Global Surgical Education - Journal of the Association for Surgical Education
https://doi.org/10.1007/s44186-022-00047-8

ORIGINAL ARTICLE

Measuring what matters: identifying assessments that reflect learning on the core surgical clerkship

Matthew F. Mikulski[1,2] · Ziv Beckerman[1,2] · Zachary L. Jacques[1] · Madison Terzo[3] · Kimberly M. Brown[1]

Received: 20 May 2022 / Revised: 2 September 2022 / Accepted: 10 September 2022

Abstract

Purpose There are various assessments used during the core surgical clerkship (CSC), each of which may be influenced by factors external to the CSC or have inherent biases from an equity lens. In particular, the National Board of Medical Examiners' Clinical Subject Exams ("Shelf") is used heavily and may not reflect clerkship curriculum or clinical learning.

Methods This is a retrospective review of medical student characteristics and assessments during the CSC from July 2017-June 2021. Assessment methods included: subjective Clinical Performance Assessments (CPA), Shelf, Objective Structured Clinical Examinations, and a short-answer in-house examination (IHE) culminating in a Final Grade (FG) of Honors/Pass/Fail. A Shelf score threshold for Honors was added in academic years 2020–2021. Descriptive, univariate, and multivariable logistic and linear regression statistics were utilized.

Results We reviewed records of 192 students. Of these, 107 (55.7%) were female, median age was 24 [IQR: 23–26] years, and most were White/Caucasian ($N=106$, 55.2%). Univariate analysis showed the number of Exceeds Expectations obtained on CPA to be influenced by surgical subspecialty taken ($p=0.013$) and academic year ($p<0.001$). Shelf was influenced by students' race ($p=0.009$), timing of CSC before or after Internal Medicine (67.9 ± 7.3 vs 72.9 ± 7.1, $p<0.001$), and Term taken (increasing from 66.0 ± 8.7 to 73.4 ± 7.5, $p<0.001$). IHE scores did not have any external associations. After adjustment with multivariable logistic and linear regressions, CPA and IHE did not have external associations, but higher scores were obtained on Shelf exam in Terms 3, 5, and 6 (by 4.62 [95% CI 0.86–8.37], 4.92 [95% CI 0.53–9.31], and 7.56 [95% CI 2.81–12.31] points, respectively. Odds of FG honors were lower when Shelf threshold was implemented (OR 0.17 [95% CI 0.06–0.50]), and increased as students got older (OR 1.14 [95% CI 1.01–1.30]) or on specific subspecialties, such as vascular surgery (OR 7.06 [95% CI 1.21–41.26]).

Conclusions The Shelf is substantially influenced by temporal associations across Terms and timing in relation to other clerkships, such as Internal Medicine. An IHE reflective of a clerkship's specified curriculum may be a more equitable summative assessment of the learning that occurs from the CSC curriculum, with fewer biases or influences external to the CSC.

Keywords Medical student education · NBME subject examination · Disparities in education

Introduction

Traditional methods of assessing and grading medical students have come under close scrutiny in recent years, through the lenses of validity evidence, impacts on trainee well-being and equity. This is especially true within the core surgical clerkship (CSC), with its perceptions of increased clinical time demands [1], call requirements [2–5], variations in teaching modalities [2, 6] and differing experiences among surgical subspecialties [7]. The equity lens is a very compelling challenge to the status quo, with growing evidence that under-represented in medicine groups experience inequity at multiple levels including

Data from this manuscript were presented at the Association for Surgical Education 2022 Annual Meeting on May 3, 2022.

✉ Matthew F. Mikulski
matthew.mikulski@austin.utexas.edu

[1] Department of Surgery and Perioperative Care, Dell Medical School, The University of Texas at Austin, Austin, TX, USA

[2] Texas Center for Pediatric and Congenital Heart Disease, UT Health Austin and Dell Children's Medical Center, 4900 Mueller Blvd, Suite 3S.003, Austin, TX 78712, USA

[3] Dell Medical School, The University of Texas at Austin, Austin, TX, USA

Published online: 19 September 2022

clerkship grading, resulting in significant barriers to successfully matching into competitive residencies [8–12].

Clerkship grades are frequently derived from a combination of subjective assessments, such as attending and resident physician clinical evaluations, and quantitative assessments, such as the National Board of Medical Examiners (NBME) Subject Examination in Surgery ("Shelf"), structured oral examinations or Objective Structured Clinical Examination (OSCE) [9, 13], and institution-specific in-house examinations (IHE) [14, 15]. The Shelf, taken at the end of the core clerkships at many medical schools, is a multiple-choice examination that allows comparisons across institutions and can serve as preparation for USLME Step 2. It is scaled to have a mean of 70% with a standard deviation of 8 [16] and offers suggestions for passing and honors cut-offs when used as part of clerkship grading. The validity evidence for its use in clerkship grading is mixed, despite its near-ubiquitous presence [14, 17]. For example, previous evaluations of the Shelf have identified minimal or weak relationships between faculty evaluations and Shelf scores [18, 19], while others demonstrated strong positive associations [20]. In addition, timing of rotations has shown to have an impact on Shelf and United States Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge scores, with those taking Surgery before Internal Medicine (IM) performing better on the IM Shelf and USMLE Step 2 [21, 22]. Together, these observations suggest that the construct being measured by the Shelf exam is not knowledge acquired through the clerkship curriculum.

Oral examinations and OSCEs, while time- and resource-intensive, have been shown to correlate with final grades received on the CSC [13] and positively correlate with USMLE Step 1 scores [1]. Recent work has further demonstrated that a structured oral examination can help decrease racial grading differences in the CSC [9]. Finally, an institutional IHE, while not generalizable to other institutions, can be a useful assessment and has been shown to be comparable to the Shelf [15].

In addition to the concerns around validity evidence for the use of the various summative assessments used in clerkship grading, particularly in a multi-tiered grading structure, our evaluation of these assessments through the lens of educational equity is still evolving. This study seeks to evaluate the factors associated with assessment outcomes on the CSC at a new medical school with an innovative curriculum. We hypothesize that the Shelf is influenced by factors external to the curriculum goals of the CSC and may not be a valid summative assessment. We further hypothesize that an internally-created IHE reflective of curriculum goals is a more fair and equitable summative assessment.

Methods

This study was approved by the Institutional Review Board of the Dell Medical School (DMS) at The University of Texas at Austin, STUDY00002003, approved November 2, 2021. DMS opened its inaugural class in 2016 with the first students entering their clinical clerkships during the 2017–2018 academic year.

