Title

A Retrospective Analysis of Medical Student Performance Evaluations, 2014-2020: Recommend with Reservations.

Permalink

https://escholarship.org/uc/item/4zj337g9

Journal

Journal of General Internal Medicine, 37(9)

Authors

Tisdale, Rebecca
Filsoof, Amy
Singhal, Surbhi
et al.

Publication Date

2022-07-01

DOI

10.1007/s11606-022-07502-8

Peer reviewed

A Retrospective Analysis of Medical Student Performance Evaluations, 2014–2020: Recommend with Reservations

Rebecca L. Tisdale, MD, MPA[1,2] , Amy R. Filsoof, MD[3], Surbhi Singhal, MD[3], Wendy Cáceres, MD[3], Shriram Nallamshetty, MD[1,3], Angela J. Rogers, MD[3], Abraham C. Verghese, MD[3], Robert A. Harrington, MD[3], and Ronald M. Witteles, MD[3]

[1]Health Services Research and Development (HSR&D), Center for Innovation to Implementation (Ci2i), VA Palo Alto Health Care System (152-MPD), 795 Willow Road, Building 324, Menlo Park, CA, USA; [2]Department of Health Policy, Stanford University, Stanford, USA; [3]Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.

BACKGROUND: The Medical Student Performance Evaluations (MSPE) is a cornerstone of residency applications. Little is known regarding adherence to Association of American Medical Colleges (AAMC) MSPE recommendations and longitudinal changes in MSPE content.

OBJECTIVES: Evaluate current MSPE quality and longitudinal changes in MSPE and grading practices.

DESIGN: Retrospective analysis.

PARTICIPANTS: Students from all Liaison Committee on Medical Education (LCME)-accredited medical schools from which the Stanford University Internal Medicine residency program received applications between 2014–2015 and 2019–2020.

MAIN MEASURES: Inclusion of key words to describe applicant performance and metrics thereof, including distribution among students and key word assignment explanation; inclusion of clerkship grades, grade distributions, and grade composition; and evidence of grade inflation over time.

KEY RESULTS: MSPE comprehensiveness varied substantially among the 149 schools analyzed. In total, 25% of schools provided complete information consistent with AAMC recommendations regarding key word/categorization of medical students and clerkship grades in 2019–2020. Seventy-seven distinct key word terms appeared across the 139 schools examined in 2019–2020. Grading practices markedly varied, with 2–83% of students receiving the top internal medicine clerkship grade depending on the year and school. Individual schools frequently changed key word and grading practices, with 33% and 18% of schools starting and/or stopping use of key words and grades, respectively. Significant grade inflation occurred over the 6-year study period, with an average 14% relative increase in the proportion of students receiving top clerkship grades.

CONCLUSIONS: A minority of schools complies with AAMC MSPE guidelines, and MSPEs are inconsistent across time and schools. These practices may impair evaluation of students within and between schools.

INTRODUCTION

The medical student performance evaluation (MSPE), a summary letter written by medical school faculty on behalf of a medical student applying to residency training programs, is a core component of a residency application. The MSPE typically contains an applicant's clinical and preclinical grading information, as well as a statement of summative performance over the applicant's time in medical school. This compilation of data weighs heavily in residency program directors' decisions regarding which candidates to invite for a limited number of interview slots and, ultimately, how to rank applicants.[1–3]

Despite this central role in residency ranking decisions, our institution[4] and others[5] have documented that prior to 2016, the MSPE was frequently incomplete and difficult to interpret. In response to these quality issues, an Association of American Medical Colleges (AAMC) task force released recommendations in 2016 to clarify and standardize data presented in the MSPEs (Supplemental Appendix I) for Liaison Committee on Medical Education (LCME)-accredited medical schools.[6] The recommendations called for inclusion of information regarding "how final grades and comparative data are derived" and to "provide school-wide comparisons if using the final 'adjective' or 'overall rating.'" While criticized as insufficient to address the deficits of the existing MSPE,[7,8] these guidelines were nonetheless expected to improve MSPE quality.

While one evaluation of MSPEs[9] and a survey of internal medicine residency program directors[10] found MSPEs improved by subjective and some objective measures since the 2016 updates, other studies reported substantial variability in MSPE content that limited their utility.[11,12] To our knowledge, no longitudinal study of MSPEs has been conducted to date. In this study, we characterize the current and recent state of the

MSPE with a year-by-year systematic analysis of MSPEs from all schools with applications submitted to our residency program between the 2014–2015 and 2019–2020 application cycles, aiming to describe trends in guideline-concordance, grading patterns, and MSPE practices over time.

