Creating a Multi-Institutional Family Medicine Clerkship Examination: Lessons Learned

Lisa M. Slatt, MEd; Beat D. Steiner, MD, MPH; David W. Hollar, PhD; Alexander W. Chessman, MD; Pamela Wiseman, MD; Julia Xin; Mari-Wells Hedgpeth, MEd

BACKGROUND AND OBJECTIVES: Multiple choice examinations assess learners’ attainment of medical knowledge. Developing multiple choice examinations that discriminate among learners is difficult and time-consuming. Many institutions avoid this effort by using the National Board of Medical Examiners (NBME) subject examinations, which can also provide comparisons to a national norm. The family medicine subject examination has been criticized, however, because the test’s content does not reflect the learning expected during the clerkship. Additionally, the test results cannot guide clerkship directors sufficiently to help students study or to improve the curriculum.

METHODS: Family medicine clerkships at three different institutions used a common 75-item examination based on the textbook Essentials of Family Medicine, Fifth Edition, for one academic year. Data were pooled and analyzed. The Rasch Item Response Theory assessed student and item performance.

RESULTS: A total of 451 students took the examination. Across the three schools: (1) item separations (Rasch) were high (8.64), indicating good spread in item difficulty, (2) person separations were lower (1.65), indicating that medical students are likely a relatively homogeneous group, (3) Rasch item reliabilities were strong (ranging from .96–.99), and (4) Rasch person reliabilities (.54–.73) were lower. True internal consistencies across items as measured by the Kuder-Richardson 20 (KR-20) reliabilities were just adequate at .71–.77.

CONCLUSIONS: By pooling resources, clerkship directors can share the creation and implementation of a written examination that has acceptable reliability and greater face validity than the NBME subject examination. They also have more control over examination content and can guide students’ learning and curriculum improvements more accurately.

From the Department of Family Medicine (Ms Slatt, Dr Steiner), Office of Medical Education (Dr Hollar, Ms Xin, Ms Hedgpeth), University of North Carolina at Chapel Hill; Department of Family Medicine, Medical University of South Carolina (Dr Chessman); and Department of Family and Community Medicine, Tulane University (Dr Wiseman).
competence that complement each other. One option to connect objectives, activities, and assessment is to use a required textbook. A textbook can deliver the clerkship’s core content, even when students spend the rotation at geographically distant sites. Requiring students to focus on one assigned textbook may penalize them, however, if the NBME subject examination covers content that is not included in that text. As a result, many family medicine clerkship directors have continued to assess their students’ clinical knowledge with MCQ examinations written by faculty ill-prepared to write good questions and without the resources to study MCQ item performance to improve the test. Such non-evaluated MCQ examinations may not assess student knowledge accurately.

The purpose of this study was to determine if a consortium of family medicine clerkship directors at three institutions could collaborate using MCQ items developed for a common set of curricular outcomes drawn from one single textbook, pool their data, and verify that the MCQ examination meets standards for both a reliable and valid instrument.

Methods
This study describes how the consortium of family medicine clerkship directors at three institutions could collaborate using MCQ items developed for a common set of curricular outcomes drawn from the Essentials of Family Medicine, Fifth Edition (Sloane, P., Slatt, L., Ebell, M., Jacques, L., Smith, M., eds. Baltimore: Lippincott Williams & Wilkins, 2008) as their required clerkship textbook would be interested in attending a special meeting to discuss forming an examination consortium. The consortium would use questions based on the Essentials textbook, and results would be pooled. The planning meeting would be held at the annual STFM Predoctoral Education Conference in January 2008. To join the consortium, schools had to agree to use all test items, to not change the wording of the items, and to report their data on an Excel spreadsheet using a common template.

Test Development
During the spring of 2008, one of the authors (BS) created a new 75-item test. Items were drawn from a pool of 100 items used the previous academic year at two institutions. Item selection was based on how well they covered the necessary content domain and medical knowledge learning objectives required for the family medicine clerkship at the author’s school. In addition, items that performed with a difficulty index of $\geq .75$ and a discrimination index of $\geq .29$ in the pilot were chosen. All 75 items were reviewed by the author. The format of the questions were all case-based, single best answer questions using examination writing guidelines proposed by the NBME.

A draft of the test was e-mailed to the consortium members in early summer 2008 for review. Changes were sent to the author, who had final authority to accept or reject suggestions. A final version of the test was e-mailed to the participants. All schools began using the test at the beginning of the 2008–2009 academic year.

To assure that the examination was addressing a range of problems and physician tasks, each item was classified in two dimensions. In one dimension, items were classified as either acute, chronic, or prevention. In the second dimension, items were classified as related to creating a differential diagnosis or committing to a management plan.

Student performance between July 2008 and July 2009 on all 75 items were collected from all three schools and analyzed.