This is a single-institution, retrospective review of medical student performance metrics during their CSC from July 2017 to June 2021, corresponding to four Academic Years. Final Grade on the CSC was defined as Honors, Pass, or Fail. Assessment methods included: NBME Shelf examination, Clinical Performance Assessments (CPA), OSCEs, and a short-answer internally-created IHE. There were two "Eras" of grading: Era 1 corresponded to Academic Years 2017–2019, Era 2 being Academic Year 2020. In Era 1, students were required to achieve a passing score on the Shelf, corresponding to 5th percentile of national performance, but the Shelf score was not part of honors criteria; in Era 2, a Shelf score threshold to obtain a final grade of honors was implemented. The CPAs consisted of 17 individual competencies (see Supplemental Material). These 17 CPAs were each assessed by individual faculty and residents then a summative assessment was determined by the grading committee using the following designations: Exceeds Expectations (EE), Meets Expectations, or Marginal Performance. Each designation for each competency is described with anchoring text. It was necessary to obtain > 10 EE to quality for a final grade of Honors. OSCEs were performed and a passing grade was required to pass the CSC, but was not a consideration for Honors. The IHE was structured as a mixture of short-answer and multiple-choice questions. It was given at the beginning of the CSC (Pre-IHE) to obtain a baseline score and again at end of the rotation (Post-IHE) to assess progress. A student was required to pass the Post-IHE to Pass the CSC, but the score itself was not otherwise weighted as part of the final grade. The IHE was discontinued after Era 1.

CSC and student characteristics

Core clinical rotations are taken during the second year of medical school, prior to the USMLE Step 1 examination. Each Academic Year consisted of six Terms of clinical rotations divided into 8-week blocks. Students could rank their order of rotation preference, but could not always be given. Students were evenly distributed throughout the six Terms. The CSC was split into 4 weeks of Acute Care Surgery (ACS) and 4 weeks of a Subspecialty surgical

service, the order of which was generated at random. Students took ACS rotations at the same hospital. Students could request a specific Subspecialty but were ultimately assigned based on availability. Available Subspecialties included: Elective General Surgery, Surgical Oncology, Pediatric General Surgery, Congenital Heart Surgery, and Vascular Surgery. During the 2019 Academic Year with COVID-19 pandemic changes, a small number of students rotated in Otolaryngology, Orthopedic Surgery, Neurosurgery, and Plastic and Reconstructive Surgery. All students participated in the same structured didactic curriculum, as well as individual service-specific conferences.

Students were considered Under-represented in Medicine (URM) if their race/ethnicity as recorded at DMS student affairs was Hispanic, African American, Pacific Islander, or Hawaiian. They were not considered URM if White or Asian. Those unreported or listed multiracial by DMS were categorized as Indeterminate.

The Medical College Admission Test (MCAT), administered by the Association of American Medical Colleges, changed significantly in April 2015. With some students having taken each of the examinations, and without a distinct 1:1 comparison of scores between old and new renditions, students were grouped into tertiles based on their relation in the old or new MCAT versions. MCAT tertiles were identified as: 1st tertile corresponding to a MCAT score of 29–32 or 505–511 for old or new, respectively, 2nd tertile for scores 33–34 or 512–515, and 3rd tertile for scores of 35–39 or 516–526.

Statistical analysis

Descriptive statistics were utilized for all variables. Categorical variables are reported as *N* (%). Continuous variables are reported as mean ± standard deviation (SD) if normally distributed or median [interquartile range (IQR)] if non-normally distributed. Univariate analysis comprised unpaired student's *T*-test and two-way Analysis of Variance for normally distributed continuous data, and Wilcoxon signed-rank test and Kruskal-Wallace tests were performed for non-normally distributed continuous data. Multivariable logistic regression was utilized for categorical outcomes, multivariable linear regressions were utilized for continuous variables. All variables included were assessed for collinearity. All statistical tests were 2-tailed and a p-value < 0.05 was considered significant. All statistics were performed using R and R Studio [23].

Results

Study population

There were 192 students reviewed over the study period. Student characteristics are outlined in Table 1. In the CSC,

Table 1 Student characteristics

Factor	Students *N* (%)
Total	192
Student demographics	
Median age [IQR] (years)	24 [23–26]
Female	107 (55.7%)
Race	
White/Caucasian	106 (55.2%)
Hispanic	20 (10.4%)
Asian	36 (18.8%)
African American	12 (6.3%)
Multiracial	8 (4.2%)
Unreported/other	10 (5.2%)
Under-represented in Medicine	
URM	34 (17.7%)
Not URM	142 (74%)
Indeterminate	16 (8.3%)
CSC characteristics	
Academic year	
2017	50 (26%)
2018	47 (24.5%)
2019	49 (25.5%)
2020	46 (24%)
Term	
1	33 (17.2%)
2	33 (17.2%)
3	33 (17.2%)
4	31 (16.1%)
5	33 (17.2%)
6	29 (15.1%)
Surgery before IM	94 (49%)
ACS before subspecialty	100 (52.1%)
Subspecialty	
Elective general	54 (28.1%)
Surgical oncology	50 (26%)
Pediatric general	39 (20.3%)
Congenital heart	21 (10.9%)
Vascular	17 (8.9%)
Other	11 (5.7%)
MCAT tertiles	
1st (low)	73 (38%)
2nd (middle)	51 (26.6%)
3rd (high)	53 (27.6%)
Mean step 1 score (± standard deviation)	241.2 ± 14.2
Assessments	
Shelf	192 (100%)
OSCE	177 (92.2%)
IHE	131 (68.2%)
CPA	192 (100%)
Final grade: honors	95 (49.5%)

students were almost evenly distributed across Academic Years, Terms, as well as ACS taken before or after their subspecialty rotation and CSC before or after IM rotation. Students most commonly took Elective General Surgery as their subspecialty rotation ($N = 54$, 28.1%).

Clinical performance assessments

The distribution of EE accrued per student on the CPA followed a non-normal distribution, with peaks at each of the extremes (Fig. 1). There were no associations between student demographics and median CPA (Table 2). There was variability in CPA based on surgical Subspecialty ($p = 0.013$), with the highest CPA occurring among those taking Vascular Surgery (median 13 [IQR: 10–17]) and the lowest among Pediatric General Surgery (median 9 [IQR: 4–12]). There was additional variation seen by Academic Year ($p < 0.001$), with the lowest scores occurring in 2017 (median 8 [IQR: 3–12]) and the highest in 2018 (median 13 [IQR: 7–17]). After adjusting for age, sex, URM status, CSC before or after IM, ACS before or after surgical subspecialty, Era 1 vs Era 2, surgical subspecialty, Term, and for good test takers by accounting for MCAT tertiles and Step 1 scores by multivariable logistic regression (Table 4), there were no associations conferring increased odds of obtaining > 10 EE on the CPAs except Step 1 scores, in which there were 1.03 odds (95% CI 1.00–1.06) for every point increase on Step 1.