METHODS

We conducted this study at Stanford University School of Medicine in March–July 2020. The Stanford University institutional review board approved this study.

MSPE Selection

For each academic year between 2014–2015 and 2019–2020, a randomly chosen MSPE from each school with at least one applicant to the Stanford Internal Medicine Residency Program was reviewed. Characteristics assessed were those consistently reported among all MSPEs from a given school in a given year; therefore, the selection of which student's MSPE from a given school to review for each year would not affect the analysis.

MSPE Data Collection

Three members of our research team (R.T., A.F., S.S.) assessed MSPEs for inclusion of a pre-specified set of characteristics of key word and clerkship grade practices as previously described.[4] These characteristics were defined based on 2016 AAMC updated recommendations for MSPE preparation.[6] Each MSPE was reviewed unblinded by a single author, with an independent review by a second author where the initial evaluating author found the MSPE to be unclear. The two authors then rigorously reexamined the MSPE in question before reaching consensus. To ensure consistency in data collection, the same team member assessed the MSPE for each school stratified alphabetically across the study period: Albert Einstein College of Medicine to Northwestern University Feinberg School of Medicine (S.S.), Oakland University Beaumont School of Medicine to University of Illinois Carle Illinois College of Medicine (R.T.), and University of Iowa Carver College of Medicine to Yale School of Medicine (A.F.). Schools were considered as separate institutions when they provided distinct data (e.g., grade distributions), even when they were grouped into single institutions by the LCME.

For key words, we noted whether key words were used at all, and if so, whether a key word key, explanation of how students were placed in key word categories, and/or distribution of key words among the student's class were present; the number of key words used; the individual words used; and whether any explanation was clearly stated. If a distribution was given, we recorded those figures. For clerkship grades, we similarly noted whether grades were used, whether a distribution of grades and/or final grade breakdown (e.g., % clinical,

% shelf exam, % OSCE) were provided, and the number of possible grades and recorded the grade distribution if given.

Data Analysis

We calculated descriptive statistics for all metrics, as well as the number of distinct key words and clerkship grades.

To assess changes over time in key word and grade practices, we compared the earliest and latest MSPE data available in our dataset for all schools with applications received in at least 2 years during the study period. When feasible, we used 2014–2015 and 2019–2020 figures, but in some cases this was not possible as our program did not receive an application from every school in every year. We excluded schools from longitudinal comparison if only 1–2 years of data were available. For schools with at least 3 years of data but where 2014–2015 or 2019–2020 data were unavailable, we used data from the closest year where data were available (e.g., for the initial year, we used data from 2015–2016 if 2014–2015 data were unavailable, or 2016–2017 if both 2014–2015 and 2015–2016 data were unavailable; for the final year, we used 2018–2019 data if 2019–2020 data were unavailable, or 2017–2018 data if both 2018–2019 and 2019–2020 data were unavailable). We used these earliest and latest values to calculate the change in proportion of students receiving the highest key word and grade for each clerkship and compared these initial and final values by school with paired t-tests. We also compared the proportion of schools using key words and using grades with Fisher's exact tests. All statistical tests were performed in R (R Foundation for Statistical Computing, Vienna, Austria).

RESULTS

Our sample totaled 149 schools over the 6-year study period, including 139 schools for the 2019–2020 application cycle. Here, we report the current practices from the 2019–2020 MSPEs, then characterize the changes in MSPEs over the study period.

2019–2020

Main outcomes for key word and grade characteristics are shown in Table 1. In total, 25% of schools met AAMC recommendations for MSPE clarity and completeness.

Key Words. Seventy-seven distinct key words were used in 2019–2020 (Table 1). Key word frequency and percentile ranges for the most commonly used words are depicted in Table 2; the most common schema was "Good, Very Good, Excellent, Outstanding" as has been previously reported.[10]

However, comparing individual schools in 2019–2020, it was not uncommon to see this key word schema shifted down—i.e., with "Very Good" as the bottom category, with or without a different word after "Outstanding" as the top category (e.g., "Exceptional" or "Distinguished"). Similarly, the distribution was at times shifted up (e.g., with

