A formative midterm test increases accuracy of identifying students at risk of failing a third year surgery clerkship

Julia Corcoran, M.D., M.H.P.E.a,*, Amy L. Halverson, M.D.a, Nancy Schindler, M.D., M.H.P.E.b

aDepartment of Surgery, Feinberg School of Medicine, Northwestern University, 251 E Huron Avenue, Galter 3-150, Chicago, IL 60611, USA; bPritzker School of Medicine, University of Chicago, Chicago, IL, USA

Abstract

BACKGROUND: Providing midclerkship feedback to identify students at risk for failing is a Liaison Committee on Medical Education standard. Objective criteria for that feedback are critical. The investigators studied the value of a formative midterm (MT) test in identifying students at risk for failing a surgery clerkship.

METHODS: A written midclerkship test, which did not contribute to the final grade, was administered (n = 155). The Bayesian specificity, sensitivity, and predictive values for clerkship failure of low MT score, low global clinical performance rating GCPR, and the combination of low MT and low GCPR were computed.

RESULTS: Low MT as a predictor of clerkship failure was sensitive (1.0) but not specific (.35). Likewise, low GCPR was sensitive (1.0) but not specific (.31). The combination of low MT and GCPR, however, was both specific (1.0) and sensitive (.87).

CONCLUSIONS: The addition of an MT test to clinical performance ratings can stratify students’ risk for clerkship failure.

The provision of formal feedback on the basis of assessment early enough in the clerkship rotation to allow sufficient time for remediation is mandated as an education standard by the Liaison Committee on Medical Education (ED-31). This education standard implies that assessment serves as the basis for that feedback. For this feedback to be useful to both faculty members and students, it must be timely and reliable. Students and faculty members should be assured that the correct students are warned of risk for failure and that appropriate learning plans are in place to support academic achievement. Furthermore, the feedback process needs to be believable by both students and faculty members so that they will “take it seriously” and act on the information appropriately.

We provide formal feedback at the midpoint of our clerkship through a student interview with either the clerkship director or the associate director. At that point in the clerkship, we have a clinical performance rating (CPR) from the first block of the clerkship and midterm
(MT) test scores available to review. Additional information available includes the clerkship log. During that interview, the clerkship director or associate director informs each student of progress to date. Students felt to be at risk for failure are asked to submit detailed learning plans within 1 week. Specific academic support may be prescribed as well. The clerkship uses a weighted composite score composed of multiple assessments for pass/fail decisions. Assessments included in the composite score are the National Board of Medical Examiners (NBME) surgery subject examination (SSE), an objective structured clinical examination, global CPRs (GCPRs), and a system of compliance and participation points. The reliability and validity of this system have been reported previously.1 Students fail the clerkship if the composite score falls below the pass/fail cut point determined annually by a standard-setting exercise, as reported previously.2 The purpose of our study was to determine the value of our purely formative MT examination in predicting students at risk for failing the surgery clerkship. We desired an indicator of potential failure that would not penalize students by contributing to the final grade. The literature identifies a clerkship pretest as predictive of academic difficulty in a medicine clerkship3 and an in-clerkship test as predictive of knowledge exam failure at the end of a medicine clerkship.4 We hypothesized that our formative MT examination would be a reliable indicator of possible academic failure in the setting of a composite score used for pass/fail decisions.

### Methods

Faculty members developed a blue printed test of 100 multiple-choice, single-best-answer items and administered it to all enrolled students (n = 155) at the midpoint of the clerkship as a proctored pencil-and-paper test using Scantron answer sheets, which were mechanically scored. We calculated the MT performance parameters (mean, standard deviation, Kuder-Richardson formula 20, and correlation with the NBME SSE) using SPSS Statistics for Mac version 21 (SPSS, Inc, Chicago, IL). We then calculated Bayesian specificity, sensitivity, and predictive values for a failing clerkship composite score of the following assessments available at MT: MT score in lowest tercile, GCPR from the first block in the lowest tercile, and the combination of MT and GCPR in the bottom tercile.

In the academic year under study (2011 to 2012), students achieving composite scores <78 failed the clerkship. The score of the MT did not contribute to the students’ final composite scores. In the formative spirit of the MT, test items were reviewed with students immediately after the examination as a group. All students met with either the clerkship director or the associate director the week after the MT to review their performance on the test and the first CPR.

### Results

#### Student performance

Of the 155 students enrolled in the clerkship for the first time, 151 (97.4%) passed and 4 (2.6%) failed the clerkship.

