Hawks and Doves in Standardized Letters of Evaluation: 6 Years of Rating Distributions and Trends in Emergency Medicine

Eric Shappell, MD, MHPE
Cullen Hegarty, MD
Sharon Bord, MD
Daniel J. Egan, MD

ABSTRACT

Background Standardized Letters of Evaluation (SLOEs) are an important part of resident selection in many specialties. Often written by a group, such letters may ask writers to rate applicants in different domains. Prior studies have noted inflated ratings; however, the degree to which individual institutions are "doves" (higher rating) or "hawks" (lower rating) is unclear.

Objective To characterize institutional SLOE rating distributions to inform readers and developers regarding potential threats to validity from disparate rating practices.

Methods Data from emergency medicine (EM) SLOEs between 2016 and 2021 were obtained from a national database. SLOEs from institutions with at least 10 letters per year in all years were included. Ratings on one element of the SLOE—the "global assessment of performance" item (Top 10%, Top Third, Middle Third, and Lower Third)—were analyzed numerically and stratified by predefined criteria for grading patterns (Extreme Dove, Dove, Neutral, Hawk, Extreme Hawk) and adherence to established guidelines (Very High, High, Neutral, Low, Very Low).

Results Of 40 286 SLOEs, 20 407 met inclusion criteria. Thirty-five to 50% of institutions displayed Neutral grading patterns across study years, with most other institutional patterns rated as Dove or Extreme Dove. Adherence to guidelines was mixed and fewer than half of institutions had Very High or High adherence each year. Most institutions underutilize the Lower Third rating.

Conclusions Despite explicit guidelines for the distribution of global assessment ratings in the EM SLOE, there is high variability in institutional rating practices.

Introduction

Standardized Letters of Evaluation (SLOEs) are an important part of resident selection in multiple specialties.[1] SLOE structures are established by national specialty groups and typically populated by a physician or team based on performance during a rotation in the student's desired specialty.

One goal of SLOEs is to differentiate between levels of trainee performance.[2-4] Many SLOEs utilize normative assessments to help stratify performance.[5-9] Effective normative assessments require a shared mental model for authors and readers. Anchors may facilitate this model, but previous work exploring aggregate data from EM,[5] orthopedic surgery,[10] and dermatology[11,12] demonstrates inflated ratings. Questions remain, however, regarding how individual institutions contribute to observed inflation. For example, is there an even slope of grade inflation from the most hawkish[13] (lower rating) to the most dovish (higher rating) programs? Or do a few dovish programs bring up the mean for an otherwise uniform group? These questions have implications for SLOE readers to understand the frequency and level of adjustment necessary to normalize ratings between institutions, and for SLOE developers, to characterize the threat to SLOE validity from inflated ratings.

The purpose of this study is to characterize the distribution of ratings by institutions on the EM SLOE global assessment item, which stratifies performance as Top 10%, Top Third, Middle Third, or Lower Third.

Methods

Data Acquisition

Anonymized data were obtained from the Council of Residency Directors in Emergency Medicine (CORD) SLOE database after approval by the CORD Board of Directors. Abstracted data included institution, submission date, and global assessment of performance from SLOEs between 2016 and 2021. Only SLOEs written by EM faculty for general EM rotations were included.

DOI: http://dx.doi.org/10.4300/JGME-D-23-00231.1

Editor's Note: The online supplementary data contains further data from the study.

Given that the global assessment item classifies performance into 1 of 4 different categories (Top 10%, Top Third, Middle Third, and Lower Third), it was determined a priori that only authors or identical author groups writing 10 or more SLOEs per study year would be included, since the distribution of ratings in small samples is expected to be irregular even if adhering to guidelines (eg, if only 2 SLOEs are submitted, it is impossible to have an even distribution across 4 categories; online supplementary data FIGURE 1). No institution had more than one author or author group meet inclusion criteria (ie, all included institutions had a single qualifying source of SLOEs). Therefore, the term "institution" is used to refer to SLOE sources for the remainder of the manuscript.

Data Analysis

Dove vs Hawk Designations: To categorize how institutions adhered to expected distributions of higher vs lower applicant rankings, we created the Dove/Hawk framework. We assigned expected distributions of 10% in Top 10% and 30% each in Top Third, Middle Third, and Lower Third based on extensive experience working with the EM SLOE as the common interpretation of these categories. Institutions were stratified with ≥10% and ≥20% above vs below the expected percentage of candidates receiving the 2 higher categories (Top 10% and Top Third) versus the other categories (Middle Third and Lower Third). Five categories were used: Extreme Dove (≥60% Higher ratings), Dove (50%-60% Higher ratings), Neutral (30%-50% Higher ratings), Hawk (20%-30% Higher ratings), and Extreme Hawk (≤20% Higher ratings).

