Assessment-based health informatics curriculum improvement

Eta S Berner,1 Amanda D Dorsey,2 Robert L Garrie,2 Haiyan Qu2

ABSTRACT

Objective Informatics programs need assurance that their curricula prepare students for intended roles as well as ensuring that students have mastered the appropriate competencies. The objective of this study is to describe a method for using assessment data to identify areas for curriculum, student selection, and assessment improvement.

Materials and Methods A multiple-choice examination covering the content in the Commission for Health Accreditation of Informatics and Information Management Education curricular facets/elements was developed and administered to 2 cohorts of entering students prior to the beginning of the program and to the first cohort after completion of the first year’s courses. The reliability of the examination was assessed using Cronbach’s alpha. Content validity was assessed by having 2 raters assess the match of the items to the Commission for Health Accreditation of Informatics and Information Management Education requirements. Construct validation included comparison of exam performance of instructed vs uninstructed students. Criterion-related validity was assessed by examining the relationship of background characteristics to exam performance and by comparing examination performance to graduate Grade Point Average (GPA).

Results Reliability of the examination was 0.91 and 0.82 (Cohort 1 pre/post-tests) and 0.43 (Cohort 2 pretest). Both raters judged 76% of the test items as appropriate. There were statistically significant differences between the instructed (Cohort 1 post-test) and uninstructed (Cohort 2 pretest) students ($t = 2.95, P < .01$), as well as between the Cohort 1 pre/post-tests ($t = 6.52, P < .001$). Neither the background variables nor the graduate GPA were significantly correlated with the examination scores.

Conclusion We found that the examination had generally good psychometric properties and the exceptions could be used to identify areas for curriculum and assessment improvement.

Keywords: education, curriculum, informatics, examination, assessment

BACKGROUND AND SIGNIFICANCE

In recent years, there has been a growing number of informatics education programs. For over 30 years, the National Library of Medicine training programs have produced informatics researchers and faculty,1 but recently, there have been new programs such as the American Council for Graduate Medical Education fellowship programs that prepare physicians for certification in the new Clinical Informatics subspecialty.2–5 the professional masters’ programs that are the target of new accreditation initiatives by the Commission on Accreditation for Health Informatics and Information Management Education (CAHIIM),6,7 and the Workforce programs facilitated by funding from the Office of the National Coordinator for Health IT and the Health Resources and Services Administration.8,9 With the growing diversity of informatics roles and functions, educational programs need to ensure that the varied curricula are designed to prepare students for the intended roles. In addition, programs need to ensure that students have mastered the appropriate competencies.

Student assessment and curriculum assessment are intertwined. Review of the psychometric properties of an examination has traditionally involved examining reliability and validity, including content, construct, and criterion-related validity.10 More recent views see construct validity as the overriding concept, with reliability reflecting the structure of the exam and content and criterion-related validity evidence falling under construct validity or content validity is included under the rubric of translational validity.11–13 Both the traditional and the more modern views emphasize the importance of systematically collecting evidence to determine if the examination measures what is intended for its given purpose.12 This evidence can also inform curriculum analysis and improvement. For instance, a valid assessment should measure the knowledge and skills that are the goals of the curriculum, and the curriculum should also embody those skills. An evaluation of the match between the assessment and the competencies that the curriculum is designed to embody (content validation) can inform improvement of both the assessment and the curriculum. The process of test development involves an iterative process of hypothesizing relationships, collecting evidence to test those hypotheses, and refining the test accordingly.12

For instance, if the assessment addresses the content of the curriculum and the curriculum is well taught, the students should improve their examination scores after going through the curriculum. If they do, it lends support to both the construct validity of the assessment and the appropriateness of the curriculum; if there is no improvement, the results can target curricular areas that may need changing.

If the assessment is valid, another reasonable hypothesis is that, prior to taking the curriculum, students who enter the program with more relevant experience should perform better on the assessment. Conversely, students who do well on a relevant pre-enrollment assessment should perform better in the program. Again, assessment of the criterion-related validity of the examination that examines these hypotheses can highlight problematic areas in the exam itself and/or the curriculum.
OBJECTIVE
In this article, we describe the examination development and validation process for a comprehensive exam for students in a professional master’s program and how we are using the results of the examination to identify areas for curricular, student selection, and assessment improvement. As part of our development process for a comprehensive examination covering the core content of our first year core curriculum, we administered the exam to 2 cohorts of entering students prior to their beginning the program and administered it again to the first cohort after they had completed the first year’s courses. We used standard psychometric test validation methods to explore aspects of the examination and curriculum that need improvement.

The methods we are using to develop the exam can be used by others with different curricula and different goals to assess both students and their informatics curricula. The lessons learned in applying general educational evaluation methods to an informatics education program will be addressed. This study was approved as Exempt by the University of Alabama at Birmingham Institutional Review Board.