Shelf examination

Mean Shelf score was 70.4 ± 8.2. There was racial variation among Shelf scores ($p = 0.009$), with Whites scoring the highest (72.0 ± 7.9) and those Unreported/Other scoring the lowest (64.7 ± 9.8). Students who had the CSC before IM scored lower than those who took IM first (67.9 ± 8.5 vs 72.9 ± 7.1, $p < 0.001$). Shelf scores increased from Terms 1 to 6 ($p < 0.001$) (Fig. 2). Students with a final grade of Honors scored higher than those who achieved Pass/Fail (74.4 ± 6.1 vs 66.5 ± 8.1, $p < 0.001$). After adjustment (Table 4), there was no longer an association between the timing of the CSC and IM. Students improved by 0.25 (95% CI 0.17–0.33) points for each point increase in Step 1 score ($p < 0.001$). Students additionally had higher scores in Terms 3, 5, and 6 by 4.62 (95% CI 0.86–8.37, $p = 0.017$), 4.92 (95% CI 0.53–9.31, $p = 0.030$), and 7.56 (95% CI 2.81–12.31, $p = 0.002$) points, respectively.

Objective structured clinical examination

The OSCE was performed at the end of the CSC. Not all students had an OSCE examination ($N = 177$, 92.2%), with the 15 missing students corresponding to the onset of the COVID-19 pandemic. There was variability in OSCE score across Terms ($p = 0.010$) with an overall downward trend from Term 1 to 6 (Fig. 2). There was variability in OSCE across Academic Years ($p < 0.001$) with a steady decrease from 73.0 ± 9.0 in 2017 to 62.5 ± 9.7 in 2020. Those who

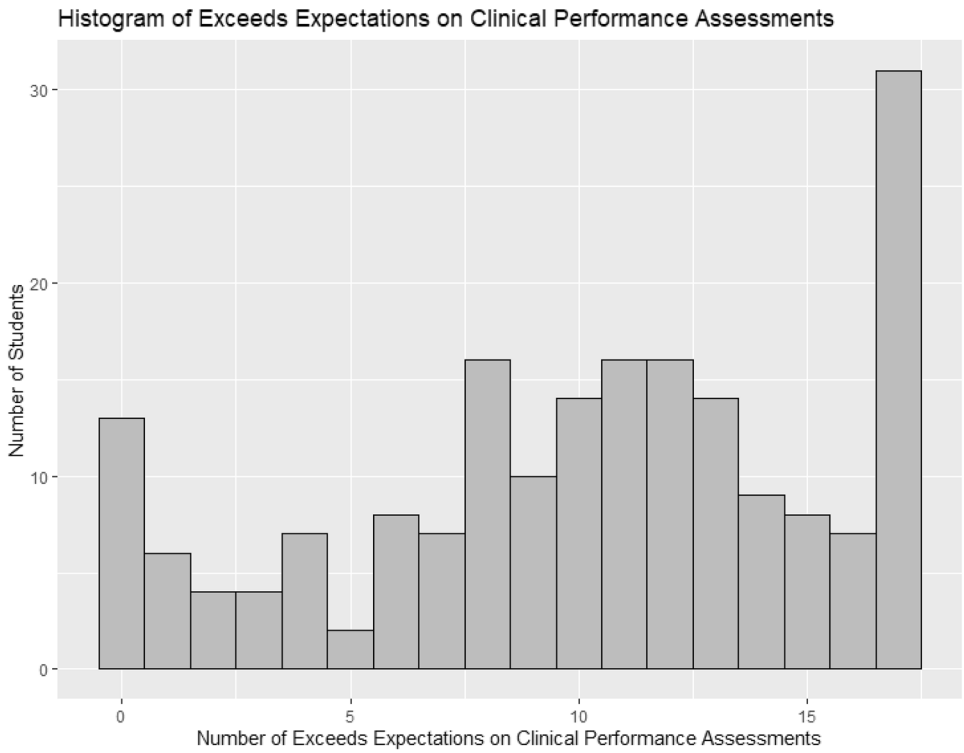

Fig. 1 Histogram of exceeds expectations on clinical performance assessments

Table 2 Univariate analyses of assessment methods

	Number of exceeds expectations on CPA (Median [IQR])	*p* value	Shelf (Mean ± SD)	*p* value	OSCE (Mean ± SD)	*p* value	IHE (Mean ± SD)	*p* value
Aggregate	11 [7–14]	–	70.4 ± 8.2	–	68.3 ± 9.9	–	74.8 ± 7.3	–
Female	11 [7–14]	0.711	69.8 ± 8.4	0.230	68.8 ± 10.5	0.484	75.8 ± 7.7	0.081
Male	11 [8–14]		71.2 ± 7.8		67.7 ± 9.2		73.6 ± 6.7	
Race								
White/Caucasian	11 [8–15]	0.627	72.0 ± 7.9	**0.009**	68.7 ± 10.0	0.317	75.9 ± 6.7	0.219
Hispanic	10 [4–15]		71.2 ± 6.5		69.1 ± 11.2		74.6 ± 7.8	
Asian	11 [8–13]		67.9 ± 8.7		67.4 ± 9.2		74.9 ± 6.9	
African American	7.5 [3.75 – 12.5]		66.8 ± 7.4		65.4 ± 12.1		70.4 ± 8.1	
Multiracial	8.5 [4.75 – 12.25]		71.8 ± 6.3		62.7 ± 5.9		72.0 ± 7.8	
Unreported/other	11 [6.75 – 12.75]		64.7 ± 9.8		73.4 ± 6.6		71.1 ± 11.0	
Under-represented in Medicine								
URM	9 [4 – 14.5]	0.236	69.1 ± 7.3	0.313	67.7 ± 11.1	0.917	72.5 ± 9.0	0.102
Not URM	11 [8–14]		71.0 ± 8.3		68.3 ± 9.8		75.6 ± 6.7	
Indeterminate	10 [4.75 – 12.5]		68.6 ± 8.8		69.0 ± 8.8		72.3 ± 7.7	
Surgery before IM	11 [7 – 14.75]	0.992	67.9 ± 8.5	**< 0.001**	69.3 ± 9.5	0.139	74.8 ± 8.1	0.993
Surgery After IM	11 [7–14]		72.9 ± 7.1		67.1 ± 10.3		74.8 ± 6.5	
ACS before Subspecialty	11 [7 – 14.25]	0.895	71.1 ± 8.5	0.230	67.8 ± 10.0	0.504	75.6 ± 6.3	0.213
ACS after Subspecialty	11 [7–14]		69.7 ± 7.8		68.8 ± 9.9		74.0 ± 8.3	
Surgical subspecialty								
Elective general	10 [3.25 – 14]	**0.013**	68.9 ± 8.1	0.265	70.4 ± 8.7	0.092	74.3 ± 7.6	0.365
Surgical oncology	12 [10–15]		71.6 ± 8.1		68.3 ± 11.8		76.5 ± 7.4	
Pediatric general	9 [4–12]		71.9 ± 7.7		69.3 ± 9.6		75.4 ± 6.9	
Congenital heart	10 [9–15]		71.6 ± 9.4		63.8 ± 9.6		72.1 ± 6.3	
Vascular	13 [10–17]		67.9 ± 7.0		67.7 ± 7.2		73.9 ± 7.9	
Other	11 [8–11]		69.1 ± 8.9		62.4 ± 7.5		–	
Term								
1	12 [7–14]	0.098	66.0 ± 8.7	**< 0.001**	68.0 ± 10.3	**0.010**	75.8 ± 7.7	0.848
2	10 [7–14]		67.3 ± 7.6		74.1 ± 9.6		75.1 ± 8.3	
3	10 [7–14]		71.3 ± 8.5		67.2 ± 11.4		75.8 ± 7.4	
4	13 [9.5 – 17]		72.2 ± 6.3		66.9 ± 9.4		73.6 ± 7.3	
5	10 [7–13]		72.7 ± 7.7		66.2 ± 7.6		74.1 ± 6.9	
6	10 [4–11]		73.4 ± 7.5		65.9 ± 8.0		73.9 ± 6.0	
Academic year								
2017	8 [3–12]	**< 0.001**	71.1 ± 8.3	0.664	73.0 ± 9.0	**< 0.001**	74.8 ± 7.9	0.462
2018	13 [7–17]		71.2 ± 8.3		68.6 ± 9.4		75.7 ± 6.7	
2019	11 [9–16]		69.6 ± 8.1		68.6 ± 8.5		73.6 ± 7.3	
2020	11.5 [8–13]		69.8 ± 8.1		62.5 ± 9.7		–	
MCAT tertiles								
1 (low)	11 [6–14]	0.357	68.9 ± 8.5	0.070	67.9 ± 9.9	0.374	73.2 ± 7.3	**0.046**
2 (middle)	12 [9 – 14.5]		71.8 ± 8.3		70.0 ± 10.7		74.7 ± 6.8	
3 (high)	10 [8–13]		71.8 ± 7.3		67.2 ± 9.2		77.2 ± 7.2	
Final grade: honors	13 [11–17]	**< 0.001**	74.4 ± 6.1	**< 0.001**	70.1 ± 9.5	**0.018**	76.5 ± 6.7	**0.005**
Final grade: pass/fail	8 [3–10]		66.5 ± 8.1		66.6 ± 10.1		72.8 ± 7.6	