Adherence Designations: The adherence to expected distributions detailed above (10% of applicants in Top 10%, 30% in other groups) were categorized using a priori cutoffs of Very High (<10% difference between expected and observed percentage of ratings for each rating group), High (<20% difference between expected and observed percentages for each rating group), Low (<20% difference between expected and observed percentages excluding Lower Third), and Very Low (≥20% difference between expected and observed percentages in any group other than Lower Third). For example, if an institution rated 10% of applicants in the Top 10% (no difference in observed vs expected), 45% in the Top Third and Middle Third (15% difference in observed vs expected), and 0% in the Lower Third (30% difference in observed vs expected), the category is Low. Lower Third ratings were excluded for Low and Very Low groups because this rating is rarely used, and its inclusion was expected to result in less informative homogenously poor adherence categorizations.

Descriptive Statistics

Rating distributions by study year were summarized using descriptive statistics. Statistical significance of rating differences across years was assessed using Poisson regression.

Given the lack of prior literature on deviation from SLOE anchors, cutoffs for adherence and grade inflation were developed based on author experience and expertise in evaluation of assessments. Statistical work was performed using Stata 16 (StataCorp LLC). This study was designated as exempt by the Institutional Review Board at Mass General Brigham.

Results

Of 40 286, 20 407 SLOEs (51%) met inclusion criteria, coming from 105 unique institutions. The remaining 49% lacked at least 10 institutional SLOEs in all study years.

Dove vs Hawk Designations

Of the Dove vs Hawk designations, the neutral category was most common (36%-49% of institutions each year, FIGURE 1A-B, TABLE). Few institutions met Hawk or Extreme Hawk criteria in any study year (total Hawk or Extreme Hawk ratings: 25 of 630, 4%, and 2 of 630, <1%, respectively), and the split between total Dove and Extreme Dove designations for all study years was relatively even (168 of 630, 27%, and 178 of 630, 28%, respectively). There were no statistically significant differences in frequency of Dove vs Hawk categorizations across years (*P* values all >.99). Most individual programs met criteria for either 2 or 3 different Dove/Hawk designations over the 6 study years (online supplementary data TABLE).

Adherence Designations

Fewer than half of institutions had Very High or High adherence each year (TABLE, online supplementary data FIGURE 2). There were no statistically significant differences in frequency of adherence categorizations ratings across years (*P* values all >.99).

Descriptive Statistics

Mean and median rating in each category were relatively stable across study years (TABLE, online supplementary data FIGURE 2). There were no statistically significant differences in frequency of rating use across years (*P* value range .89 to >.99). The range