This study received approval from the Institutional Review Board at the University of North Carolina at Chapel Hill. One of the lead editors of the textbook (LS) sent out the e-mail to invite schools into the collaborative and also served as project manager. While these roles may represent a conflict of interest, she was not directly involved in the statistical analysis of the data.

Analysis
To assess whether the 75-item test was robust, we performed a sequential set of analyses: (1) Calculations of item difficulty and discrimination indices, (2) Rasch analysis, a theoretical approach that compares item responses based on two factors: item difficulty and differences in abilities of the test takers. For the latter method, item difficulties were mapped across levels of student knowledge/performance.

The difficulty index (the $P$ value) measures the proportion of examinees who answered the item correctly. It can range between 0.0 and 1.0, with a higher value indicating that a greater proportion of examinees responded to the item correctly, and it was thus an easier item measure. A high difficulty index (eg, 0.95) reflects an easy question.

The item discrimination index measures how well an item is able to distinguish between examinees who are knowledgeable and those who are not. It looks at the relationship between an examinee’s performance on the given item (correct or incorrect) and the examinee’s score on the overall test. For an item that is highly discriminating, in general the examinees who responded to the item correctly also did well on the test, while in general the examinees who responded to the item incorrectly also tended to do poorly on the overall test. The possible range of the discrimination index is -1.0 to 1.0; an item with discrimination below 0.0 suggests a problem. When an item is discriminating negatively, overall the most knowledgeable examinees are getting the item wrong.
and the least knowledgeable examinees are getting the item right. A negative discrimination index may indicate that the item is measuring something other than what the rest of the test is measuring. For this study, items with a discrimination index above 0.30 reflect the capacity of the item to distinguish between high and lower ability students.

There is a relationship between an item’s difficulty index and its discrimination index. If an item has a very high (or very low) P value, the potential value of the discrimination index will be much less than if the item has a mid-range P value.

A useful approach when reviewing a set of item discrimination indexes is to also view each item’s P value at the same time. For example, if a given item has a discrimination index below 0.1, but the item’s P value is greater than 0.9, you may interpret the item as being easy for almost the entire set of examinees and probably for that reason not providing much discrimination between high ability and low ability examinees.

A statistical power analysis detects if there was an effect, given that the effect actually exists and relies on the sample size to achieve adequate power to detect the effect. For our power analysis, we used Cohen’s tables and G*Power software. We set our power at 0.85 to detect if there was an effect, given that the effect actually exists.

Item Response Theory (IRT) is a modern testing theory. It is used for test development, test score equalities, and identification of biased items. IRT is based on the idea that the probability of a correct/keyed response to an item is a mathematical function of the person taking the test and item parameters. For this study, the Rasch IRT model was selected, and the WINSTEPS software was used. For items and individuals (persons), the model incorporated these measures: mean, standard deviations, reliability, and Kuder-Richardson Formula 20 (KR-20). This is a measure of internal consistency for dichotomous choices. A high KR-20 coefficient (eg, >0.90) indicates that the items precisely measure a single measurement concept (ie, that the items agree with each other and can be used in various combinations). Values can range from 0.00 to 1.00 (sometimes expressed as 0 to 100), with high values indicating that the examination is likely to correlate with alternate forms (a desirable characteristic).

**Results**

Thirteen schools attended the STFM Predoctoral Education Conference recruitment meeting in January 2008. Three schools agreed to the conditions for joining the consortium, and Table 1 illustrates their characteristics. Eight schools chose to use only a subset of the 75 questions that reflected the objectives for their specific clerkships.

Of the 75 questions, 39 dealt with acute concerns, 24 related to chronic conditions, and 12 addressed prevention topics. For physician tasks, there were 11 differential diagnosis questions and 64 management questions.

The components of the Rasch IRT model measurements for overall test item and student performance patterns (persons) are summarized in Table 2. Item separations (Rasch) were high (8.64) across the three schools, indicating good spread in difficulties of items. Person separations (eg, 1.65 for all three schools combined) were lower, but again, medical students tend to be relatively homogeneous in terms of high ability/knowledge. Rasch item reliabilities were strong (.96–.99) across schools, although Rasch person reliabilities (.54–.73) across schools were lower, again indicative of the students being very similar with high knowledge/abilities. True internal consistencies across items for the test were reflected by the test Kuder-Richardson 20 (KR-20) reliabilities, which were just adequate at .71–.77 across the three schools (Table 2).

Mean scores per institution are shown on a 100-point scale in Table 3, compared to Rasch standardized means shown in Table 2. On the 100-point scale, school means were significantly different (F2,445=51.3, P<.000). School A students had significantly higher mean scores (74.5) than Schools B and C (both 65.7).

