The American Journal of Surgery 220 (2020) 328–333

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

Oral examinations in undergraduate medical education – What is the ‘value added’ to evaluation?

Luise I.M. Pernar [a, *], Reza Askari [b], Elizabeth M. Breen [c]

[a] Boston Medical Center, Boston, MA, USA
[b] Brigham and Women’s Hospital, Boston, MA, USA
[c] Lahey Clinic, Burlington, MA, USA

ARTICLE INFO

Article history:
Received 24 October 2019
Received in revised form
19 December 2019
Accepted 30 December 2019

Keywords:
Surgery
Clerkship
Oral examination
Evaluation
Assessment

ABSTRACT

Background: Given the long tradition of oral examinations in surgical training, surgical clerkships continue to use oral examinations to evaluate medical students even though the value of oral examination at the post-graduate level has been questioned. The key issue in the context of undergraduate surgical training then is to understand value of the oral examination in assessment. The goal of this study is to clarify what oral examinations do, or appear to, test and how this complements other methods of assessment.

Methods: The study is a retrospective, qualitative study of comments provided by examiners on the oral examination score sheets evaluating performance of students completing their core surgery clerkship at an academic medical center. Through immersion in and initial familiarization with the data we develop a scheme of codes for labeling the data for subsequent synthesis. Using these inductive codes, all comments were reviewed and analyzed to determine what qualities examiners detect, or naturally comment on, when administering and scoring the oral examinations.

Results: Thirteen substantive codes (Communication, Critical Thinking, Decisiveness, Demeanor, Differential Diagnosis, Focus, Knowledge, Management, Organization, Pace, Prompting, Thoroughness, and Work Up) and three valence codes (Negative, Neutral, and Positive) were developed and used to code the data. The most universal code was ‘Knowledge’, used by 43 (100%) of examiners; the most frequently used code was ‘Work Up’, applied to the comments 437 (21.1%) times. Overall, positive valence was attached to 1146 (55.2%) of codes and negative valence to 879 (42.3%) codes. The most discriminating codes in grading were ‘Demeanor’, ‘Focus’, and ‘Organization’.

Conclusions: Oral examinations provide rich opportunity for testing qualities readily tested on other examinations but also many intangible qualities that are otherwise less well or not well tested. As such, the ‘value-added’ by oral examinations likely justifies their continued use in the evaluation of surgical trainees. The identification of testable qualities should aid in the development of a standardized scoring rubric, the use of which may aid in minimizing subjectivity and bias in what otherwise is a rich assessment tool.

Introduction

The American Board of Surgery (ABS) is tasked with the certification of surgeons. To this end the Board initially proposed, “that the examinations will be divided into two parts of which one will be written and the other will be practical, covering bedside work, clinical, and operative features.“[1] In its original iteration the practical examination consisted of observation of the candidate surgeon in the clinic as well as in the operating room.[2] This method was abandoned in 1952 as too many candidates required observation to make the process feasible.[2] Ultimately, the Board shifted technical skill assessment to training institutions and the second portion of the practical examination has since become strictly an oral examination, the Certifying Examination (CE). According to the ABS, the “CE is designed to assess a candidate’s surgical judgement, clinical reasoning skills and problem-solving ability”.[3] However, the

[*] Corresponding author. Department of Surgery, 1 Boston Medical Center Place, Boston Medical Center, Collamore D-501, Boston, MA, 02118, USA.
E-mail address: Luise.Pernar@bmc.org (L.I.M. Pernar).

https://doi.org/10.1016/j.amjsurg.2019.12.031
0002-9610/© 2020 Elsevier Inc. All rights reserved.

contents of the CE may not be disclosed and we are not aware of publically available evaluation criteria.

