Global Surgical Education - Journal of the Association for Surgical Education (2023) 2:101
https://doi.org/10.1007/s44186-023-00173-x

ORIGINAL ARTICLE

Item analysis of general surgery multi-institutional mock oral exam: opportunities for quality improvement

Jerome Andres[1] · Ivy A. Huang[1] · Areti Tillou[1] · Justin P. Wagner[1] · Catherine E. Lewis[1] · Farin F. Amersi[2] · Timothy R. Donahue[1] · Formosa C. Chen[1] · James X. Wu[1]

Received: 1 June 2023 / Revised: 2 September 2023 / Accepted: 25 September 2023 / Published online: 28 October 2023

Abstract

Purpose Mock oral examinations (MOE) prepare general surgery residents for the American Board of Surgery Certifying Exam by assessing their medical knowledge and clinical judgement. There is no standard accepted process for quality analysis among MOE content items. Effective questions should correlate with mastery of MOE content, as well as exam passage. Our aim was to identify opportunities for question improvement via item analysis of a standardized MOE.

Methods Retrospective review of testing data from the 2022 Southern California Virtual MOE, which examined 64 general surgery residents from six training programs. Each resident was assessed with 73 exam questions distributed through 12 standardized cases. Study authors indexed questions by clinical topic (e.g. breast, trauma) and competency category (e.g. professionalism, operative approach). We defined MOE passage as mean percentage correct and mean room score within 1 standard deviation of the mean or higher. Questions were assessed for difficulty, discrimination between PGY level, and correlation with MOE passage.

Results Passage rate was 77% overall (49/64 residents), with no differences between postgraduate year (PGY) levels. PGY3 residents answered fewer questions correctly vs PGY4 residents (72% vs 78%, $p < 0.001$) and PGY5 residents (72% vs 82%, $p < 0.001$). Out of 73 total questions, 17 questions (23.2%) significantly correlated with MOE passage or failure. By competency category, these were predominantly related to patient care (52.9%) and operative approach (23.5%), with fewer related to diagnosis (11.8%), professional behavior (5.9%), and decision to operate (5.9%). By clinical topic these were equally distributed between trauma (17.7%), large intestine (17.7%), endocrine (17.7%), and surgical critical care (17.7%), with fewer in breast (11.8%), stomach (11.8%), and pediatric surgery (5.9%). We identified two types of ineffective questions: 1) questions answered correctly by 100% of test-takers with no discriminatory ability ($n = 3$); and 2) questions that varied inversely with exam passage ($n = 11$). In total, 19% (14/73) of exam questions were deemed ineffective.

Conclusions Item analysis of multi-institutional mock oral exam found that 23% of questions correlated with exam passage or failure, effectively discriminating which examinees had mastery of MOE content. We also recognized 19% of questions as ineffective that can be targeted for improvement.

Keywords American Board of Surgery Certifying Exam · Item analysis · Mock oral examination · Exam question development · Resident assessment

Introduction

The American Board of Surgery Certifying Exam (ABS-CE) constitutes the final step of becoming a board-certified general surgeon. The purpose of the exam is to verify that an individual is competent and safe for independent practice by assessing knowledge and decision-making. Mock oral exams (MOEs) are simulations of the ABS-CE for surgical trainees. MOEs closely resemble the certifying exam, and

✉ Jerome Andres
jtandres@mednet.ucla.edu

[1] Department of Surgery, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

[2] Department of Surgery, Cedars-Sinai Medical Center, Los Angeles, CA, USA

MOE performance has also been shown to predict passage of the ABS-CE [1, 2].

The structure and content of the ABS-CE are confidential. Therefore, most MOEs are written by local content experts not affiliated with the American Board of Surgery and lack standardization. Since MOEs must discern an examinee's mastery of surgical knowledge using a limited number of questions, each question should be regularly assessed for effectiveness [3, 4]. Effective questions should incorporate more advanced skills such as synthesis of ideas and problem-solving with less questions assessing lower-order cognitive functions such as recall of information [5]. Exam questions should avoid focusing on tedious details and instead emphasize practical applications of knowledge [6]. Previous studies have shown that item analysis of multiple-choice examinations can distinguish questions by reliability, discriminative ability, and overall difficulty [7–10].

