithfully assign students to a category. The current norm-based ranking system uses strict percentile cutoffs, meaning absolute adherence could cause 2 students of almost identical ability to be placed into different rankings. Proper norm-based ranking would use standard deviation from the mean,[35] which is not feasible for the EM SLOE, as it requires precise numerical scores, such as with multiple-choice tests. Criterion-based rankings with descriptions would not eliminate ranking inflation, but writers may have an easier time placing students into categories that contain a description of the typical student in that category (eg, "independently creates treatment plans that do not require modification"). This would add more meaningful contextualization of the applicant for residency programs as well as create a more equitable evaluation system for students.

Switching from a norm-referenced to a criterion-based system may also help to combat bias on the SLOE. A study of language use in narrative assessments found that female and underrepresented in medicine (UiM) medical students had significantly more personality attributes described, compared to competency-based language used for male and non-UiM students.[36] Changing to a criterion-based system grounded with competency descriptors will force writers to consider the chosen competencies when assessing students rather than relying on personality attributes and may therefore decrease implicit bias in ranking. This would need to be further studied but would present an opportunity to examine a potential method to systematically reduce bias in medical assessments.

Whether or not the evaluation system changes, bias on the SLOE requires further study. Gender bias has been examined by multiple studies, with mixed results, trending toward favoring female applicants. However, racial bias in SLOE rankings has not been examined. Studies in other domains, including induction into the Alpha-Omega-Alpha (AOA) honor society, MSPE letters, and clerkship grades have all shown evidence of racial bias that negatively affects UiM groups.[37–39] Due to the documented existence of bias and the outsized importance the SLOE has on residency applications, future studies must assess what effect race has on the SLOE rankings.

Further complicating the response process is the lack of interrater reliability. While there will always be a degree of variability in workplace-based assessment, the large differences in each institution's clerkship make a standardized comparison between them difficult. While there is a published national curriculum for EM clerkships,[40] significant differences between clerkships remain.[41] Importantly, differences include how assessments are performed, with variations in whether residents are allowed to assess students; if a written test is used for assessment and, if so, which one; and whether direct observation is a requirement of assessment.[41] Key clerkship differences are further illustrated by the wide variability of grading practices, in which some clerkships are pass/fail, some give grades but not honors, and some use a range of 3- to 5-point scales.[18] These factors make creating a "standardized" letter to compare students across the country very difficult, if not impossible. To address this, stakeholders need to push for further standardization among clerkship curricula. Additionally, consensus on how assessments are performed and

by whom should be published. Finally, using a standardized shift assessment, so that SLOEs are based on the same inputs across clerkships, will create a more reliable assessment. The National Clinical Assessment Tool created by a consensus conference at CORD is a potential tool that could be widely adopted to assist with this process.[42] This tool will need further evidence for validity prior to its widespread use. Leaders in EM education need to push for the study, and if it demonstrates evidence for validity, adoption of this tool, as well as the inclusion of an item on the SLOE to indicate whether or not it is used during the clerkship so that applications reviewers can make their own assessment about validity.

Next, relation to other variables for the EM SLOE remains understudied. Without larger, more robust studies in this domain, it is difficult to know whether the SLOE is actually predictive of future success in residency and therefore serving its original purpose. Our results demonstrate that the focus of study on the EM SLOE has been weighted heavily toward the inputs, despite the predictive value perhaps being even more important. The new eSLOE format creates a large database to perform multi-institutional studies comparing it to other variables; performing these studies will be a necessary step to provide further evidence for validity for the EM SLOE.

Taking steps to improve and study the EM SLOE will become even more important to both EM and to all specialties using or considering a standardized letter after the recent decision by the Federation of State Medical Boards and National Board of Medical Examiners to make the USMLE Step 1 to be reported as pass/fail.[43] Previous surveys have shown that Step 1 was either the third most important factor or factor of "middle importance" to interviewing and ranking for matching.[3,4] It would be reasonable to predict that by removing another objective variable, the SLOE will gain even more importance to program directors and future residents. This could have even more significant effects in other specialties currently using or considering adopting the SLOE, as each specialty values the USMLE Step 1 score differently. If the SLOE continues to be utilized by program directors as the most important factor in medical students' applications, further improvement to make it the best tool possible is required.