Despite the long tradition of use of oral examinations in surgical training there is a relative paucity of literature regarding the oral examination, particularly when it comes to evaluating what the examination measures. Generally speaking it is thought that the oral examination is useful to determine an examinee's ability to apply clinical knowledge[4–6] and to test how facile the examinee is in tackling clinical problems.[7]However, the oral examination's true usefulness in this regard has come under scrutiny. Concerns about subjectivity and bias have been raised [5,6,8] and include that an examinee's manner of dress and speech can significantly alter the examination's outcome. Also, participation in 48-h oral board review courses, thought to be too brief to confer actual knowledge, improves pass rates for second time examinees beyond pass rates for first time examinees.[9]

The issues facing the oral board at the graduate level also affects undergraduate medical education. Given the long tradition of oral examinations in surgical training, some surgical clerkships continue to use oral examinations to evaluate medical students. For medical students, as for surgeons taking the CE, the question what the oral examination tests also remains unresolved. The key issues in this context are to describe what the purported value of the oral examination is and to clarify if oral examinations do, or should, test knowledge or more intangible qualities.

Given the lack of clarity, we elected to review and qualitatively analyze available oral examination score sheets from our Department where oral examinations have long been conducted for medical students. The goal was to determine what qualities examiners look for when administering the oral examinations and what they include in considering grades. By inference, these likely represent the qualities oral examinations lend themselves to evaluate and the analysis therefore should aid in defining the value of the oral examination when used in the context of evaluation.

Methods

Oral examinations

All students completing their third year core surgery clerkship at Brigham and Women's Hospital (BWH) are required to participate in an oral examination exercise at the end of the clerkship. Two examiners, recruited for this exercise from the general and vascular surgery faculty staff cadre, examine each student independently. For the purposes of the examination examiners are given two broad topic areas, specifically, gastroenterology/oncology (GI-Onc) and trauma/fluids/electrolytes (TFL). Examiners are also provided case stems with suggested questions and lines of inquiry; however, they are empowered to tailor the examination and adjust questions based on how the examination unfolds. Each examination is intended to last approximately 20 min.

At the completion of the oral examination, the examiner awards a grade on a four-tiered grade scale (High Honors, Honors, Satisfactory, or Unsatisfactory) and provides free-form written comments to describe the examined student's performance. Completed grade sheets are submitted to the clerkship coordinator in the Surgery Education Office (SEO), become part of the students' permanent file, and are factored into the final grade awarded for the clerkship. During the time period for which data was collected for this study, the oral examination counted for 5% of the final grade.

Data collection

All oral examination score sheets returned to the SEO for students completing their clerkships between December 2005 and July 2011 were collected and de-identified for further analysis. For analysis, all grades and free-form comments from the score sheets were transcribed into a text document. The free-form comments were qualitatively analyzed using codes developed through extensive data immersion .[10,11] Specifically, the comments were extensively reviewed and keywords and phrases were extracted to develop codes. Additionally, valence codes were specified to modify the theme codes. Instructions for coding and valence determinations are shown in Table 1.

Using these instructions, comments were coded using qualitative data software (ATLAS.ti, ATLAS.ti Scientific Software Development GmbH, Berlin, Germany). A training set of 20 comments was randomly selected and coded independently by three analysts. In a group consensus conference any discrepancies in coding were discussed until agreement was reached regarding coding. After this training, a single analyst coded the remainder of all comments. In parallel, to ensure consistency in coding, the other two analysts coded a 100-phrase subset of all available comments. These codes were compared to the coding done by the analyst coding the entire set to ensure the codes were applied as agreed in the consensus conference and to calculate inter-rater reliability. For this 100-phrase set, inter-rater reliability was calculated to be 0.83 for all analysts. All data analysis discussed below was performed on the entire coded set.

Data analysis

All counts of comment codes and attached valences were tabulated and tallied. While comments designated as 'Junk', i.e. those comments that were not substantive, were coded and tallied, the analysis was performed only taking into consideration those comments that were substantive and 'Junk' comments were excluded from additional analysis.

Results

Students

Between December 2005 and July 2011, 24 12-week core surgery clerkships were completed at BWH. Because of a curriculum change taking place in June of 2008, several clerkships partially overlapped between March 2008 and October 2008. Taking into account one oral exam missed for illness, 315 students were examined and 629 oral examination score sheets were available for review. The average age of students was 26.5 (SD 2.4; range 23–38) and 50.6% were male. Other demographic data, such as race or ethnicity, was not available for review.