The objective of this study was to perform a thorough item analysis of a large, multi-center MOE. We hypothesized that item analysis would identify questions that best measure content mastery, predict exam passage, and distinguish between high- and low-performing residents.

Methods

Data from a 2022 standardized multi-institutional general surgery MOE with 64 participating residents from 6 institutions were retrospectively reviewed. Exam questions were written by participating faculty on a volunteer basis. All cases were reviewed and revised by study authors for clarity and content. Examinees were asked 73 questions across 12 standardized cases, and examiners would enter whether students would receive a "pass" or "fail" after every question. In the examination, each case consisted of discrete questions. In our analysis, each question was considered a separate item that could be graded as "pass," "fail," and sometimes "critical fail." For each item, examiners were given a rubric with specific criteria for each grade. An item was considered correctly answered if both examiners agreed the examinee passed. At the end of each case, examiners graded the overall performance as "pass," "borderline," or "fail" and were not given specific criteria on the definition of each. Thus, the number of items answered correctly was not directly tied to passage or failure of the entire case. Resident levels (PGY-3, PGY4, PGY-5) were blinded to examiners to avoid bias and unnecessary expectations by amount of training. Furthermore, examiners were paired only with examinees from a differing institution to remove preconceived notions of the resident due to personal familiarity. "Room score" was defined as the mean of the 4 case scores. MOE passage was defined as both of percentage questions correct and mean room score 1 standard deviation below mean or higher. Study authors (JW, JW, FC) categorized questions by clinical topic (surgical critical care, skin and soft tissue, large intestine, stomach, pediatric surgery, breast, endocrine, biliary, trauma) and clinical competency (diagnosis, decision to operate, operative approach, professionalism, patient care, medical knowledge).

Rates of passage were reported for all participants and stratified by PGY using Microsoft Excel. Independent two-sample t-tests were used to compare rates of item passage among pairs of PGY-levels. For all analyses, $p \leq 0.05$ was considered statistically significant.

Item analysis was performed for each test question. We assessed whether answering correctly was associated with MOE passage and whether answering incorrectly was associated with MOE failure using Fisher's exact test. We defined questions as "effective" if there was a significant correlation between answering correctly and passing the MOE because it revealed that the examinee was more likely to have mastery over the MOE content overall. We defined questions as "ineffective" that: 1) had no discriminatory ability (all examinees answered correctly, or all examinees answered incorrectly) and did not have a critical fail option or 2) answering the question correctly was correlated with exam failure.

Results

Exam characteristics and examinee performance

A total of 64 resident examinees, PGY 3–5, from six general surgery residency programs participated in the MOE. Rate of overall MOE passage was 76.7% (49/64) and total percent of items answered correctly was 78.0% (Table 1). There were no statistically significant differences in pass rates by clinical year. By PGY, pass rates were 73.3% (11/15), 78.3% (18/23), and 76.9% (20/26) for PGY-3, 4, and 5, respectively. Overall percentage of items correct was 71.7% for PGY-3s, 77.6% for PGY-4s, and 81.8% for PGY-5s. Differences in

Table 1 Number of exam participants, pass rates, and percent questions correct by PGY

	PGY-3	PGY-4	PGY-5	Total
# Residents	15	23	26	64
Pass rate (%)	73.3% (11/15)	78.3% (18/23)	76.9% (20/26)	76.6% (49/64)
Questions correct (%)	71.7% (728/1015)	77.6% (1142/1472)	81.8% (1474/1801)	78.0% (3344/4288)

performance between PGY-levels were all statistically significant (p < 0.01).