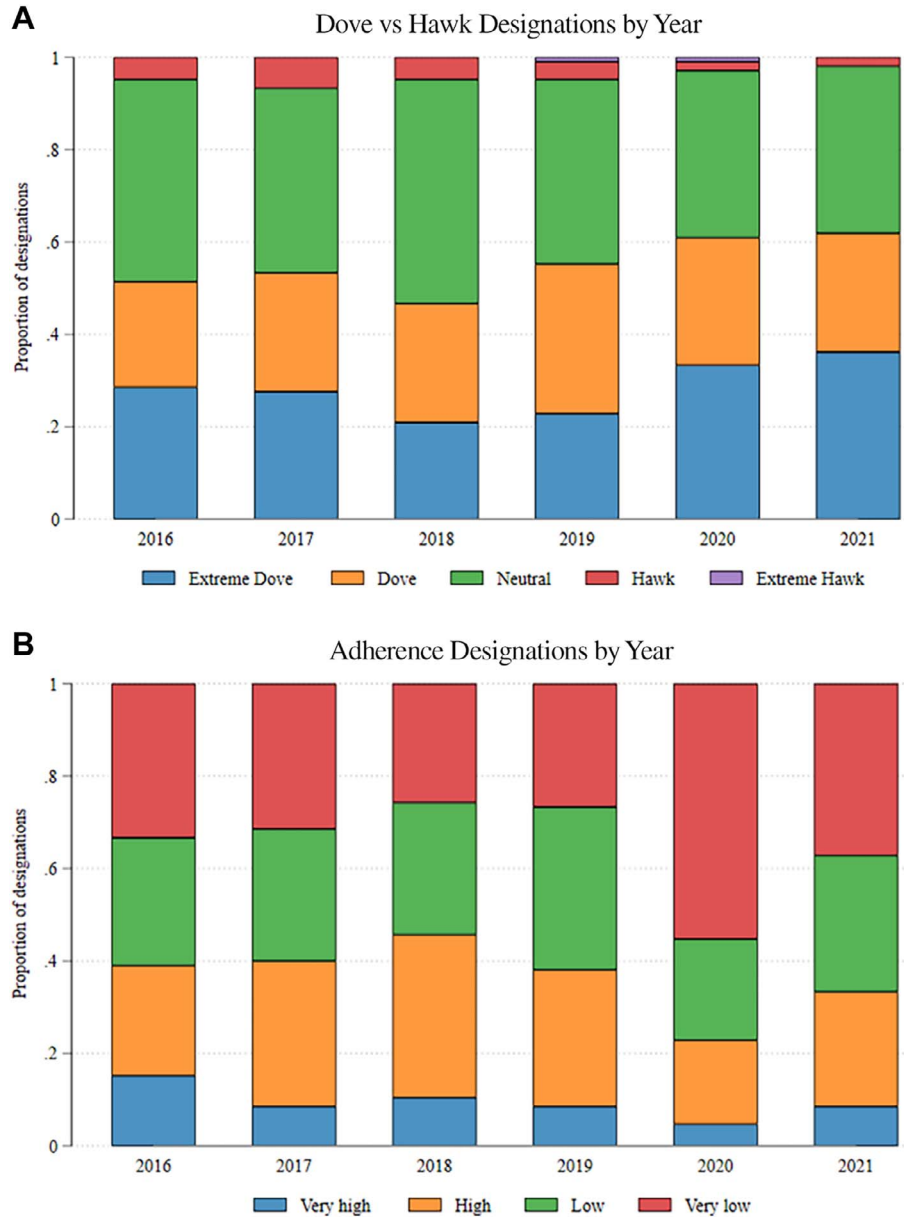

FIGURE 1A-B
Emergency Medicine Standardized Letter of Evaluation Dove vs Hawk and Adherence Designations by Year (2016-2021)

of use of the different ratings across institutions, however, was broad (TABLE, online supplementary data FIGURE 3).

Discussion

There is profound variability in the distributions of ratings used by individual programs completing the EM SLOE (online supplementary data FIGURE 3), with up to 87% variability in use of the Middle Third ranking in 2020 (TABLE). Averaged across study years, over half of institutions met criteria for Dove or Extreme Dove categories and had adherence designations of Low or Very Low. These findings demonstrate a threat to the validity of the EM SLOE and underscore the importance of contextualization by residency recruitment committees when interpreting SLOEs. This study is limited in that it examines only EM SLOEs from institutions that produce 10 or more SLOEs per year. How results from SLOEs from lower-volume sites or other specialties would compare is unclear.

As competency-based assessments increasingly replace norm-referenced assessments, rating inflation will become more difficult to quantify. If differentiation of levels of trainee performance remains a primary purpose of SLOEs, developers must monitor and

TABLE
Institutional Emergency Medicine Standardized Letter of Evaluation Dove vs Hawk Designations, Adherence Designations, and Rating Distributions by Year (2016-2021)[a]

Category	Key	2016	2017	2018	2019	2020	2021
Dove vs Hawk designations							
Extreme Doves	n (%)	30 (29)	29 (28)	22 (21)	24 (23)	35 (33)	38 (36)
Doves	n (%)	24 (23)	27 (26)	27 (26)	34 (32)	29 (28)	27 (26)
Neutral	n (%)	46 (44)	42 (40)	51 (49)	42 (40)	38 (36)	38 (36)
Hawks	n (%)	5 (5)	7 (7)	5 (5)	4 (4)	2 (2)	2 (2)
Extreme Hawks	n (%)	0 (0)	0 (0)	0 (0)	1 (1)	1 (1)	0 (0)
Adherence designations							
Very High	n (%)	16 (15)	9 (9)	11 (10)	9 (9)	5 (5)	9 (9)
High	n (%)	25 (24)	33 (31)	37 (35)	31 (30)	19 (18)	26 (25)
Low	n (%)	29 (28)	30 (29)	30 (29)	37 (35)	23 (22)	31 (30)
Very Low	n (%)	35 (33)	33 (31)	27 (26)	28 (27)	58 (55)	39 (37)
Rating distributions							
Top 10%	Mean % (±SD)	15 (±9)	15 (±8)	14 (±7)	13 (±7)	17 (±11)	16 (±8)
	Range %	39 (0-39)	43 (0-43)	42 (0-42)	42 (0-42)	53 (0-53)	40 (0-40)
Top Third	Mean % (±SD)	37 (±13)	36 (±11)	36 (±11)	37 (±12)	36 (±14)	39 (±13)
	Range %	70 (7-77)	51 (14-65)	63 (14-77)	65 (12-77)	77 (10-87)	70 (13-83)
Middle Third	Mean % (±SD)	37 (±12)	37 (±11)	39 (±11)	39 (±10)	38 (±14)	36 (±12)
	Range %	59 (9-68)	60 (7-67)	67 (9-76)	69 (13-82)	87 (0-87)	67 (0-67)
Lower Third	Mean % (±SD)	11 (±10)	11 (±9)	11 (±9)	11 (±9)	9 (±9)	10 (±9)
	Range %	37 (0-37)	36 (0-36)	38 (0-38)	44 (0-44)	36 (0-36)	35 (0-35)