#### Midterm test parameters

The mean MT score was 58.1 ± 7.38, with scores ranging from 42 to 88. The MT reliability by Kuder-Richardson formula 20 (Cronbach’s α) was .65. Pearson’s correlation coefficient for students’ performance on the MT with their later performance on the NBME SSE was .48 (P < .01, 2 tailed). Pearson’s correlation coefficient for students’ performance on the MT with their final composite score was .44 (P < .01, 2 tailed).

#### Bayesian predictions

The predictive values of low performance on the MT, GCPR, and MT plus GCPR for clerkship failure are presented in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Positive predictive value (likelihood to fail)</th>
<th>Negative predictive value (likelihood to pass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low MT score</td>
<td>1.0</td>
<td>.35</td>
<td>.07</td>
<td>1.0</td>
</tr>
<tr>
<td>Low GCPR</td>
<td>1.0</td>
<td>.31</td>
<td>.08</td>
<td>1.0</td>
</tr>
<tr>
<td>Low MT score plus low GCPR</td>
<td>1.0</td>
<td>.87</td>
<td>.17</td>
<td>1.0</td>
</tr>
</tbody>
</table>

GCPR = global clinical performance rating; MT = midterm.
to provide feedback and direct added academic support would put approximately one-third of students at risk, as indicated by the positive predictive values of .35 and .31. The combination of low performance on both the MT and the GCPR was both sensitive and specific. The positive predictive value in this group was .17, indicating that this one-sixth of students should be scrutinize for the need of additional academic support. On the basis of this information, sound counseling is provided to students at the clerkship midpoint about the potential for failing the surgery clerkship.

**Comments**

In our study, the use of 2 indicators of performance at the midpoint of the clerkship provided sensitive and specific information about potential failure in a surgical clerkship that uses a composite score for pass/fail decisions. Measurement experts encourage the use of multiple appropriate assessments to increase the reliability and validity of grading decisions. Kane and Case recommended the use of composite scoring for the assessment of achievement in complex integrated behaviors that require multiple competencies. With that in mind, it makes sense that the use of a single achievement measure would be less specific than the use of a combination of 2 achievement measures to identify students at risk for failure. By choosing measures that accurately reflect how a student will be graded, the validity of the process is supported. To generalize this study to other clerkship situations, we suggest using more than a single measure of achievement to provide midclerkship feedback and that those measures reflect the types of assessments that will be used for the final grading decisions. Because our clerkship uses knowledge testing in the form of the NBME SSE and clinical application in the form of an objective structured clinical examination and CPRs, a proxy for clinical performance and a proxy for knowledge are good choices.

The consequences for students who fail a required 3rd-year clerkship are serious, including permanent marks on official transcripts, which follow the individuals into the residency match, licensure, and board applications. At our institution, failure of the surgery clerkship in the 2nd half of the 3rd year precludes graduation on time, incurring another year of academic expense and losing a year of employment income. Reliable information must be available to the faculty to inform students, and the information needs to be believable by students so that it will be valued.

For medical students, who are a particularly high achieving group of individuals, a failing grade may be internalized inappropriately, making identification and acceptance of a remediation pathway difficult. A single poor score on an MT test can be explained away because of the amount of call or a bad day. A single poor CPR can likewise be written off as a personality clash. The combination of low scores on 2 measures is harder to dismiss and, in our experience, tends to be taken more seriously by students. We ask all students who perform in the lower tercile on both the MT and the GCPR to turn in learning plans, including a timeline of study, study contents, and planned resources. The responsibility to achieve these goals lies with the student. For students with very severe deficits, recommendation for evaluation by the academic support services in the dean’s office is made.

Another benefit of using multiple measures is the potential to identify areas of strength and areas to work on improvement for middle-achieving students. For students with high achievement on CPRs and low test achievement, MT feedback serves as a reminder that broad surgical knowledge is necessary. Conversely, students who have strong funds of knowledge measured on the MT but low CPRs can be encouraged to step out of their comfort zones and participate more fully in clinical activities.

This benefit for the middle-level achievers points out a limitation of this study. The MT is itself an intervention. By providing MT assessments, formal feedback, and, when indicated, forcing the creation of a true learning plan, we cannot design a blinded study. With the Liaison Committee on Medical Education ED-31 requirement of formal, critical feedback with sufficient time to allow correction of deficiencies, a randomized crossover study design is neither feasible nor ethical.

**Conclusions**

We advocate the use of multiple measures not only for clerkship grade assignment but also for midclerkship feedback and identification of at-risk or struggling students. The addition of a written test to CPRs at the midpoint of a surgery clerkship that uses a composite score for pass/fail decision making is both sensitive and specific. The combination allows the identification and remediation of a specific subpopulation within the clerkship cohort.

**References**