There are limitations to our study's findings. During our data collection process we did not include poster presentations and abstracts, meaning there could be further evidence for validity for the EM SLOE that was not discovered. Second, many studies examining the same aspect of the SLOE have differing results, which can make consistent conclusions on these aspects of validity difficult. Third, the nature of this review is inherently subjective regarding each individual study examined. Despite this limitation, applying Messick's framework for validity evidence to the whole should add reliability to our results.

Other specialties should take note of the current challenges facing the EM SLOE and edit or create their own standardized letters accordingly. First, stakeholders should consider the drawbacks of using norm-based percentile rankings and consider using criterion-based descriptive categories. Next, evaluators must be aware of the implicit and systemic bias that exist within assessments and work to address this in any standardized letter. Additionally, specialties need to examine current clerkship differences and advocate for the standardization of the clerkship experience, particularly the assessment portion. Finally, specialties should perform early study on the relation to other variables to provide further evidence for validity for their standardized letters.

Conclusions

There is little evidence for validity for the EM SLOE regarding response process, internal structure, or relation to other variables.

References

1. Keim SM, Rein JA, Chisholm C, et al. A standardized letter of recommendation for residency application. *Acad Emerg Med*. 1999;6(11):1141–1146. doi:10.1111/j.1553-2712.1999.tb00117.x
2. Breyer MJ, Sadosty A, Biros M. Factors affecting candidate placement on an emergency medicine residency program's rank order list. *West J Emerg Med*. 2012;13(6):458–462. doi:10.5811/westjem.2011.1.6619
3. Love JN, Smith J, Weizberg M, et al. Council of Emergency Medicine Residency Directors' standardized letter of recommendation: the program director's perspective. *Acad Emerg Med*. 2014;21(6):680–687. doi:10.1111/acem.12384
4. Negaard M, Assimacopoulos E, Harland K, Van Heukelom J. Emergency medicine residency selection criteria: an update and comparison. *AEM Educ Train*. 2018;2(2):146–153. doi:10.1002/aet2.10089
5. Girzadas DV Jr, Harwood RC, Dearie J, Garrett S. A comparison of standardized and narrative letters of recommendation. Comparative study. *Acad Emerg Med*. 1998;5(11):1101–1104. doi:10.1111/j.1553-2712.1998.tb02670.x
6. Love JN, Ronan-Bentle SE, Lane DR, Hegarty CB. The Standardized Letter of Evaluation for postgraduate training: A concept whose time has come? *Acad Med*. 2016;91(11):1480–1482. doi:10.1097/ACM.0000000000001352