Grades

The most frequently awarded grade for the oral examination was Honors (266, 42.3%), followed by High Honors (203, 32.3%), and Satisfactory (150, 23.8%); an Unsatisfactory grade was rarely given (9, 1.4%). One (0.2%) oral examination was scored without a grade being awarded. Of note, the full grade scale, excluding Unsatisfactory, since it was so infrequently awarded, was used by 26 (60.5%) examiners. Two (4.7%) examiners graded using only one grade; 1 (2.3%) examiner only awarded High Honors and 1 (2.3%) only awarded Satisfactory grades.

Codes

Table 2 shows representative examples of comments that were coded using the identified codes and valences.

Table 1
Theme codes identified through data immersion and corresponding coding instructions for qualitative analysis.

Codes - Qualities	This could should be applied to any
Communication	Instance in which the student's ability to communicate (either as listener or as speaker) is commented upon. Non-verbal communication should also be coded using this code.
Critical Thinking	Reference to a student's demonstration of (or failure to demonstrate) critical thinking. This encompasses any instances in which there is (or the student fails to appropriately show) judgment, appraisal of a situation, ability to integrate information, incorporation of new data etc.
Decisiveness	Mention of the student's ability or readiness (or not) to commit to a course of inquire or action in any activity. NOTE that PROMPTING is a separate code and any mention that can be coded using PROMPTING should be coded thusly and NOT using this code.
Demeanor	Mention regarding the student's individual representation and their interaction with the examiner or how they handled the examination situation.
Differential Diagnosis	Mention of a student's ability (or inability) to formulate an appropriate differential diagnosis for a problem.
Focus	Mention of how concise or goal-directed (or not) etc. a student is in relation to any activity including, but not limited to, the gathering of, interpretation of, or acting on information.
Junk	Situation where there is a general statement made (either positive or negative) that has no pertinent feedback potential or cannot be otherwise classified. Comments such as 'good job', 'outstanding job' fall under this code.
Knowledge	Any instance of commentary on a student's knowledge (or demonstration of lack thereof). NOTE that there are separate codes for WORK UP, DIFFERENTIAL DIAGNOSIS, MANAGEMENT, and CRITICAL THINKING. Any specific mention of or reference to these should be coded thusly and NOT using this code.
Management	Mention of a student's attempt (or failure to attempt) to manage the problem they have been presented with. This includes any intervention ranging from observation over placement of tubes or lines to administration of fluids or blood products and, finally, operative management.
Organization	Mention of the student's organization (or lack thereof) in, but not limited to, approach, differential diagnosis, thought, or management plan.
Pace	Mention of a student's speed, expeditiousness, slowness etc. in relation to any activity including, but not limited to, the gathering of, interpretation of, or acting on information.
Prompting	Mention of the need for prompting, hinting, cajoling (or not) etc. to get the student to make progress in any activity including, but not limited to, the gathering of, interpretation of, or acting on information.
Thoroughness	Mention of how accurate, comprehensive (or not) etc a student is in relation to any activity including, but not limited to, the gathering of, interpretation of, or acting on information.
Work Up	Mention of a student's selection of questions, laboratory investigation, imaging, or any other modality pursued by the student to make a diagnosis. Any mention of trauma algorithms etc. would also be appropriately coded using this code.
Codes – Valences	**This code should be applied if**
Negative	The valance of the statement otherwise coded is negative or entails a suggestion for improvement. This code should also be applied if a course of action is corrected for the initially wrong course of action.
Neutral	A comment is neither positive nor negative but simply is stated or reported on neutrally.
Positive	The valance of the statement otherwise coded is positive. Beware of statements such as 'was better with prompting' - this should be coded using the codes PROMPTING and NEGATIVE because the implication is that prompting was necessary.

Counts

All 43 (100%) examiners made at least one comment that was coded for knowledge. In contrast, only 13 (30.2%) provided a comment that was coded for decisiveness. All use counts are summarized in Table 3.

While the code 'Knowledge' was used by all examiners at least once, this code overall was not the most frequently applied. Of all substantive comments, the code 'Work Up' was applied 437 (21.1%) times, making it the most frequent. 'Knowledge' was applied 253 (12.2%), making it the third most frequent code after 'Management' (352, 17%); 'Decisiveness' was the least frequently applied code (23, 1.1%). The use frequency data are summarized in Table 3.