All bolded items correspond to statistically significant values

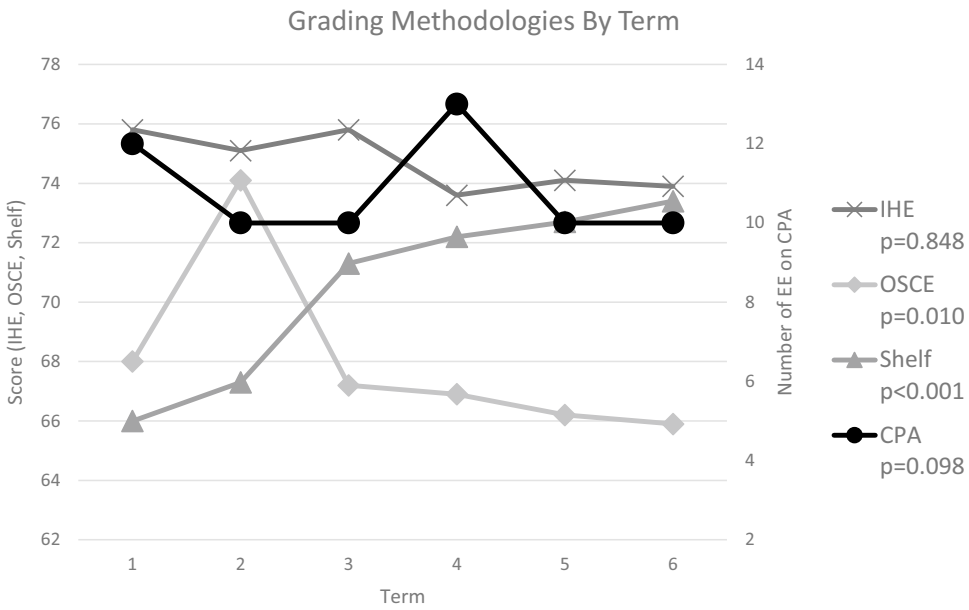

Fig. 2 Assessment methodologies by term

achieved a final grade of Honors on average had a higher OSCE than those who achieved Pass/Fail (70.1 ± 9.5 vs 66.6 ± 10.0, $p = 0.018$). After adjustment (Table 4), associations with Term persisted, with those in Term 2 scoring 7.01 points higher (95% CI 2.22–11.80). Additional associations included an increase by 0.62 points (95% CI 0.20–1.05) for each year older a student was, a decrease in score by − 8.37 (95% CI − 11.97 to − 4.77) in Era 2, and a decrease in score by − 6.01 points (95% CI − 11.28 to − 0.75) if the student's subspecialty was Congenital Heart Surgery.

In-house examination

Pre-IHE and Post-IHE, along with calculated difference between Pre-IHE and Post-IHE per student, are depicted in Table 3. The Pre-IHE showed that students who had IM before CSC achieved a higher initial score (44.6 ± 9.4 vs 38.3 ± 8.1, $p < 0.001$), but the association was no longer present after the Post-IHE (74.8 ± 6.5 vs 74.8 ± 8.1, $p = 0.993$). There was additional variation by Term among the Pre-IHE ($p = 0.005$), with the lowest being among Term 3 (37.2 ± 10.6) and the highest among Term 5 (47.1 ± 8.2). This association was no longer present after the Post-IHE ($p = 0.848$). A student's MCAT tertile had no association on the Pre-IHE, but there was an increase in Post-IHE score as MCAT tertiles increased (73.2 ± 7.3 vs 74.7 ± 6.8 vs 77.2 ± 7.2, $p = 0.046$). On both the Pre-IHE and Post-IHE, students who achieved a final grade of Honors scored higher than those who achieved Pass/Fail (44.4 ± 8.4 vs 37.8 ± 9.2, $p < 0.001$ and 76.5 ± 6.7 vs 72.8 ± 7.6, $p = 0.005$, respectively).

After adjustment (Table 4), there were no associations seen on the Post-IHE, except a 0.11 (95% CI 0.01–0.21) point higher score on the Post-IHE for every point increase on the Step 1 examination.