Table 1 2019–2020 Main Outcomes

s	N (%)
Total schools	139
Complete information on key words and clerkship grades*	35 (25%)
Complete key word information, incomplete clerkship grade information	15 (11%)
Incomplete key word information, complete clerkship grade information	51 (37%)
Incomplete key word and clerkship grade information	38 (27%)
Any key word data	95 (68%)
Key word key and some context[†]	89 (64%)
Complete key word data[‡]	50 (36%)
Number of distinct key words[§]	77
Percentage of students awarded top key word designation, when applicable	Median: 25% of students Range: 1%-50%
Any grade data	134 (96%)
Clerkship grade distributions	125 (89%)
Complete clerkship grade data[‖]	86 (61%)
Percentage of students awarded top internal medicine clerkship grade	Median: 34% of students Range: 2–75%

*Clerkship grade distributions with clear explanation of grade breakdown; key word distributions/key with clear explanation of how students assigned to key word category
[†]Includes key word and distribution among categories, plus a key word key and/or some explanation of how students are assigned to categories
[‡]Includes key word key with clear explanation of how students are placed in key word categories and distribution among categories
[§]E.g., "Good/Recommend" was counted as equivalent to "Good," as was "Good Candidate"
[‖]Includes clerkship grade distributions with clear explanation of grade breakdown

"Satisfactory" added in the bottom position, below "Good"). For the 55 schools in the 2019–2020 sample using the word "Outstanding," it was the top word in 35 (64%), but for the remaining 20 schools it denoted the second group or lower. A student called "Outstanding" could potentially be in any percentile from 32 to 99, depending on the school; the label "Superior" would place that student somewhere between the 21st and 99th percentiles, and a "Very Good" student might land anywhere from the 0th to 71st percentiles. No school included any key words with negative connotations, though the word "Satisfactory" always represented the bottom category when used.

We considered hypothetical 90th and 10th percentile students. A 90th percentile student in 2019–2020 would have no key word 30% of the time and would be considered "Outstanding" 24% of the time, "Exceptional" 8% of the time, and in another mathematically defined top grouping (e.g., top quartile, top quintile) 19% of the time. A 10th percentile student in 2019–2020 would have no key word 30% of the time and a mathematically defined metric 19% of the time, with the remaining key words dominated by "Good" (16%) and "Very good" (12%). Interestingly, over a quarter (27%) of students in the 10th percentile would not be in the bottom key word category, largely due to the 8 of the 139 schools in the sample including zero students in their bottom category; 5% would not be in either the bottom or second-from-bottom categories, and 4% of the time they would be assigned "Excellent."

Clerkship Grades. The variability in key word data extended to clerkship grades. Figure 1 displays the distribution in percentage of students receiving the top internal medicine clerkship grades among medical schools. The median percentage of students awarded this top grade was 34% with a range from 2 to 75% (Table 1).

The mean, median, and mode number of grades used per school was 4 (Supplemental Figure 1), though there was significant variability across schools.

Changes over Time

The number of schools analyzed per year totaled 149 distinct schools over the study period (Supplemental Table 1). Of these, 8 schools had only 1–2 years of data over the study period and were thus excluded from the comparison of change between 2014–2015 and 2019–2020 academic years (N=141 schools included in longitudinal analysis). Of these 141 schools, 124 schools (87.9%) had 2014–2015 and 2019–2020 MSPEs, with the remaining 17 schools' data coming from previous and/or subsequent years as detailed in the methods.

Key Words. The overall proportion of schools using key words did not change significantly between 2014–2015 and 2019–2020 (66% and 68%, respectively; $p = 0.7$). Clarity among those using key words increased: the proportion of key word-using schools providing a distribution increased significantly (83% versus 94% of key word users; $p = 0.02$), as did the proportion of key word-using schools providing a clear explanation of how students received their key word designation (31% versus 55% of key word users; $p = 0.002$). The timing of this improvement in percentage of schools providing a key word distribution, shown in Supplemental Figure 2, suggests that it likely represents a response to the 2016 AAMC recommendations. Importantly, even with these improvements, only slightly more than half of key word users at the end of the study provided clear explanations consistent with AAMC recommendations.[6]

No change occurred in the percentage of students who received the top key word over the study period: 21% of students received this top word at the beginning of the study, and 22% at the end. However, practices varied widely across schools over the timeframe examined. Among the 131 schools with at least two data points for students in the top word category over the study period, some schools kept roughly similar distributions across the study period, but many others displayed large-magnitude changes in the percent of students awarded the top key word—ranging from decreasing that percentage by 22 absolute percentage points to increasing it by 31 absolute percentage points.