Valences

Overall, 1146 (55.2%) of applied codes had a positive valence and 879 (42.3%) had negative valence. 'Knowledge' most frequently co-occurred with the positive valence code (186, 73.5%), while 'Prompting' most frequently co-occurred with the negative valence code (156, 87.2%). The code 'Pace' was used in a negative way 36 (65.5%) times; when used negatively comments referred to a student acting too slowly 22 (61.1%) times and too quickly 14 (38.9%) times. All valence counts are shown in Fig. 1 and summarized in Table 3.

Themes

The codes were also organized thematically in to three groups. The codes Work up, Management, Knowledge, and Differential Diagnosis were groups together under the theme Clinical Knowledge. The codes Thoroughness, Prompting, Critical Thinking, Organization, Pace, and Decisiveness were groups together under the theme Cognitive Processes and Behaviors Associated with Surgical Decision Making. Finally, Demeanor and Communication comprised the theme of Interpersonal Skills (Table 3). The themes were encountered in the comments in decreasing frequency with Clinical Knowledge most frequent and Interpersonal Skills least frequent.

Code and valence relationship to grades

Table 4 shows the code counts, by valence, applied to the provided comments broken down by the awarded grades. The codes 'Demeanor', 'Focus', and 'Organization', overall used mostly in a positive way (see Table 3), were most frequently used in a positive way when a High Honors grade was awarded and most frequently used in a negative way when a Satisfactory grade was awarded. 'Communication', also a code predominantly used in a positive way (Table 3) most often accompanied a High Honors grade when used in grade comments; however, negative use was predominantly associated with an Honors grade and not a grade of Satisfactory or less. While 'Prompting' and 'Decisiveness' are overall predominantly used in a negative way (Table 3), the grade awarded on score sheets in which these codes were most frequently applied was Honors. Likewise, while 'Knowledge' and 'Differential Diagnosis' are predominantly used in a positive way (Table 3), the grade most frequently awarded when these codes are applied to the corresponding comments is also Honors.

Discussion

In this analysis we identified thirteen substantive areas on which examiners naturally comment when scoring oral

Table 2
Examples of comments for identified codes.

Code	Negative Valence	Positive Valence
Communication	Needs to work on her presentation - frequently said “whatever' or used hands to put quotes around terms	Articulates her thought process very concisely
Critical Thinking	Difficulty applying principles to practical solutions OR Could not demonstrate an understanding of basic principles of fluid management and patient assessment OR Appeared to have specific algorithm memorized - once off track couldn't recover	Able to think through problems OR Very good thought process OR Very good [...] judgment - beyond years
Decisiveness	Perseverated on irrelevant areas	Very [...] decisive
Demeanor	Sitting too forward and moves around a lot in chair; looked uncomfortable.	At ease, confident, very likeable
Differential Diagnosis	Completely missed differential diagnosis (didn't think of UTI or pregnancy or PID)	Extremely knowledgeable about differential diagnosis in each case
Focus	Not [...] focused on solving problems presented OR History was unfocused	Nicely demonstrated focus[ed] history and physical exam skills OR Able to focus on problems succinctly
Knowledge	[Needs to] Work on fund of knowledge OR Very poor fund of knowledge	Knows everything in the textbook
Management	Failed to do serial abdominal exams on appendicitis patient admitted to the hospital OR Missed management of scalp laceration and importance of preventing hypothermia OR Left out radiation with partial mastectomy	Sound principles trauma and fluid management OR Excellent thought process regarding patient management
Organization	Very unsystematic approach OR Needs to prioritize and organize the history, physical, and testing	Approached clinical problems in a very systematic logical way
Pace	**Too Fast** Needs to slow down OR Need[s] to make sure not to rush ahead too fast **Too Slow** Took too much time	Her pace [...] matched the urgency of the clinical situation
Prompting	Despite tremendous prompting [...] could not demonstrate an understanding of basic principles	Hit all points without prompting
Thoroughness	He did know the basic elements of work-up, but histories were incomplete OR Forg[ot] to do a thorough H&P	Very thorough OR Does not miss details OR Very thorough; thinks through scenario completely
Work Up	Did not know all the avenues for GI bleed work up OR Forgot to do physical exam, needed prompting to get key elements such as GU history in female with abdominal pain OR Had a difficult time navigating through a basic work-up	Assessed and evaluated patient appropriately and performed primary and secondary surveys expeditiously OR Very thorough, orderly and to the point work up OR Excellent history and physical OR Outstanding on ABCs, primary and secondary survey, importance of systematic exam, was efficient in getting labs, baseline x-ray to look for bullet.