Final grade

There were 95 (49.5%) students who achieved a final grade of Honors. There was only 1 (0.5%) student who Failed the CSC. Throughout each of the assessment methods, students who achieved Honors had higher scores than those who achieved Pass/Fail (Table 2). After adjustment (Table 4), there were multiple associations with achieving a final grade of Honors identified including: 1.14 odds (95% CI 1.01–1.30) for each year older a student was, 0.17 odds (95% CI 0.06–0.50) in Era 2, 7.06 odds (95% CI 1.21–41.26) if taking Vascular Surgery as a Subspecialty, and 1.07 odds (95% CI 1.03–1.10) for each point increase in Step 1 score.

Discussion

In this exploration of factors associated with different assessment methods used in a surgical clerkship, we found that students' performance on the NBME Shelf exam was associated with race/ethnicity and the term on which the student took the CSC. Performance on the IHE did not have any associations with race/ethnicity or term, suggesting that it is assessing learning that occurs during the CSC itself. When used as part of honors criteria, a Shelf cut-off resulted in fewer students receiving honors. While these associations have not manifested an overall disparity in final grades received in the CSC for URM students, our numbers are small and we cannot say for certain that this association would not eventually emerge.

Table 3 In-house examination scores: beginning and end of rotation

	Pre-IHE (Mean ± SD)	*p* value	IHE (Mean ± SD)	*p* value
All	41.4 ± 9.3	–	74.8 ± 7.3	–
Female	42.1 ± 9.4	0.412	75.8 ± 7.7	0.081
Male	40.7 ± 9.3		73.6 ± 6.7	
Race				
White/Caucasian	42.1 ± 9.0	0.474	75.9 ± 6.7	0.219
Hispanic	43.2 ± 9.3		74.6 ± 7.8	
Asian	39.1 ± 9.7		74.9 ± 6.9	
African American	41.9 ± 10.2		70.4 ± 8.1	
Multiracial	34.5 ± 8.7		72.0 ± 7.8	
Unreported/other	42.5 ± 10.7		71.1 ± 11.0	
Under-represented in medicine				
URM	42.0 ± 9.7	0.929	72.5 ± 9.0	0.102
Not URM	41.3 ± 9.2		75.6 ± 6.7	
Indeterminate	40.9 ± 10.3		72.3 ± 7.7	
Surgery before IM	38.3 ± 8.1	**< 0.001**	74.8 ± 8.1	0.993
Surgery after IM	44.6 ± 9.4		74.8 ± 6.5	
ACS before subspecialty	41.8 ± 9.7	0.644	75.6 ± 6.3	0.213
ACS after subspecialty	41.0 ± 8.9		74.0 ± 8.3	
Surgical subspecialty				
Elective general	40.4 ± 8.3	0.485	74.3 ± 7.6	0.365
Surgical oncology	43.5 ± 9.6		76.5 ± 7.4	
Pediatric general	41.7 ± 8.7		75.4 ± 6.9	
Congenital heart	38.6 ± 11.3		72.1 ± 6.3	
Vascular	42.1 ± 10.3		73.9 ± 7.9	
Other	–		–	
Term				
1	39.1 ± 6.2	**0.005**	75.8 ± 7.7	0.848
2	40.0 ± 9.5		75.1 ± 8.3	
3	37.2 ± 10.6		75.8 ± 7.4	
4	44.8 ± 8.4		73.6 ± 7.3	
5	47.1 ± 8.2		74.1 ± 6.9	
6	41.7 ± 8.5		73.9 ± 6.0	
Academic year				
2017	41.4 ± 9.7	**< 0.001**	74.8 ± 7.9	0.462
2018	44.5 ± 8.9		75.7 ± 6.7	
2019	37.2 ± 7.8		73.6 ± 7.3	
2020	–		–	
MCAT tertiles				
1 (low)	41.7 ± 8.9	0.594	73.2 ± 7.3	**0.046**
2 (middle)	41.7 ± 7.9		74.7 ± 6.8	
3 (high)	40.0 ± 9.4		77.2 ± 7.2	
Final Grade: Honors	44.4 ± 8.4	**< 0.001**	76.5 ± 6.7	**0.005**
Final Grade: Pass/Fail	37.8 ± 9.2		72.8 ± 7.6	

All bolded items correspond to statistically significant values

Subjective assessment: clinical performance assessment

This assessment is a means of quantifying the subjective evaluation of students from faculty and residents across all clerkships. On the CSC, we saw variation in the number of EE achieved across Academic Years, especially a large change from a median of 8 [IQR: 3–12] in 2017 to 13 [IQR: 7–17] in 2018, with leveling off to 11–11.5 in 2019–2020. This could reflect the growing pains and faculty adjusting to the grading system, but this would need validation over additional years.

As with many subjective evaluations, there are individuals who grade more moderately than others. This may be reflected in variation in the number of EE achieved across Subspecialties on univariate analysis. There was a wide variation, from a median of 9 to 13, perhaps reflecting more lenient grading by the Vascular surgeons and more stringent evaluation on the Pediatric General Surgery service. However, this association did not persist in multivariable analysis adjusting for other potential confounding factors and none of the Subspecialty services were associated with increased odds of obtaining > 10 EE necessary to achieve Honors. Importantly, the CSC allowed students to document examples of their clinical engagement through reflective writing to be considered evidence of demonstrating the anchoring text for "exceeds expectations" in several domains assess on the CPA. This was found to be particularly helpful in assessing competencies that are often not directly observed by faculty and resident assessors, such as "Inter-professional Collaboration" or "Health Systems Context", and was considered a counter-measure to the differences in subjective assessments across attendings and services. While CPA rater training included explanations of the rating tool and discussion of student actions/behaviors that would align with these anchoring texts for each of the assessment categories, there is always potential for variability in responses. With this knowledge, re-calibrating rater training across all services, including Vascular and Pediatrics could be undertaken to reduce variance in student assessments coming from rater use of the assessment tool.

With no factors conferring an increased odds of obtaining > 10 EE besides Step 1 scores, the CPA appeared to be a fairly unbiased and reliable assessment method. Quantifying the subjective nature of clinical evaluations may be worthwhile to implement at institutions with strictly subjective evaluations. Prior research has indicated that the Medical School Performance Evaluation, which utilizes summary words from subjective faculty assessments, may have racial biases [24], so having both a quantitative and subjective assessment of students from faculty may be worthwhile.