[a] N = 105.
Note: No differences in designations or ratings across years are statistically significant.

respond to potential threats to validity in national rating trends.

Conclusions

Despite explicit guidelines for the distribution of global assessment ratings in the EM SLOE, there is high variability in institutional rating practices.

References

1. 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
2. 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. *Acad Emerg Med*. 2013;20(9):926-932. doi:10.1111/acem.12214
3. American Orthopaedic Association. Electronic Standardized Letter of Recommendation (eSLOR). Accessed March 16, 2023. https://www.aoassn.org/standardized-electronic-letter-of-recommendation-eslor/
4. Kaffenberger JA, Mosser J, Lee G, et al. A retrospective analysis comparing the new standardized letter of recommendation in dermatology with the classic narrative letter of recommendation. *J Clin Aesthet Dermatol*. 2016;9(9):36-42.
5. 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
6. Otolaryngology Program Directors Organization. Otolaryngology—head & neck surgery residency standardized letter of recommendation form. Accessed March 16, 2023. https://cdn.ymaws.com/opdo-hns.org/resource/resmgr/oto_standardized_letter_of_r.pdf
7. American Orthopaedic Association. Council of Orthopaedic Residency Directors standardized letter of recommendation form. Accessed March 16, 2023. https://aoaproduction.wpenginepowered.com/wp-content/uploads/2020/12/AOA-CORD-SLOR-2017-V8162.pdf
8. American Council of Academic Plastic Surgeons. Plastic surgery residency recommendation form: independent applicant. Accessed March 16, 2023. https://acaplasticsurgeons.org/multimedia/files/2019-ACAPS-Plastic-Surgery-Recommendation-Form_Independent.pdf

9. Association of Professors of Dermatology. Dermatology standardized letter of recommendation. Accessed March 16, 2023. https://www.dermatologyprofessors.org/files/2013%20Annual%20Meeting/DERMATOLOGY_SLOR.pdf
10. Pacana MJ, Thier ZT, Jackson JB 3rd, Koon DE Jr, Grabowski G. More than one-third of orthopaedic applicants are in the top 10%: the standardized letter of recommendation and evaluation of orthopaedic resident applicants. *Clin Orthop Relat Res*. 2021;479(8):1703-1708. doi:10.1097/CORR.0000000000001707
11. Abidi NY, Wanner B, Brown M, et al. Characterization of the 2019 micrographic surgery and dermatologic oncology standardized letter of recommendation. *Dermatol Surg*. 2021;47(3):327-332. doi:10.1097/DSS.0000000000002812
12. 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.
13. Faherty A, Counihan T, Kropmans T, Finn Y. Inter-rater reliability in clinical assessments: do examiner pairings influence candidate ratings? *BMC Med Educ*. 2020;20(1):147. doi:10.1186/s12909-020-02009-4

Eric Shappell, MD, MHPE, is Associate Program Director and Assistant Professor, Department of Emergency Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA; **Cullen Hegarty, MD,** is Program Director, HealthPartners Institute/Regions Hospital Emergency Medicine Residency Program, and Professor, Department of Emergency Medicine, University of Minnesota Medical School, Minneapolis, Minnesota, USA; **Sharon Bord, MD,** is Clerkship and Sub-Internship Director and Assistant Professor, Department of Emergency Medicine, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; and **Daniel J. Egan, MD,** is Program Director and Associate Professor, Department of Emergency Medicine, Massachusetts General Hospital, Brigham and Women's Hospital, and Harvard Medical School, Boston, Massachusetts, USA.

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

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

The authors would like to acknowledge the Council of Residency Directors in Emergency Medicine Board of Directors and Kevin Hamilton for providing access to the data for this study.

Corresponding author: Eric Shappell, MD, MHPE, Massachusetts General Hospital, Boston, Massachusetts, USA, eshappell@mgh.harvard.edu

Received March 31, 2023; revisions received August 1, 2023, and April 1, 2024; accepted April 2, 2024.