examinations used for student evaluation. The developed codes suggest that the oral examination lends itself to assessing students' facility in several of the Accreditation Council on Graduate Medical Education (ACGME) competencies (Table 5) including Patient Care, Medical Knowledge, Interpersonal and Communication Skills, Systems-based Practice, and Professionalism. Several of these areas can be readily tested on less time-intensive examinations, such as multiple choice-based tests, but many, particularly those pertaining

Table 3
Theme Code Use Counts; Valence Counts and Percentages do not add up to Whole Counts and 100%, respectively, as Neutral Counts were Omitted from this Table.

Code	Theme	Used by N (%) Examiners	Code Counts, N (%)	Relative Positive Co-occurrence, N (%)	Relative Negative Co-occurrence, N (%)	Ratio Positive/Negative Co-occurrence (Rounded)
All Codes			2076	1146 (55.2)	879 (42.3)	1:1
Knowledge	Clinical Knowledge	43 (100)	253 (12.2)	186 (73.5)	61 (24.1)	3:1
Management	Clinical Knowledge	41 (95.3)	352 (17)	199 (56.5)	132 (37.5)	1.5:1
Work Up	Clinical Knowledge	38 (88.4)	437 (21.1)	254 (58.1)	173 (39.6)	1.5:1
Critical Thinking	Cognitive Process/ Behavior	37 (86)	156 (7.5)	90 (57.7)	60 (38.5)	1.5:1
Prompting	Cognitive Process/ Behavior	34 (79.1)	179 (8.6)	19 (10.6)	156 (87.2)	1:8
Thoroughness	Cognitive Process/ Behavior	34 (79.1)	218 (10.5)	124 (56.9)	93 (42.7)	1:1
Demeanor	Interpersonal Skills	34 (79.1)	108 (5.2)	70 (64.8)	38 (35.2)	2:1
Organization	Cognitive Process/ Behavior	32 (74.4)	118 (5.7)	71 (60.2)	47 (39.8)	1.5:1
Differential Diagnosis	Clinical Knowledge	31 (72.1)	97 (4.7)	63 (64.9)	31 (32.0)	2:1
Pace	Cognitive Process/ Behavior	22 (51.2)	55 (2.6)	19 (34.5)	36 (65.5)	1:2
Focus	Cognitive Process/ Behavior	21 (48.8)	45 (2.2)	24 (53.3)	21 (46.7)	1:1
Communication	Interpersonal Skills	18 (41.9)	35 (1.7)	24 (68.6)	11 (31.4)	2:1
Decisiveness	Cognitive Process/ Behavior	13 (30.2)	23 (1.1)	3 (13.0)	20 (87.0)	1:6

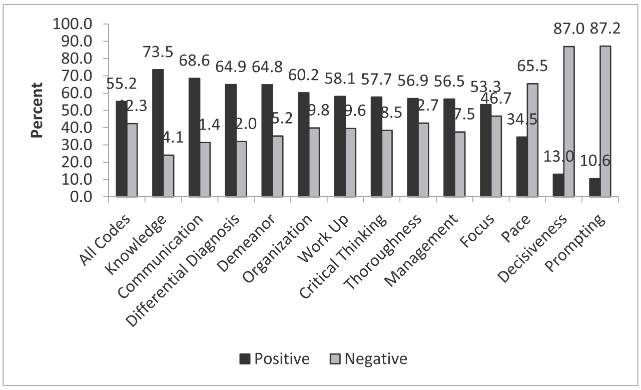

Fig. 1. Valence counts (expressed as percentages) for individual codes.