Just as notable was the frequency of change in whether an individual school used key words, henceforth referred to as *churn*, across the study period (Table 3). For those eleven schools both introducing and discontinuing use of key words over the study period, this could mean either that they used key

Table 2 Most Common Key Words and Characteristics, 2019–2020 (N = 97)

Key word or characteristic	Frequency (%)	Percentile range	Frequency representing top word (% of overall frequency)	Frequency representing bottom word (% of overall frequency)
Outstanding	55 (56.7)	32–99	35 (63.6)	0
Very good	54 (55.7)	0–71	0	8 (14.8)
Excellent	51 (52.6)	0–87	0	0
Good	42 (43.3)	0–31	0	36 (85.7)
Mathematical term*	25 (25.8)	N/A	N/A	N/A
Exceptional	15 (15.5)	65–99	13 (86.7)	0
Satisfactory	10 (10.3)	0–43	0	10 (100.0)
Superior	9 (9.3)	21–99	3 (33.3)	0
Distinguished	4 (4.1)	76–99	4 (100.0)	0
Use of "Good/Very Good/Excellent/Outstanding" schema	23 (23.7)	N/A	N/A	N/A
Zero students in bottom category	8 (8.2)	N/A	N/A	N/A

*E.g., quartile, quintile

words at the outset, then discontinued using them, then restarted, or that they did not use key words at the outset, introduced them, then discontinued using them during the study period. Over one-third of schools made at least one change in use of key words over the study period.

Clerkship Grades. The majority of schools (>95%) in all study years used grades, and of those using grades, the majority provided grade distributions (Supplemental Table 2). Initially, a minority of schools provided information regarding grade breakdown, but this increased significantly from 11.2% of schools that used grades providing grade breakdowns to 60.5% of schools that used grades providing grade breakdowns between 2016–2017 and 2017–2018, likely in response to AAMC guideline changes.[6]

Churn was present, but to a lesser degree than for key words. Nine schools never used grades, eleven schools began to use grades, and eleven schools stopped using grades over the course of the study period; five additional schools both started and stopped using grades (in either order) over the study period, for a total of 18% of sample schools changing the use of grades.

To assess for the presence of grade inflation, we examined the change in proportions of students receiving the top grade in all core clerkships (Supplemental Table 3). The percentage of students receiving the top clerkship grade varied widely across schools; for the internal medicine clerkship, for example, 2–83% of students received the top internal medicine clerkship grade depending on the year and school.

Figure 2 shows the change from the initial average percentage of students receiving the top grade across clerkships. These relative changes ranged from 11% inflation over the study period for internal medicine to 17% for pediatrics, neurology, and obstetrics/gynecology, for an average of 14% inflation across all clerkships. Alongside this trend of grade inflation, a significant minority of schools demonstrated deflation in their grades over the study period, with negative changes as large as −61 percentage points (Supplemental Table 3). Variability was high, as reflected in the large standard deviations for change in students receiving the top grade for all clerkships.
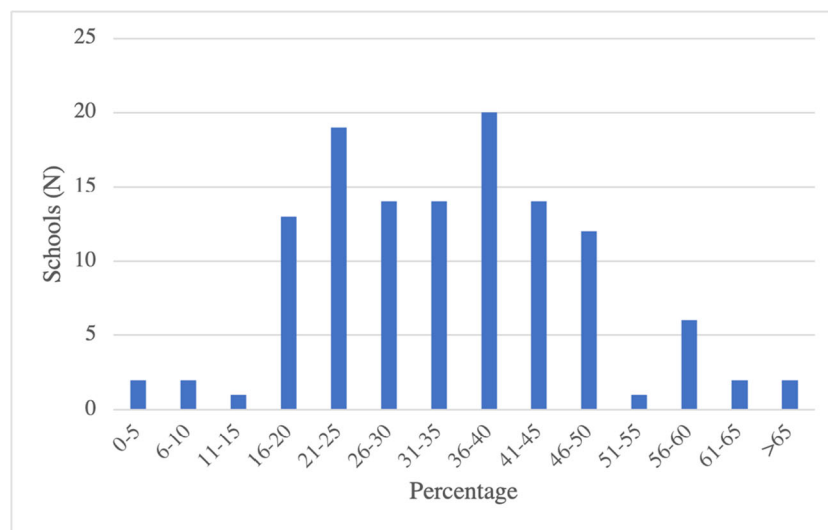