Effective items

Item analysis identified 17 items (23.2%) that were significantly correlated with MOE passage overall. With each question considered individually, residents answering correctly had a pass rate of ranging from 79.7 to 93.8%. Conversely, residents who answered one of these items incorrectly had a pass rate that ranged from 0 to 65.6% (Fig. 1, 2). These items with high discriminatory ability were relatively evenly distributed by clinical topic. Topics included endocrine, surgical critical care, trauma, and large intestine (17.7% each) with fewer in stomach (11.8%), breast (11.8%), and pediatric surgery (5.9%). By clinical competency, most of these items pertained to patient care (52.9%) and operative approach (23.5%), with fewer related to diagnosis (11.8%), decision to operate (5.9%), and professionalism (5.9%) (Fig. 3).

Ineffective items

Item analysis identified 14 items with zero or negative discriminatory ability. First, there were three items with 100% correct rate, and therefore had no discrimination between pass and fail outcomes. Of note, none of these questions had the possibility of "critical fail" that contained elements of knowledge deemed essential to safe practice. Although questions that all examinees answer correctly are not necessarily ineffective, these are unable to discriminate between residents with low rates of content mastery from those with high rates of content mastery within our local cohort. Second, eleven items had a higher pass rate for examinees who answered incorrectly than for those who answered correctly, although these findings did not reach statistical significance (Fig. 4). By clinical topic, ineffective items had a greater proportion of stomach (35.7%) and large intestine (21.4%) with lesser representation of surgical critical care, pediatric surgery, breast, and endocrine (7.1% each). Clinical competencies tested by these items were mostly related to diagnosis (42.9%) and decision to operate (21.4%). Operative approach (14.3%), patient care (14.3%), and professionalism (7.1%) were less represented (Fig. 5).

Discussion

Our item analysis of mock oral examination questions found that only 23.3% of questions were associated with passing the exam, and also identified 19.2% as ineffective It is