Table 4 Multivariable linear and logistic regressions of assessment methods

	OR of CPA > 10	*p* value	Point difference in Shelf Score	*p* value	Point difference in OSCE	*p* value	Point difference in IHE	*p* value	OR FG Honors	*p* value
Age (years)	1.07 (0.96 to 1.20)	0.228	0.11 (− 0.2 to 0.42)	0.501	0.62 (0.20 to 1.05)	**0.005**	0.14 (− 0.26 to 0.53)	0.503	1.14 (1.01 to 1.30)	**0.040**
Female	1.00 (0.49 to 2.05)	0.994	0.51 (− 1.64 to 2.65)	0.645	0.10 (− 2.90 to 3.10)	0.947	2.12 (− 0.54 to 4.78)	0.121	0.81 (0.37 to 1.78)	0.608
URM										
Not URM	Ref.	–	Ref.	–	Ref.	–	Ref.	–	Ref.	–
URM	0.86 (0.35 to 2.14)	0.748	0.42 (− 2.41 to 3.25)	0.771	− 1.38 (− 5.28 to 2.53)	0.491	− 2.07 (− 5.62 to 1.49)	0.257	1.04 (0.37 to 2.90)	0.941
Indeterminate	1.19 (0.31 to 4.64)	0.799	− 1.92 (− 5.87 to 2.03)	0.343	3.51 (− 2.54 to 9.57)	0.257	− 5.1 (− 10.94 to 0.73)	0.090	0.76 (0.18 to 3.19)	0.702
Surgery before IM	1.84 (0.66 to 5.13)	0.242	− 2.59 (− 5.53 to 0.35)	0.086	0.65 (− 3.30 to 4.61)	0.747	− 1.12 (− 4.81 to 2.57)	0.553	0.51 (0.17 to 1.48)	0.213
ACS before subspecialty	0.73 (0.37 to 1.43)	0.356	1.74 (− 0.31 to 3.78)	0.098	− 1.33 (− 4.19 to 1.53)	0.364	1.54 (− 1.14 to 4.23)	0.263	0.81 (0.38 to 1.73)	0.584
After AY 2019	1.52 (0.63 to 3.70)	0.353	0.51 (− 2.08 to 3.11)	0.699	− 8.37 (− 11.97 to − 4.77)	**< 0.001**	–	–	0.17 (0.06 to 0.50)	**0.001**
Surgical subspecialty										
Elective general	Ref.	–	Ref.	–	Ref.	–	Ref.	–	Ref.	–
Surgical oncology	2.17 (0.82 to 5.77)	0.120	0.81 (− 2.10 to 3.73)	0.585	− 1.86 (− 5.84 to 2.12)	0.361	0.23 (− 3.31 to 3.78)	0.897	1.65 (0.55 to 4.99)	0.372
Pediatric general	0.46 (0.17 to 1.21)	0.115	0.77 (− 2.20 to 3.73)	0.614	1.30 (− 2.87 to 5.47)	0.543	− 0.78 (− 4.51 to 2.94)	0.681	0.94 (0.32 to 2.71)	0.902
Congenital heart	0.77 (0.24 to 2.46)	0.656	0.75 (− 2.91 to 4.40)	0.690	− 6.01 (− 11.28 to − 0.75)	**0.027**	− 4.66 (− 9.44 to 0.12)	0.059	0.40 (0.10 to 1.57)	0.191
Vascular	3.05 (0.72 to 12.98)	0.132	0.01 (− 4.22 to 4.24)	0.996	− 5.06 (− 10.92 to 0.8)	0.093	1.38 (− 3.65 to 6.41)	0.592	7.06 (1.21 to 41.26)	**0.030**
Other	0.52 (0.09 to 2.96)	0.464	− 1.16 (− 6.56 to 4.23)	0.673	0.84 (− 7.33 to 9.01)	0.841	–	–	1.14 (0.15 to 8.78)	0.903
Term										
1	Ref.	–	Ref.	–	Ref.	–	Ref.	–	Ref.	–
2	0.56 (0.17 to 1.80)	0.331	0.15 (− 3.43 to 3.74)	0.933	7.01 (2.22 to 11.8)	**0.005**	− 1.38 (− 5.68 to 2.92)	0.531	0.63 (0.17 to 2.36)	0.488
3	0.75 (0.22 to 2.61)	0.653	4.62 (0.86 to 8.37)	**0.017**	− 1.34 (− 6.36 to 3.69)	0.603	− 0.59 (− 5.13 to 3.96)	0.800	0.70 (0.17 to 2.94)	0.631
4	2.09 (0.48 to 9.16)	0.326	3.99 (− 0.12 to 8.10)	0.059	− 0.74 (− 6.25 to 4.76)	0.792	− 2.42 (− 7.44 to 2.60)	0.347	2.20 (0.49 to 9.88)	0.304
5	0.82 (0.19 to 3.53)	0.788	4.92 (0.53 to 9.31)	**0.030**	− 0.99 (− 7.19 to 5.22)	0.755	− 2.14 (− 7.87 to 3.59)	0.466	1.16 (0.23 to 5.88)	0.856
6	2.15 (0.42 to 10.97)	0.359	7.56 (2.81 to 12.31)	**0.002**	2.04 (− 4.79 to 8.86)	0.560	− 0.81 (− 7.33 to 5.71)	0.808	1.11 (0.19 to 6.28)	0.910
MCAT tertiles										
1	Ref.	–	Ref.	–	Ref.	–	Ref.	–	Ref.	–
2	2.15 (0.85 to 5.40)	0.104	1.05 (− 1.63 to 3.72)	0.445	1.22 (− 2.51 to 4.96)	0.522	1.88 (− 1.62 to 5.38)	0.296	2.45 (0.90 to 6.70)	0.080
3	0.65 (0.27 to 1.58)	0.342	0.86 (− 1.83 to 3.55)	0.531	− 3.26 (− 7.01 to 0.49)	0.090	3.23 (− 0.18 to 6.64)	0.067	1.21 (0.46 to 3.23)	0.697
Step 1 score	1.03 (1.00 to 1.06)	**0.042**	0.25 (0.17 to 0.33)	**< 0.001**	0.09 (− 0.02 to 0.20)	0.124	0.11 (0.01 to 0.21)	**0.039**	1.07 (1.03 to 1.10)	**< 0.001**

All bolded items correspond to statistically significant values

Temporal associations

Due to the number of patients, students, and clinical sites, there will always be differences in the order in which students take clerkships. Some will take IM before CSC, on the CSC itself, some will take different sub-specialties in different orders, with diverse demands and different overlap with curricular learning objectives. However, problems arise when these factors affect students' grades, which ultimately affects students' access to competitive residencies. For example, students taking the CSC *after* IM achieved 5 points higher than their counterparts on the Shelf (72.9 ± 7.1 vs 67.9 ± 8.5, $p < 0.001$), which has been previously reported [21, 22]. In addition, the steady increase in Shelf score across Terms by an average of 7.4 points overall is further evidence that these exams are at least in part measuring progress toward competency, as is appropriate for a licensing exam which is ultimately pass/fail. However, when there are Shelf score thresholds to achieve honors, these could have substantial implications on a student's immediate grade and eventual career. This was further supported on multivariable analysis. While the association between CSC before or after IM and Shelf score itself no longer became significant, students continued to obtain higher Shelf scores during later Terms. Furthermore, there was a negative impact on the odds of achieving a final grade of Honors in Era 2 [OR 0.17 (95% CI 0.06–0.50], when the Shelf cutoff was implemented. The Shelf exam, therefore, significantly limited Honors in Era 2 after the cutoff was implemented. Students are well aware of this pattern and use it as part of their decision-making around selecting clerkship order. This ends up having a negative impact on students doing surgery early who then decide to pursue surgery as a specialty because of their experience on the CSC.