Fig. 1 Distribution of percent of students receiving top medicine clerkship grade, 2019–2020

Table 3 Churn in Key Word Usage (2014–2020)

Category	N (total = 146*)
Ever used key word	115
Never used key word	31
Consistently used key word	65
Introduced use of key word	27
Discontinued use of key word	12
Both introduced and discontinued use of key word	11

Only schools with at least two observations included; three schools had only one observation each. For schools with missing data, we assumed patterns before earliest observation and after latest observation continued; e.g., for a school with observations only in 2017–2018 and 2019–2020, but where key words were used in both those years, this was counted as a school that likely used key words throughout

DISCUSSION

In this systematic analysis of MSPEs from 149 medical schools between 2014–2015 and 2019–2020, we find great variation in MSPE practices both within and across institutions and marked grade inflation over time, as well as overall poor adherence to AAMC recommendations.

Variation in Practices

In both our cross-sectional 2019–2020 results and over time, a striking degree of variability within and across schools emerges: nearly 80 distinct key word terms in a single year and a correspondingly large number of ranking systems, a range of 2 to 83% of students receiving the top internal medicine clerkship grade, and more than a third of schools changing whether they used key words over the study period.

This variability makes comparative evaluation of students difficult, as exemplified by our exploration of key words. For example, "superior" only represents the top key word in 33.3% of cases, whereas "outstanding" represents the top key word in 63.6% of cases. This discrepancy by school in key word categorization for a given caliber of student underscores the arbitrariness of these terms. Even seemingly straightforward mathematical descriptions are suspect. One school placed 10% of students in a "top ten" category implying that this represented the top ten *students*, in a published class size of 220; another's bottom "quartile" contained just 4% of students.

Furthermore, even an experienced program director could be mistaken, given the change within a school over time (e.g., in whether a school is a relatively "hard" or "easy" grader). This lack of standardization across schools may disproportionately disadvantage smaller or lesser-known medical schools that send fewer future residents to a given residency program, since program directors are less likely to have direct experience with those schools' students on which to base their assessments of future students, as well as smaller residency programs receiving fewer applications for comparison from each given medical school.

Grade Inflation

We identified significant clerkship grade inflation across all core clerkships over a relatively short study period. Grade inflation in itself causes a trend toward narrower grade distributions that may complicate differentiation among students.[14,15] However, it is the variability discussed above—in this case, occurring simultaneously with grade deflation seen at some schools—that impedes program directors' efforts to interpret grades. This variability means that program directors cannot assume that all schools inflate grades year over year, or that school X tends to be an easy or hard grader, and interpret a given student's grades in that context. This further emphasizes the need for data reporting to be standardized within and between medical schools.[8,13]

Adherence to AAMC Recommendations

The changes in AAMC guidelines for MSPEs released in 2016 stressed inclusion of the derivation of final grades and comparative data, as well as school-wide comparisons. We note some changes that were likely in response to these updated recommendations: the number of schools providing grade breakdown information increased in the years immediately following guideline release, as did the number of schools including a distribution of key words/rankings and an explanation of how students received their key word designation. In our 2016 MSPE analysis, we found that 51% of schools met AAMC guidelines for the MSPE in 2013–2014;[4] completeness of 2019–2020 MSPEs using the former criteria shows 59% of schools meeting those old guidelines, suggesting some improvement.

Still, it is disappointing that nearly 5 years after new AAMC guidelines were released, just 25% of schools provided complete key word and grade information based on these updated recommendations. Whether the recommendations went far enough has already been called into question,[7,8,10] but it appears that implementation of even these relatively conservative recommendations has been limited.