7. National Resident Matching Program. Results of the 2020 NRMP Program Director Survey. https://www.nrmp.org/main-residency-match-data/. Accessed April 23, 2021.
8. Messick S. Validity. In: Linn RL, ed. *The American Council on Education/Macmillan series on higher education. Educational Measurement*. 3rd ed. Macmillan Publishing Co Inc; 1989:13–103.
9. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc*. 2015;13(3):141–146. doi:10.1097/XEB.0000000000000050
10. Li S, Fant AL, McCarthy DM, Miller D, Craig J, Kontrick A. Gender differences in language of standardized letter of evaluation narratives for emergency medicine residency applicants. *AEM Educ Train*. 2017;1(4):334–339. doi:10.1002/aet2.10057
11. Miller DT, McCarthy DM, Fant AL, Li-Sauerwine S, Ali A, Kontrick AV. The standardized letter of evaluation narrative: differences in language use by gender. *West J Emerg Med*. 2019;20(6):948–956. doi:10.5811/westjem.2019.9.44307
12. Grall KH, Hiller KM, Stoneking LR. Analysis of the evaluative components on the standard letter of recommendation (SLOR) in emergency medicine. *West J Emerg Med*. 2014;15(4):419–423. doi:10.5811/westjem.2014.2.19158
13. Love JN, Deiorio NM, Ronan-Bentle S, et al. Characterization of the Council of Emergency Medicine Residency Directors' standardized letter of recommendation in 2011–2012. Multicenter Study. *Acad Emerg Med*. 2013;20(9):926–932. doi:10.1111/acem.12214
14. Pelletier-Bui A, Van Meter M, Pasirstein M, Jones C, Rimple D. Relationship between institutional standardized letter of evaluation global assessment ranking practices, interviewing practices, and medical student outcomes. *AEM Educ Train*. 2018;2(2):73–76. doi:10.1002/aet2.10079
15. Jackson JS, Bond M, Love JN, Hegarty C. Emergency medicine standardized letter of evaluation (SLOE): findings from the new electronic SLOE format. *J Grad Med Educ*. 2019;11(2):182–186. doi:10.4300/JGME-D-18-00344.1
16. Hegarty CB, Lane DR, Love JN, et al. Council of emergency medicine residency directors standardized letter of recommendation writers' questionnaire. *J Grad Med Educ*. 2014;6(2):301–306. doi:10.4300/JGME-D-13-00299
17. Harwood RC, Girzadas DV, Carlson A, et al. Characteristics of the emergency medicine standardized letter of recommendation. *Acad Emerg Med*. 2000;7(4):409–410. doi:10.1111/j.1553-2712.2000.tb02253.x
18. Hall MM, Dubosh NM, Ullman E. Distribution of honors grades across fourth-year emergency medicine clerkships. *AEM Educ Train*. 2017;1(2):81–86. doi:10.1002/aet2.10018
19. Beskind DL, Hiller KM, Stolz U, et al. Does the experience of the writer affect the evaluative components on the standardized letter of recommendation in emergency medicine? *J Emerg Med*. 2014;46(4):544–550. doi:10.1016/j.jemermed.2013.08.025
20. Boysen-Osborn M, Andrusaitis J, Clark C, et al. A retrospective cohort study of the effect of home institution on emergency medicine standardized letters of evaluation. *AEM Educ Train*. 2019;3(4):340–346. doi:10.1002/aet2.10374
21. Girzadas DV Jr, Harwood RC, Davis N, Schulze L. Gender and the council of emergency medicine residency directors standardized letter of recommendation. *Acad Emerg Med*. 2004;11(9):988–991. doi: 10.1197/j.aem.2004.03.024
22. Andrusaitis J, Clark C, Saadat S, et al. Does applicant gender have an effect on standardized letters of evaluation obtained during medical student emergency medicine rotations? *AEM Educ Train*. 2020;4(1):18–23. doi:10.1002/aet2.10394
23. Girzadas DV Jr, Harwood RC, Delis SN, et al. Emergency medicine standardized letter of recommendation: predictors of guaranteed match. *Acad Emerg Med*. 2001;8(6):648–653.
24. Hayden SR, Hayden M, Gamst A. What characteristics of applicants to emergency medicine residency programs predict future success as an emergency medicine resident? *Acad Emerg Med*. 2005;12(3):206–210. doi:10.1197/j.aem.2005.01.002
25. Oyama LC, Kwon M, Fernandez JA, et al. Inaccuracy of the global assessment score in the emergency medicine standard letter of recommendation. Multicenter study. *Acad Emerg Med*. 2010;17(suppl 2):38–41. doi:10.1111/j.1553-2712.2010.00882.x
26. Bhat R, Takenaka K, Levine B, et al. Predictors of a top performer during emergency medicine residency. Multicenter study. *J Emerg Med*. 2015;49(4):505–512. doi:10.1016/j.jemermed.2015.05.035
27. Wang RF, Zhang M, Alloo A, Stasko T, Miller JE, Kaffenberger JA. Characterization of the 2016–2017 dermatology standardized letter of recommendation. *J Clin Aesthet Dermatol*. 2018;11(3):26–29.
28. Kominsky AH, Bryson PC, Benninger MS, Tierney WS. Variability of ratings in the otolaryngology standardized letter of recommendation. *Otolaryngol Head Neck Surg*. 2016;154(2):287–293. doi:10.1177/0194599815623525
29. Kang HP, Robertson DM, Levine WN, Lieberman JR. Evaluating the standardized letter of recommendation form in applicants to orthopaedic surgery residency.