**Discussion**

This study shows that a consortium of family medicine clerkship directors at three institutions can successfully pool data from a common MCQ examination based on a textbook used by the members, and faculty can prove that the test is reliable, while ensuring the content validity. The Rasch IRT assessments of

<table>
<thead>
<tr>
<th>Institution</th>
<th>Type of School</th>
<th>Length of Clerkship in Weeks</th>
<th>Number of Students</th>
<th>Method of Test Administration</th>
<th>Time Allotted for the Examination</th>
<th>% That Examination Contributed to Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>School A</td>
<td>Public</td>
<td>4</td>
<td>158</td>
<td>Written</td>
<td>90</td>
<td>15</td>
</tr>
<tr>
<td>School B</td>
<td>Public</td>
<td>6</td>
<td>148</td>
<td>Computer</td>
<td>90</td>
<td>15</td>
</tr>
<tr>
<td>School C</td>
<td>Private</td>
<td>6</td>
<td>144</td>
<td>Written</td>
<td>120</td>
<td>30</td>
</tr>
</tbody>
</table>
the test indicate that it has excellent item separation and internal consistency of items, and the person separation is adequate given the high academic skills of this population (Table 2). Further, item spreads were comparable from school to school. The item discrimination and difficulty indices indicate avenues for specific item improvements for future clerkships and testing as well as for selection of items for shorter tests. These statistics should be reassuring to medical student deans, faculty, and students concerned about the quality of multiple test examinations not created by the NBME.

One negative aspect of a consortium's examination relates to the limited bank of questions. Our consortium did not create a pool of questions larger than the examination itself, so there was no alternate version offered to students having to retake the examination. Also the need to use the same version of the examination throughout the year decreases test security. This is not the case with the NBME subject examination question bank, which is extensive.

Unlike the NBME subject examination, a consortium-based test can provide results that inform curricular change. One school had a higher mean of 74.5 compared to the other two schools, with identical means of 64.7. This difference did not likely result because of different emphasis on the examination, because the one school that counted the examination for 30% rather than 15% of the grade was one of the two schools with a lower score. We hope to investigate this finding more fully to determine if the difference in the mean can be attributed to student or curriculum characteristics. If we could isolate features of a curriculum at one school that improves scores, then the consortium's examination results might be of benefit to other schools in redesigning their curricula.

This consortium proves that it is possible to create a family medicine clerkship examination with robust test characteristics. Creating such an examination offers the advantages of tailoring the assessment to the specific outcomes and teaching resources of the clerkship. However, the effort and cost required to create and maintain such an examination are substantial and for many institutions may not be worthwhile.

ACKNOWLEDGEMENTS: The authors thank the following individuals for their help in carrying out this project: Doug Bower, MD; Kelly Everard, PhD; Katie Margo, MD; Susan Stangl, MD; Anne Walsh, PA-C; and Terri Whoehrle, MD. The Institutional Review Board at the University of North Carolina at Chapel Hill reviewed this project and granted exempt status. Data from this study were presented at the 2010 Society of Teachers of Family Medicine Annual Spring Conference in Vancouver.

Table 2: Overall Rasch IRT Measurement Summary

<table>
<thead>
<tr>
<th>Statistic</th>
<th>All Items</th>
<th>School A</th>
<th>School B</th>
<th>School C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>69.8</td>
<td>51.7</td>
<td>48.6</td>
<td>49.3</td>
</tr>
<tr>
<td>SD</td>
<td>9.68</td>
<td>9.68</td>
<td>9.68</td>
<td>9.68</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
<td>1.54</td>
<td>3.01</td>
<td>2.93</td>
<td>2.94</td>
</tr>
<tr>
<td>Rasch Separation</td>
<td>8.64</td>
<td>1.65</td>
<td>1.07</td>
<td>1.64</td>
</tr>
<tr>
<td>Reliability</td>
<td>.99</td>
<td>.73</td>
<td>.54</td>
<td>.73</td>
</tr>
<tr>
<td>Kuder Richardson -20</td>
<td>—</td>
<td>.77</td>
<td>—</td>
<td>.73</td>
</tr>
</tbody>
</table>

Table 3: Mean Scores, Percentage Answered Correctly Per Institution

<table>
<thead>
<tr>
<th>Institution</th>
<th>Mean (Based on 100 Points)</th>
<th>n</th>
<th>Standard Deviation</th>
<th>Standard Error of Mean</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>School A</td>
<td>74.5</td>
<td>158</td>
<td>8.15</td>
<td>.64</td>
<td>75</td>
</tr>
<tr>
<td>School B</td>
<td>65.7</td>
<td>149</td>
<td>8.97</td>
<td>.73</td>
<td>75</td>
</tr>
<tr>
<td>School C</td>
<td>65.7</td>
<td>144</td>
<td>9.16</td>
<td>.76</td>
<td>75</td>
</tr>
<tr>
<td>Total</td>
<td>68.8</td>
<td>451</td>
<td>9.68</td>
<td>.46</td>
<td>75</td>
</tr>
</tbody>
</table>
CORRESPONDING AUTHOR: Address correspondence to Ms Slatt, University of North Carolina, Department of Family Medicine, 590 Manning Drive, Chapel Hill, NC 27599-7595. 919-966-3912. slatt@med.unc.edu.

References