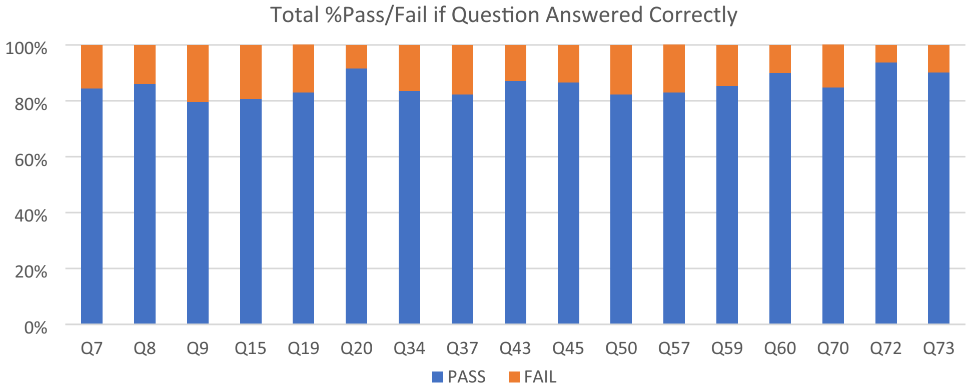


Fig. 1 Percentage correct on predictive questions if answered correctly

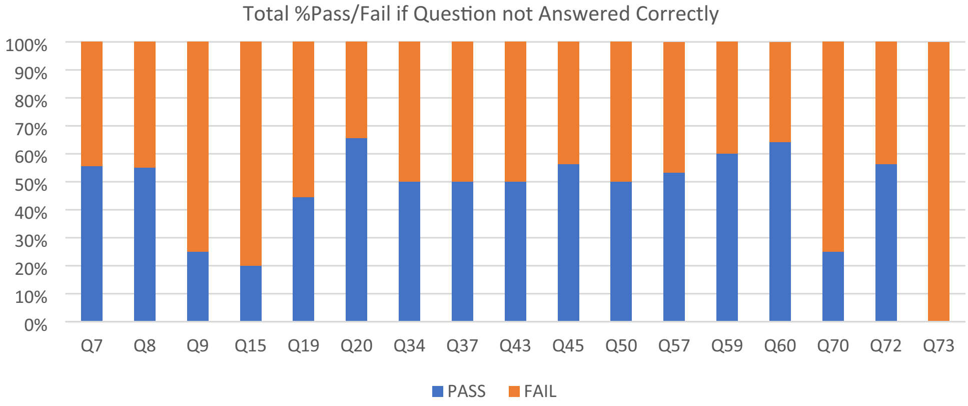


Fig. 2 Percentage correct on predictive questions if not answered correctly

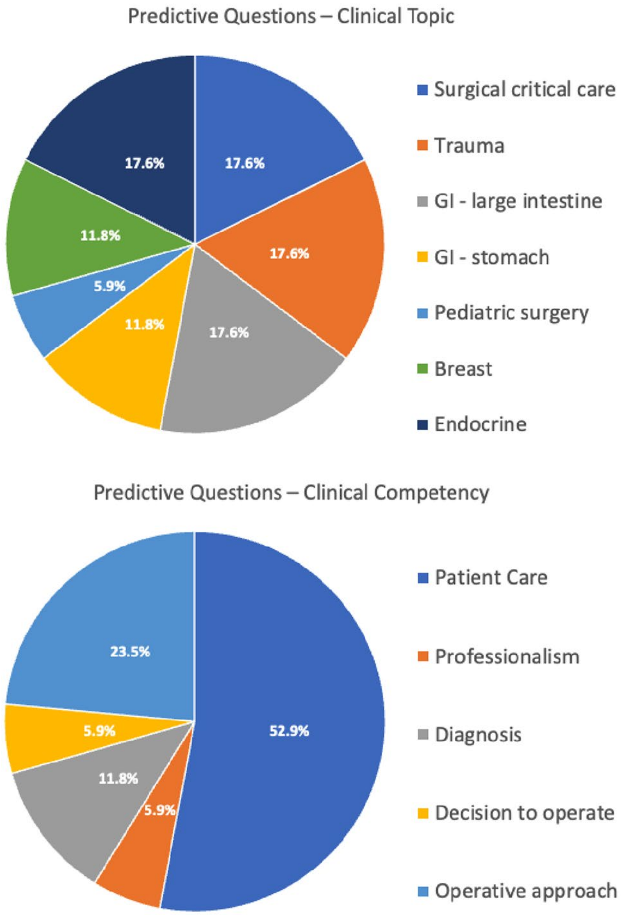


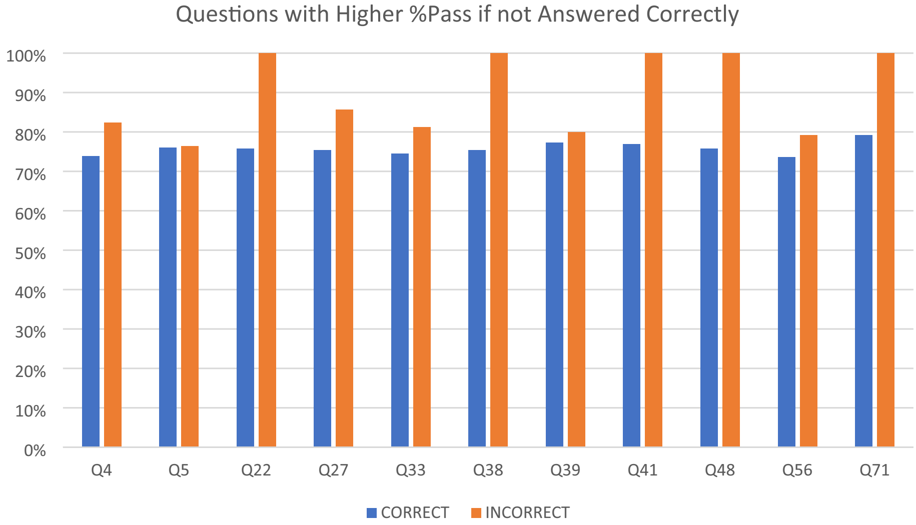
Fig. 3 Distribution of predictive questions by clinical topic and clinical competency

crucial to identify and expand the most effective questions and to minimize or eliminate the least effective ones. Item analysis is a means of quality assurance and improvement for question writing when there is no other standard available. Our study is the first to describe item analysis in MOEs.