J Am Acad Orthop Surg. 2020;28(19):814–822. doi:0.5435/JAAOS-D-19-00423
30. Dudek NL, Marks MB, Wood TJ, et al. Quality evaluation reports: can a faculty development program make a difference? *Med Teach*. 2012;34(11):e725–e731. doi:10.3109/0142159X.2012.689444
31. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874–881. doi:10.7326/0003-4819-140-11-200406010-00008
32. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74–79. doi:10.1007/s11606-008-0842-3
33. Weitz G, Vinzentius C, Twesten C, Lehnert H, Bonnemeier H, König IR. Effects of a rater training on rating accuracy in a physical examination skills assessment. *GMS Z Med Ausbild*. 2014;31(4):Doc41. doi:10.3205/zma000933
34. Pelletier-Bui A, Rimple D, Pasirstein M, Van Meter M. SLOE lower third ranking: is it the kiss of death? *West J Emerg Med*. 2016;17(4.1)
35. Chan WS. A better norm-referenced grading using the standard deviation criterion. *Teach Learn Med*. 2014;26(4):364–365. doi:10.1080/10401334.2014.945031
36. Rojek AE, Khanna R, Yim JWL, et al. Differences in narrative language in evaluations of medical students by gender and under-represented minority status. *J Gen Intern Med*. 2019;34(5):684–691. doi:10.1007/s11606-019-04889-9
37. Wijesekera TP, Kim M, Moore EZ, Sorenson O, Ross DA. All other things being equal: exploring racial and gender disparities in medical school honor society induction. *Acad Med*. 2019;94(4):562–569. doi:10.1097/ACM.0000000000002463
38. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekroud A, Moore EZ. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. *PLoS One*. 2017;12(8):e0181659. doi:10.1371/journal.pone.0181659
39. Boatright D, Ross D, O'Connor P, Moore E, Nunez-Smith M. Racial disparities in medical student membership in the Alpha Omega Alpha Honor Society. *JAMA Intern Med*. 2017;177(5):659–665. doi:10.1001/jamainternmed.2016.9623
40. Manthey DE, Ander DS, Gordon DC, et al. Emergency medicine clerkship curriculum: an update and revision. *Acad Emerg Med*. 2010;17(6):638–643. doi: 10.1111/j.1553-2712.2010.00750.x
41. Khandelwal S, Way DP, Wald DA, et al. State of undergraduate education in emergency medicine: a national survey of clerkship directors. *Acad Emerg Med*. 2014;21(1):92–95. doi: 10.1111/acem.12290
42. Jung J, Franzen D, Lawson L, et al. The National Clinical Assessment Tool for Medical Students in the Emergency Department (NCAT-EM). *West J Emerg Med*. 2018;19(1):66–74. doi:10.5811/westjem.2017.10.34834
43. United States Medical Licensing Examination. Changes to pass/fail score reporting for Step 1. https://www.usmle.org/incus/. Accessed April 23, 2021.

Both authors are with the University of Chicago Medical Center. **Paul Kukulski, MD,** is Assistant Professor and Assistant Clerkship Director, Section of Emergency Medicine, Department of Medicine; and **James Ahn, MD, MHPE,** is Associate Professor and Program Director, Section of Emergency Medicine, Department of Medicine.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

Findings from this study were previously presented as an abstract at the Council of Emergency Medicine Program Directors Academic Assembly, New York, NY, March 8–11, 2020.

Corresponding author: Paul Kukulski, MD, University of Chicago Medical Center, paul.kukulski@uchospitals.edu

Received September 21, 2020; revisions received February 4, 2021, and April 13, 2021; accepted April 15, 2021.