By contrast, an IHE reflective of a program's curriculum may be more fair summative assessment than the Shelf. This was supported by the fact that the Pre-IHE *did* reflect the temporal differences expected by a student's experience level in that the mean Pre-IHE score was higher in the final three Terms than any of the first three Terms ($p = 0.005$) and that students who had the CSC after IM were 6.3 points higher on the Pre-IHE, which was not dissimilar to the Shelf exam. However, by the end of the rotation and implementation of a dedicated curriculum, the Post-IHE showed no differences in score across Terms ($p = 0.848$) and those who had CSC before IM scored the exact same as those who had CSC after IM (74.8 ± 8.1 vs 74.8 ± 6.3, $p = 0.993$). Any temporal differences were, therefore, evened out. Given that the Post-IHE still followed a normal distribution, it could very easily be used to identify high-performing students deserving of Honors after demonstrating learning that the CSC is supposed to foster through its unique curriculum, not that gleaned from other clerkships throughout the year.

Equity

There is evidence to suggest that there are racial groups and individuals who systemically perform better on standardized multiple-choice examinations due to multiple societal disparities [25]. This was initially seen with Racial differences on the Shelf exam, with Whites having the highest scores and African Americans as well as Unreported/Other races scoring the lowest. We, therefore, attempted to adjust for racial and demographic confounding factors in our multivariable models by incorporating age, sex, and URM status. We found that sex and URM status did not display any impact on assessment methods, though student age was associated with improved OSCE scores and higher odds of a final grade of Honors. As the OSCE is administered as a formative assessment and did not have an impact on final grades, we hypothesize that the variation in OSCE scores may come from students not preparing for it as they would a summative assessment. Increased age may reflect greater emotional maturity and personal experience resulting in higher OSCE scores. When grouped according to URM status, no associations were found on univariate or multivariable analysis. At a relatively small state school in a state whose racial distributions are different than that of the national population, these finding require validation on a larger scale.

"Good test takers"

Even without racial disparities, some students may be good "test takers" whose scores do not correlate with their real-life clinical acumen. This is an important consideration, because performance on multiple-choice standardized tests may be over-represented in assessments that end up as part of a residency application, and few correlations exist between this dimension of performance and the other critical competency domains in GME [9, 12, 13, 25]. One area in which this has been seen is the disconnect between Shelf scores and the subjective evaluation of students by faculty on the CSC [18, 19]. We, therefore, created our multivariable models to incorporate standardized tests performed before and after the CSC with MCAT tertile and Step 1 grade, respectively. A student's MCAT tertile had no associations with any assessment method in the model, but a student's Step 1 score had multiple associations including higher CPA, Shelf score, IHE score, and increased odds of Honors on the CSC.

The lack of associations with MCAT tertile may reflect a level playing field: all of the students were likely deserving of their admission to medical school. Since admission, they all undergo the same curriculum and teaching; therefore, the differences in undergraduate teaching are potentially eliminated with the unified curriculum during medical school. This unified curriculum is potentially reflected by seeing no

differences in their various performance assessments on the CSC when stratified by MCAT score.

The consistent associations on the assessment methods with Step 1, however, may reflect a different paradigm. It may show that the USMLE Step 1 accurately captured those students who performed better on a multitude of assessment metrics. This agrees with previous studies [1, 26]. However, when looking at the *degree* of impact, we see that there was a much higher increase in Shelf score (0.25 points per increase in Step 1 score) compared to the IHE (0.11 points per increase in Step 1 score). This again reflects that an IHE may be a more fair summative assessment in that good test takers had less of an impact on scores than that seen on the Shelf.

While the IHE was abandoned at our institution prior to this study due to the demands for psychometric analysis that exceeded the resources available, it highlights the need for assessments that reflect curriculum objectives. One effort toward that goal is the collaboration between the NBME and the Association for Surgical Education (ASE) and American College of Surgeons (ACS) Medical Student Core Curriculum to align the Shelf exam with a curriculum developed by surgical educators. The authors look forward with much anticipation to the findings and analysis of that partnership. Since the Shelf exam, despite its current shortcomings highlighted above, is a useful checkpoint against the content covered by a clerkship that will ultimately tested on USMLE Step 2 examinations, it has been continued at our institution. As efforts to align the Shelf with the ASE/ACS curricula mature, there may be a time when one can say that the Shelf exam reflects the curriculum and, therefore, is an appropriate and equitable assessment tool for grading.

Limitations

The limitations of this study stem from its retrospective, single-institution nature at a relatively small medical school. In addition, due to the adjustments necessary in the creation of a new medical school and through the disruptions caused by the COVID-19 pandemic, the learning experiences of each student may not be equivalent, and therefore, the assessment methods not generalizable across the students examined. Generalization of our results to other larger schools with more clinical sites will need to be measured. Further research would be required to make more far-reaching conclusions.

Conclusions

There are various assessment methods available to grade medical students on the CSC, each of which has its distinct benefits and drawbacks. No one assessment should be used as a summative assessment in isolation and each program must be cognizant of what assessments best meet its curricular objectives. An equity lens in performance assessments should continue to be emphasized medical education. The Shelf, a near ubiquitous examination across medical schools, is substantially influenced by temporal associations across Terms and timing in relation to other clerkships, such as IM, and by "good test takers". An IHE, or other assessment that specifically reflects a clerkship's specified curriculum, may be a more equitable summative assessment of the learning that occurs from the CSC curriculum, with fewer biases or influences external to the CSC.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44186-022-00047-8.

Author contributions All listed authors are qualified for authorship and all who are qualified to be authors are listed as authors on the byline.

Funding The authors did not receive support from any organization for the submitted work.

Data availability The data sets generated during and/or analyzed during the current study are not publicly available due to them containing private student information, but are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical approval This study was approved by the Institutional Review Board of the Dell Medical School at The University of Texas at Austin, STUDY00002003, approved November 2, 2021.