Bias in Data Sources

Any discussion of data forming the basis of residency interview and ranking decisions is incomplete without consideration of the level of bias intrinsic to that data. In particular, increasing recognition of deep bias in clerkship grades[16,17] rightly raises the question of whether more schools should move away from using clerkship grades at all. Indeed, many prestigious institutions, including UCSF, Harvard, and Vanderbilt, have discontinued the use of clerkship grades in recent years, citing concerns over imprecision and bias.[18] A growing literature demonstrates the extent to which bias disproportionately affects underrepresented minority students, both in clerkship grades and in the descriptive words used throughout an MSPE.[17,19]
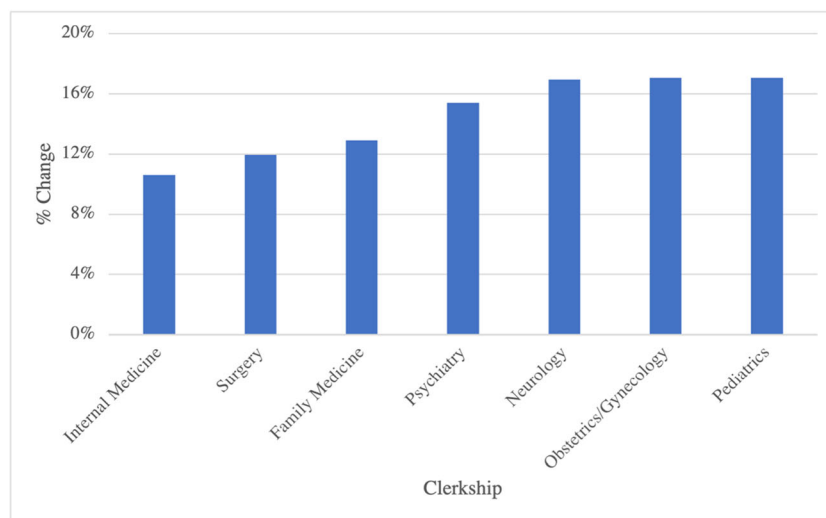
Fig. 2 Percent change from baseline percentage of students receiving top grade from the start to the end of the study (2014–2020), by clerkship

We are sympathetic to this argument and committed to rooting out and addressing bias all along the medical training pathway. While we note that clerkship grades may be highly imperfect and biased, the same is true of the other data sources program directors use for decision-making. With the recent conversion of USMLE Step 1 to reporting as pass/fail[20] and increasingly vague MSPEs, residency selection committees are left to rely on fewer data points (which in themselves are biased) when forming an overall impression of a candidate.

Given the limitations of the MSPE in its current form as detailed above and the erosion of other data sources, what information ought program directors use to make the necessary decisions about interviewing and ranking? Frameworks for improving inclusivity of resident recruitment provide a useful starting point,[21] but data is still required. Letters of recommendation are nearly universally glowing; one-page personal statements give a limited snapshot; research and work experience can look quite similar across the majority of medical students.[22] For those invited for an interview, interviews may provide the most biased of all the available data[23–26]—perhaps surpassed only by a phone call from a program director to a medical school contact for assistance in understanding which applicants really would be a "good fit," at least from that given contact's (by definition biased) perspective.

Limitations

Our sample size over the study period was almost, but not fully, complete; we did not receive applications from 6 of the 155 LCME-accredited schools (4%), and they were therefore not included in this study. For some schools, we received applications for only a subset of years of the study period (e.g., a student applied from that school in 2014–2015 and 2016–2017 but not 2015–2016). As this study focused on adherence to AAMC recommendations, which apply only to these LCME-accredited schools, we did not analyze

applications from osteopathic schools or international medical schools. The three members of our research team who assessed MSPEs were unblinded, and we did not perform formal inter-rater reliability assessment.

Conclusions

Years after updated AAMC MSPE guidelines were released, a minority of schools comply with these recommendations, and MSPEs are neither transparent nor consistent across time and schools. Significant grade inflation occurred over the study period, and variation in grade inflation/deflation was observed between schools. Current MSPE practices impair comparison of students within and between medical schools, leaving programs to rely on other biased sources of applicant data for interview offers and ranking decisions. Whichever characteristics a given residency program most values in its potential applicants, this lack of transparency and consistency in reporting of medical student performance impedes its search—and has the potential to especially disadvantage students from smaller or less academically prestigious schools without a longer track record of matching students to particular residency programs. Improved compliance with AAMC MSPE guidelines could improve the residency application process for applicants and programs alike.

Corresponding Author: *Rebecca L. Tisdale, MD, MPA; Health Services Research and Development (HSR&D), Center for Innovation to Implementation (Ci2i), VA Palo Alto Health Care System (152-MPD), 795 Willow Road, Building 324, Menlo Park, CA 94025, USA (e-mail: rtisdale@stanford.edu).*

Declarations:

Conflict of Interest: *All authors report no relevant conflicts of interest.*

REFERENCES

1. 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 Ed.* 2019 Jun;11(3):268-73.