Table 5

ACGME Competency	Code(s)
Patient Care	Critical Thinking, Management, Work Up
Medical Knowledge	Differential Diagnosis, Knowledge
Practice-based Learning and Improvement	–
Interpersonal and Communication Skills	Communication, Demeanor
Professionalism	Demeanor
Systems-based Practice	Management

to Interpersonal and Communication Skills and Professionalism cannot. The oral examination also lends itself to evaluating students on thought process metrics, namely their focus, organization, thoroughness, pacing, decisiveness, and need for prompting, which are otherwise not well measured but do add to a holistic assessment of students' performance. In fact, of the three codes that appear to be most discriminating, distinguishing high from low performers on the oral examination, 'Demeanor', 'Focus', and 'Organization', suggesting these kinds of qualities may drive performance overall, even on examinations that do not directly measure them. Oral examinations thus may lay bare performance modifiers not otherwise testable. Additionally, in contrast to global evaluations, which tend to overestimate clinical skills[12] and predominantly highlight good performance[13], comments provided in the context of oral examinations are lauding and criticizing performance in near equal measure. This parallels the proportion of supportive and critical comments provided on another evaluation tool relying on immediate, timely feedback on discreetly observed performance, namely the mini-clinical examination exercise (mini-CEX).[13]

Several limitations must be acknowledged. First, the results are drawn from examinations conducted at a single institution and may in some ways represent a biased view of what qualities the Department seeks in its students. Second, the study relies on qualitative data analysis, which is an inherently subjective way in which to analyze data. We hoped to minimize subjectivity bias by using a training process with several analysts, consolidating opinions in a consensus conference, and by conducting crosscheck analysis. The high inter-rater reliability attained suggests that our measures were likely successful. Additionally, since this study is a retrospective review of qualitative comments provided by examiners without specific guidance as to what qualities to focus on, we could not examine if code use truly predicted a grade awarded. Finally, there is evidence that evaluation of performance on oral examinations is influenced not only by the factual content of answers and medical knowledge displayed but also by factors such as manner of speech and appearance, gender, and ethnicity[5,8,9]; Since we did not have detailed demographic data on the students and did not observe examinations, we cannot comment on how bias may have played in to the oral examinations reviewed in this analysis.

Table 4
Codes (and Code Valences) Applied to the Corresponding Comments Provided on the Oral Examination Score Sheets by Grades Awarded; Counts and Percentages do not add up to Whole Counts and 100%, respectively, as Neutral Counts were Omitted from this Table.

		Grades and Corresponding Code Counts, N (%)			
Code (Total Counts, N)	**Valence**	**High Honors**	**Honors**	**Satisfactory**	**Unsatisfactory**
Work Up (352)	Positive	86 (19.7)	132 (30.2)	36 (8.2)	0 (0)
	Negative	8 (1.8)	72 (16.5)	86 (19.7)	6 (1.4)
Management (352)	Positive	67 (19)	99 (28.1)	32 (9.1)	1 (0.3)
	Negative	11 (3.1)	55 (15.6)	58 (16.5)	7 (2)
Knowledge (253)	Positive	70 (27.7)	89 (35.2)	27 (10.7)	0 (0)
	Negative	4 (1.6)	28 (11.1)	22 (8.7)	6 (2.4)
Thoroughness (218)	Positive	69 (31.7)	45 (20.6)	10 (4.6)	0 (0)
	Negative	12 (5.5)	40 (18.3)	39 (17.9)	2 (0.9)
Prompting (179)	Positive	14 (7.8)	4 (20.6)	1 (0.6)	0 (0)
	Negative	11 (6.1)	82 (45.8)	57 (31.8)	6 (3.4)
Critical Thinking (156)	Positive	42 (26.9)	37 (23.7)	11 (7.1)	0 (0)
	Negative	2 (1.3)	27 (17.3)	27 (17.3)	4 (2.6)
Organization (118)	Positive	37 (31.4)	32 (27.1)	2 (1.7)	0 (0)
	Negative	5 (4.2)	20 (16.9)	22 (18.6)	0 (0)
Demeanor (108)	Positive	32 (29.6)	30 (27.8)	7 (6.5)	1 (0.9)
	Negative	6 (5.6)	11 (10.2)	16 (14.8)	5 (4.6)
Differential Diagnosis (97)	Positive	23 (23.7)	32 (33)	8 (8.2)	0 (0)
	Negative	3 (3.1)	11 (11.3)	15 (15.5)	2 (2.1)
Pace (55)	Positive	10 (18.2)	9 (16.4)	0 (0)	0 (0)
	Negative	6 (10.9)	19 (34.5)	11 (20)	0 (0)
Focus (45)	Positive	13 (28.9)	11 (24.4)	0 (0)	0 (0)
	Negative	2 (4.4)	7 (15.6)	12 (26.7)	0 (0)
Communication (35)	Positive	12 (34.3)	11 (31.4)	1 (2.9)	0 (0)
	Negative	3 (8.6)	5 (14.3)	2 (5.7)	1 (2.9)
Decisiveness (23)	Positive	1 (4.3)	2 (8.7)	0 (0)	0 (0)
	Negative	2 (8.7)	8 (34.8)	8 (34.8)	2 (8.7)