References

1. Barnum TJ, Halverson AL, Helenowski I, Odell DD. All work and no play: addressing medical students' concerns about duty hours on the surgical clerkship. Am J Surg. 2019;218:419–23. https://doi.org/10.1016/J.AMJSURG.2018.12.012.
2. Hinojosa-Gonzalez DE, Farias JS, Tellez-Giron VC, Aguirre-Villarreal D, Brenes-Castro D, Flores-Villalba E. Lower frequency of call shifts leads to higher attendance, higher academic performance, and less burnout syndrome in surgical clerkships. J Surg Educ. 2021;78:485–91. https://doi.org/10.1016/J.JSURG.2020.07.043.
3. Ziegler T, Adibfar A, Abbasian A, Jiang SX, Rutka JT, Gawad N. Propagating the "SEAD": exploring the value of an overnight call shift in the surgical exploration and discovery (SEAD) program. J Surg Educ. 2020;77:104–14. https://doi.org/10.1016/J.JSURG.2019.08.011.
4. Skube SJ, Ramaswamy A, Chipman JG, Acton RD. Medical student perceptions of 24-hour call. J Surg Educ. 2019;76:387–92. https://doi.org/10.1016/J.JSURG.2018.09.002.
5. Scott O, Novak C, Forbes K. Medical student perceptions of on-call modalities: a focus group study. Teach Learn Med. 2018;31:34–43. https://doi.org/10.1080/10401334.2018.1480957.

6. McLean SF, Horn K, Tyroch AH. Case based review questions, review sessions, and call schedule type enhance knowledge gains in a surgical clerkship. J Surg Educ. 2013;70:68–75. https://doi.org/10.1016/J.JSURG.2012.07.005.
7. Chai AL, Matsushima K, Sullivan ME, Inaba K, Demetriades D. Acute care surgery education in US medical schools: a systematic review of the current literature and report of a medical student experience. J Surg Educ. 2020;77:316–22. https://doi.org/10.1016/J.JSURG.2019.09.008.
8. Onumah CM, Lai CJ, Levine D, Ismail N, Pincavage AT, Osman NY. Aiming for equity in clerkship grading: recommendations for reducing the effects of structural and individual bias. Am J Med. 2021;134:1175-1183.e4. https://doi.org/10.1016/J.AMJMED.2021.06.001.
9. Caldwell KE, Zarate Rodriguez JG, Hess A, Han BJ, Awad MM, Sacks BC. Standardized oral examinations allow for assessment of medical student clinical knowledge and decrease racial grading differences in a surgery clerkship. Surgery. 2022;171:590–7. https://doi.org/10.1016/J.SURG.2021.11.005.
10. Ross PT, Lypson ML, Byington CL, Sánchez JP, Wong BM, Kumagai AK. Learning from the past and working in the present to create an antiracist future for academic medicine. Acad Med. 2020. https://doi.org/10.1097/ACM.0000000000003756.
11. Ufomata E, Merriam S, Puri A, Lupton K, LeFrancois D, Jones D, Nemeth A, Snydman LK, Stark R, Spagnoletti C. A policy statement of the society of general internal medicine on tackling racism in medical education: reflections on the past and a call to action for the future. J Gen Intern Med. 2021;36:1077–81. https://doi.org/10.1007/S11606-020-06445-2/FIGURES/1.
12. Teherani A, Hauer KE, Fernandez A, King TE, Lucey C. How small differences in assessed clinical performance amplify to large differences in grades and awards: a cascade with serious consequences for students underrepresented in medicine. Acad Med. 2018;93:1286–92. https://doi.org/10.1097/ACM.0000000000002323.
13. Merrick HW, Nowacek G, Boyer J, Robertson J. Comparison of the objective structured clinical examination with the performance of third-year medical students in surgery. Am J Surg. 2000;179:286–8. https://doi.org/10.1016/S0002-9610(00)00340-8.
14. Kelly WF, Papp KK, Torre D, Hemmer PA. How and why internal medicine clerkship directors use locally developed, faculty-written examinations: results of a national survey. Acad Med. 2012;87:924–30. https://doi.org/10.1097/ACM.0B013E318258351B.
15. Veale P, Woloschuk W, Coderre S, McLaughlin K, Wright B. Comparison of student performance on internally prepared clerkship examinations and NBME subject examinations. Can Med Educ J. 2011;2:e81–5. https://doi.org/10.36834/CMEJ.36560.
16. National Board Of Medical Examiners® Subject examination program comprehensive basic science examination score interpretation guide. 2020. https://www.nbme.org/sites/default/files/2020-01/Basic%20Score%20Interpretation%20Guide%20and%20Norms.pdf. Accessed 22 Feb 2022.
17. Myers JA, Vigneswaran Y, Gabryszak B, Fogg LF, Francescatti AB, Golner C, Bines SD. NBME subject examination in surgery scores correlate with surgery clerkship clinical experience. J Surg Educ. 2014;71:205–10. https://doi.org/10.1016/J.JSURG.2013.07.003.
18. Farrell TM, Kohn GP, Owen SM, Meyers MO, Stewart RA, Meyer AA. Low correlation between subjective and objective measures of knowledge on surgery clerkships. J Am Coll Surg. 2010;210:680–3. https://doi.org/10.1016/J.JAMCOLLSURG.2009.12.020.
19. Hermanson B, Firpo M, Cochran A, Neumayer L. Does the National Board of Medical Examiners' surgery subtest level the playing field? Am J Surg. 2004;188:520–1. https://doi.org/10.1016/J.AMJSURG.2004.07.036.
20. Reid CM, Kim DY, Mandel J, Smith A, Bansal V. Correlating surgical clerkship evaluations with performance on the National Board of Medical Examiners examination. J Surg Res. 2014;190:29–35. https://doi.org/10.1016/J.JSS.2014.02.031.
21. Dong T, Copeland A, Gangidine M, Schreiber-Gregory D, Ritter EM, Durning SJ. Factors associated with surgery clerkship performance and subsequent USMLE step scores. J Surg Educ. 2018;75:1200–5. https://doi.org/10.1016/J.JSURG.2018.02.017.
22. Ouyang W, Cuddy MM, Swanson DB. US medical student performance on the NBME subject examination in internal medicine: do clerkship sequence and clerkship length matter? J Gen Intern Med. 2015;30:1307–12. https://doi.org/10.1007/S11606-015-3337-Z/TABLES/5.
23. R Core Team. R: a language and environment for statistical computing. 2019.
24. Low D, Pollack SW, Liao ZC, Maestas R, Kirven LE, Eacker AM, Morales LS. Racial/ethnic disparities in clinical grading in medical school. Teach Learn Med. 2019;31:487–96. https://doi.org/10.1080/10401334.2019.1597724.
25. Jencks C, Phillips M. The black-white test score gap. Washington, DC: Brookings Institution Press; 2011.
26. Zahn CM, Saguil A, Artino AR, Dong T, Ming G, Servey JT, Balog E, Goldenberg M, Durning SJ. Correlation of National Board of Medical Examiners scores with united states medical licensing examination step 1 and step 2 scores. Acad Med. 2012;87:1348–54. https://doi.org/10.1097/ACM.0B013E318266A13BD.