2. Negaard M, Assimacopoulos E, Harland K, Van Heukelom J. Emergency medicine residency selection criteria: an update and comparison. *AEM Ed and Training.* 2018 Apr;2(2):146-53.

3. Association of American Medical Colleges. "Medical Student Performance Evaluation (MSPE)." Accessed 17 Dec 2020 at https://www.aamc.org/professional-development/affinity-groups/gsa/medical-student-performance-evaluation.

4. Hom J, Richman I, Hall P, et al. The State of Medical Student Performance Evaluations: Improved Transparency or Continued Obfuscation? *Acad Med.* 2016;91(11):1534–9.

5. Boysen-Osborn M, Yanuck J, Mattson J, et al. Who to interview? Low adherence by US medical schools to medical student performance evaluation format makes resident selection difficult. *West J Emerg Med.* 2017;18(1):50.

6. Association of American Medical Colleges. Recommendations for revising the medical student performance evaluation (MSPE). May 2017. Downloaded 23 Nov 2020 from https://www.aamc.org/media/23311/download.

7. Hauer KE, Giang D, Kapp ME, Sterling R. Standardization in the MSPE: key tensions for learners, schools, and residency programs. *Acad Med.* 2020 Dec 29;96(1):44-9.

8. Catalanotti JS, Andolsek KM, Berger JS. The AAMC medical student performance evaluation task force recommendations: do they go far enough?. *J Grad Med Ed.* 2017;9(1):1-4.

9. Hook L, Salami AC, Diaz T, Friend KE, Fathalizadeh A, Joshi ART. The revised 2017 MSPE: better, but not "outstanding". *J Surg Educ.* 2018;75(6):e107–e111.

10. Brenner JM, Arayssi T, Conigliaro RL, Friedman K. The revised Medical School Performance Evaluation: does it meet the needs of its readers?. *J Grad Med Ed.* 2019;11(4):475-8.

11. Brenner JM, Bird JB, Brennder J, et al. Current State of the Medical Student Performance Evaluation: A Tool for Reflection for Residency Programs. *J Grad Med Ed.* 2021;13(4):576-580

12. Thimm TN, Kiefer CS, Aloi MS, et al. Does the Medical Student Performance Evaluation Change the Decision to Invite Residency Applicants? *West J Emerg Med.* 2021;22(5):1102-1109.

13. Boysen Osborn MB, Mattson J, Yanuck J, et al. Ranking Practice Variability in the Medical Student Performance Evaluation: So Bad, It's "Good". *Acad Med.* 2016;91(11):1540.

14. Schiel KZ, Everard KM. Grade inflation in the family medicine clerkship. *Fam Med.* 2019;51(10):806-10.

15. Grover S, Swisher-McClure S, Sosnowicz S, et al. Grade inflation in medical student radiation oncology clerkships: missed opportunities for feedback?. *Int J Rad Onc* Biol* Physics.* 2015;92(4):740-4.

16. Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. *Acad Med.* 2012;87(8):1070–1076.

17. 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;94:562–569.

18. Hauer KE, Lucey CR. Core clerkship grading: the illusion of objectivity. *Acad Med.* 2019;94(4):469-72.

19. 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.

20. United States Medical Licensing Examination (USMLE). "Change to pass/fail score reporting for Step 1". Accessed 17 Mar 2021 from https://www.usmle.org/incus/.

21. Gonzaga AM, Appiah-Pippim J, Onumah CM, Yialamas MA. A framework for inclusive graduate medical education recruitment strategies: meeting the ACGME standard for a diverse and inclusive workforce. *Acad Med.* 2020 May 1;95(5):710-6.

22. Green M, Jones P, Thomas JX Jr. Selection criteria for residency: Results of a national program directors survey. *Acad Med.* 2009;84(3):362-7.

23. Quintero AJ, Segal LS, King TS, Black KP. The personal interview: assessing the potential for personality similarity to bias the selection of orthopaedic residents. *Acad Med.* 2009;84(10):1364-72.

24. Stephenson-Famy A, Houmard BS, Oberoi S, Manyak A, Chiang S, Kim S. Use of the interview in resident candidate selection: a review of the literature. *J Grad Med Ed.* 2015;7(4):539-48.

25. Burkhardt JC. What Can We Learn From Resident Selection Interviews?. *J Grad Med Ed.* 2015;7(4):673-5.

26. Rivera L. Hiring as cultural matching: the case of elite professional service firms. *Am Sociol Rev.* 2012;77(6):999–1022.