Conducting oral examinations is time and labor intensive but despite these limitations and questions about the subjective nature of oral examinations[5,6,9]; they remain part of the traditional cannon of evaluative methods in all stages of surgical training. The results show here that the 'value-added' by oral examinations can be substantial if subjectivity could be limited. Data from this study could then be used to develop a new oral examination scoring rubric anchored in the herein identified qualities that clearly can be reasonably measured by an oral examination to standardize the oral examination process for medical students. Such a rubric would hopefully allow the oral examination to be more meaningful in evaluation.

Funding

This work was supported by a Partners Center of Expertise in Medical Education Grant and the Center for Surgery and Public Health Brigham and Women's Hospital Tracy Arthur Cabot Fellowship.

Acknowledgements

The authors wish to thank Katherine Corso, MS, and Amy Sullivan, PhD for aid in developing coding schemes.

References

1. Griffin WO. *The American Board of Surgery in the 20th Century – Then and Now*. Philadelphia, PA: ABS Office of the Secretary; 2004.
2. Schwartz RW, Sloan DA, Griffin WO, Bland KI, Rhodes RS, Strodel WE. The necessity, practicality, and feasibility of modern educational and evaluative methods for residency training: economic and governing body perspectives. The 1994 Association for Academic Surgery Panel on Education. *Curr Surg*. 1997;54:261–269.
3. General Surgery Certifying Exam. http://www.absurgery.org/default.jsp?certcehome. Last accessed 12/19/2019.
4. Measurement Research Associates. Standardized oral examinations. http://www.measurementresearch.com/media/standardizedoral.pdf. Last accessed 9/7/2011.
5. Haq I, Higham J, Morris R, Dacre J. Effect of ethnicity and gender on performance in undergraduate medical examinations. *Med Educ*. 2005;39(11):1126–1128.
6. Fernandez A, Wang F, Braveman M, Finkas LK, Hauer KE. Impact of student ethnicity and primary childhood language of communication skill assessment in a clinical performance examination. *J Gen Intern Med*. 2007;22(8):1150–1160.
7. Ullman Y, Fodor L, Meilick B, Eshach H, Ramon Y, Meilick A. The oral board examination for plastic surgery: seeking a better way. *Med Teach*. 2006;28:360–364.
8. Burchard KW, Rowland-Morin PA, Coe NP, Garb JL. A surgery oral examination: interrater agreement and the influence of rater characteristics. *Acad Med*. 1995;70(11):1044–1046.
9. Stahlfeld KR. Is the American board of surgery certifying examination archaic? *Curr Surg*. 2004;61:405.
10. Ritchie J, Spencer L, O'Connor W. Carrying out qualitative analysis. In: Ritchie J, Lewis J, eds. *Qualitative Research Practice*. London, UK: SAGE Publications; 2009:219–262.
11. Pope C, Ziebland S, Mays N. Qualitative research in Health Care: analysing qualitative data. *BMJ*. 2000;320(7227):114–116.
12. Silber CG, Nasca TJ, Paskin DL, Eiger G, Robeson M, Veloski JJ. Do global rating forms enable program directors to assess the ACGME competencies? *Acad Med*. 2004;79(6):549–556.
13. Pernar LI, Peyre SE, Warren LE, et al. Mini-clinical evaluation exercise as a student assessment tool in a surgery clerkship: lessons learned from a 5-year experience. *Surgery*. 2011;150(2):272–277.