Consistent with the results of our study, previous studies have found that item analysis is effective in identifying multiple-choice questions according to discriminatory ability [8–10]. Moderate difficulty questions were most discriminatory whereas excessively difficult or easy questions were less effective [8]. Removal of ineffective questions with negative discrimination (i.e. low achievers performed better than high achievers) from subsequent exams increases validity and reliability [9]. Similarly, another study found that revision or replacement of low-discrimination questions led to increased discrimination and exam quality in subsequent iterations of the exam [10]. Taken together with our results, item analysis is a value practice for general surgery programs writing their own mock oral examination questions to increase the reliability and efficiency of their exam. Easiest items to target for removal or revision are items with negative discrimination or no discriminatory ability. Finally, it is worth noting that a poorly performing item may simply have a poorly written scoring rubric, with nothing to do with topic or the examinee. Nevertheless, item analysis can at least identify problematic items for further scrutiny.

Although eliminating ineffective questions is easy, it is more difficult to discern what makes a question more effective than others. Effectiveness for a general surgery mock oral exam is defined by its ability to simulate the ABS CE, but the content of ABS CE cannot be shared with individuals writing the mock oral exam. Unsurprisingly, mock oral examinations are widely utilized in surgical education, but have untested reliability and validity [11, 13]. In our study, we found that effective questions were relatively evenly distributed throughout clinical topics. We did note that when effective questions were stratified by clinical competency, 76% were related to patient care and operative approach. In

Fig. 4 Questions which displayed greater pass rates if not answered correctly



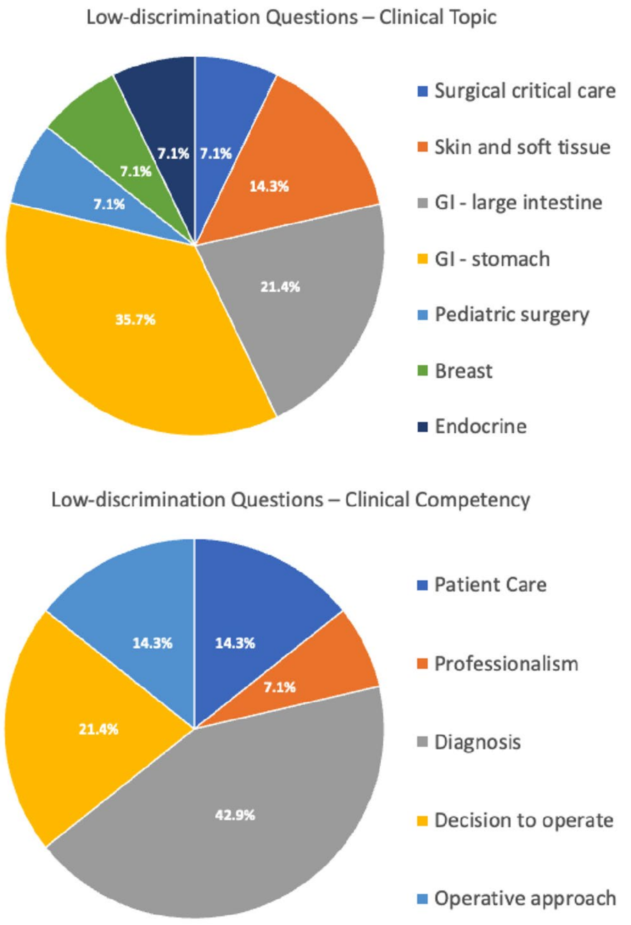


Fig. 5 Distribution of low-discrimination questions by clinical topic and clinical competency

comparison, 64% of ineffective questions related to diagnosis and decision to operate. Thus, future question writers should focus on patient care and operative approach, which are typically learned in the senior years of training. While other areas are important, it is possible these would be better assessed in a different format.

In our analysis of question performance by PGY level, clinical topic, and clinical competency, we found statistically significant differences in performance across all PGY-levels. This finding supports the validity of our MOE as the questions, taken in aggregate, assess knowledge and skills that improve as examinees attain higher levels of training. We also identified 13 questions that may be particularly effective at discerning topics that discriminate among levels of training. Distribution of effective items by clinical topic was relatively even, whereas the greatest proportions of ineffective items were identified among topics of stomach and large intestine. This may be explained in part by the clarity of rubrics provided for these items. It is also plausible that questions regarding stomach and large intestine tested knowledge that is attained at lower PGY-levels, and thus had low discriminatory value. Regarding clinical competencies, patient care and operative approach contained greater proportions of effective items, likely because these are skills that significantly improve with increasing levels of training and clinical exposure. There were relatively greater proportions of ineffective items among categories of diagnosis and decision to operate, which may be in part explained by residency curricula prioritizing diagnostic skills and indications for surgical management earlier in training than other competencies.

Our study has several limitations. First, we cannot assess the true validity of the mock oral examination as it compares to the American Board of Surgery Certifying Exam, and our results are based on exam passage of the mock oral exam. Our study lacks sufficient longitudinal data to analyze the correlation of performance on our MOE with ABS-CE passage. However, with subsequent iterations of the MOE and increasing data available, assessing this relationship may be a promising area of future research. Despite this, multiple studies have found that participating in practice exams significantly improves exam performance [4, 12]. Whether item analysis improves the validity of general surgery MOE, and whether more valid mock oral exams lead to better performance on the ABS-CE are two areas of possible future research. Second, this is based on a limited sample of residents from southern California general surgery programs. A larger study incorporating a nation-wide sample may yield more generalizable results.

Conclusion

Many general surgery residency programs use MOE to prepare residents for the ABS-CE. Therefore, MOEs must be optimized with questions that are reliable, discriminatory, and predictive of overall performance. Item analysis can identify both effective and ineffective questions to guide future exam development.

Data availability Authors are willing to share example of exam contents and de-identified data upon request. Please contact corresponding author Jerome Andres.

Declarations

Conflict of interest No disclosures.

References

1. Fingeret AL, et al. Sequential participation in a multi-institutional mock oral examination is associated with improved American Board of Surgery certifying examination first-time pass rate. J Surg Educ. 2016;73(6):e95–103.
2. Aboulian A, et al. The public mock oral: a useful tool for examinees and the audience in preparation for the American Board of Surgery Certifying Examination. J Surg Educ. 2010;67(1):33–6.
3. Adesope OO, Trevisan DA, Sundararajan N. Rethinking the use of tests: a meta-analysis of practice testing. Rev Educ Res. 2017;87(3):659–701.
4. Balch WR. Practice versus review exams and final exam performance. Teach Psychol. 1998;25(3):181–5.
5. Jones KO, et al. Relationship between examination questions and bloom's taxonomy. In: 2009 39th IEEE frontiers in education conference. IEEE; 2009.
6. Whitcomb ME. The teaching of basic sciences in medical schools. Acad Med. 2006;81(5):413–4.
7. Talebi GA, et al. Item analysis an effective tool for assessing exam quality, designing appropriate exam and determining weakness in teaching. Res Dev Med Educ. 2013;2(2):69–72.
8. Rao C, et al. Item analysis of multiple choice questions: assessing an assessment tool in medical students. Int J Educ Psychol Res. 2016;2(4):201.
9. Khilnani AK, Thaddanee R, Khilnani G. Development of multiple choice question bank in otorhinolaryngology by item analysis: a cross sectional study. Int J Otorhinolaryngol Head Neck Surg. 2019;5(2):449.
10. Smith EB, et al. Auditing radexam: employing psychometrics to improve exam quality. Acad Radiol. 2021;28(10):1389–98.
11. Houston P, Kearney RA, Savoldelli G. The oral examination process—gold standard or fool's gold. Can J Anesth. 2006;53(7):639.
12. Jacobsohn E, Alan Klock P, Avidan M. Poor inter-rater reliability on mock anesthesia oral examinations. Can J Anesth. 2006;53(7):659.
13. Saab SS, et al. Validity study of an end-of-clerkship oral examination in obstetrics and gynecology. J Surg Educ. 2023;80(2):294–301.