MATERIALS AND METHODS
Curriculum design
We described the evolution of the Master of Science in Health Informatics program previously. The current program has a Core/Track Model wherein all students begin the first year of the program with the “core” informatics courses (comprised of 6 Informatics courses plus 3 leadership seminars, all based on the existing CAHIIM Facets and Curricular Components). After completing the core courses, the students begin their second year of coursework in their respective specialty tracks: Data Analytics, User Experience, or Health Informatics Management Leadership. After identifying the courses in the core curriculum, the faculty set about the task of mapping the content of each course to the CAHIIM standards for Health Informatics Master’s Degrees. This mapping process led to a systematic way of ensuring consistency of content being taught, and also demonstrated to our regional accrediting body, the Southern Association of Colleges and Schools, that our student learning outcomes were both reasonable and measurable. The creation of the core/track model also necessitated a method whereby we could assess student learning outcomes after the students had completed the courses in the first year core curriculum, and before they entered the specialty tracks.

Examination design
A 3-person faculty committee was tasked with developing a comprehensive multiple-choice examination to be given to entering masters’ students as a pretest and again as a post-test after they took the first year core curriculum. Using the CAHIIM curricular elements as a guide, we consulted 3 sources of questions related to general informatics or health information technology competencies. Sources included the questions developed as part of the Office of the National Coordinator for Health IT curriculum materials sample questions available as a study guide for the Healthcare Information and Management Systems Society Certified Professional in Health Information & Management Systems (HIMSS CPHIMS) exam, and questions from health information technology and health informatics textbooks. We aimed to select questions that measured higher cognitive levels—i.e., analytic or application questions rather than simple recall—but that was not always possible. The exam development team mapped the questions to the CAHIIM facets/elements, edited them when needed to improve their quality, and included those where there was consensus among the 3 informatics faculty as to the appropriateness of the questions. We did not specify a certain number of questions per CAHIIM facet/element. Generally, the judgment of appropriateness was made on the basis of the content being included in the CAHIIM facets/elements, rather than on the cognitive level.

The exam was administered in August 2014 to students entering our master’s program (referred to below as First Cohort or Cohort 1) and again to the same students after they had completed their core courses. The same exam, with minor edits on some questions, was administered to a second cohort of students (Cohort 2) who entered the program in August 2015. After the first examination was administered, we reviewed the exam item analysis statistics (difficulty index (proportion of students getting the item correct) and discrimination index (point-bi-serial correlation between total score on the exam and getting the item correct)) and deleted, gave credit for all answers, or changed the answer keys on any questions that were judged to be ambiguous or very unclear, then rescored the exam. This approach is a recommended best practice in test administration. The student scores were based on the second administration, and descriptive statistics (mean, median, standard deviation, minimum, and maximum scores) were computed.

The faculty committee reviewed all test questions prior to the second administration and made some changes to improve the test. Because we wanted to keep the pre- and post-test basically the same, as we were preparing the post-test, we were very careful to eliminate or modify only questions that on review had serious problems. For the post-test for Cohort 1, 3 questions that were on the pre-test were eliminated, 3 minor typographical errors were fixed, and 1 question was significantly reworded. More changes for improvement were made on the pretest for Cohort 2. Five questions that had been eliminated from the both the pre- and post-test for Cohort 1 on the first scoring had minor wording changes for clarification and were used with Cohort 2. In addition, 11 questions on the tests for Cohort 1 had minor wording changes for Cohort 2, and 2 questions on the Cohort 1 tests were eliminated for Cohort 2. Since we are still exploring the exam performance, the examination is currently given in a “low stakes” mode, where it does not count for advancement in the curriculum or affect student grades. The final number of items on the 3 exams was 146 for the pretest for Cohort 1, 145 for the post-test for Cohort 1, and 147 for the pretest for Cohort 2.

Reliability of all administrations of the examination was assessed using Cronbach’s alpha. This statistic is standard output on many test scoring programs and was produced by the program we used.

Content validity
In addition to the informal content validation (consensus of 3 informatics faculty based on discussion), a formal assessment of content validation of the examination was done using the following procedures described by Martuza. Two Health Informatics faculty members who were not involved in the examination development rated the match of each item to the CAHIIM requirements using the following scale: item was not at all relevant (to CAHIIM content), item was slightly relevant, item was relevant, and item was very relevant. The items that were rated not and slightly relevant were considered “not a match,” and relevant and very relevant were considered a “match” of exam items and CAHIIM content. Inter-rater agreement and percent of “matches” were calculated. Although Martuza did not include it, we also calculated Cohen’s kappa as a measure of agreement. The higher the percentage of agreement and the percentage of matches, the better the content validity.

Construct validity
As discussed above, construct validity involves assembling evidence to support a particular use of a test, e.g., as a measure of the content
of our core curriculum. We looked at several types of evidence as measures of construct validity. For several of these measures, we used the contrasted groups approach. As described in DeVon et al., this approach to construct validity involves comparing groups known, or expected, to differ in the construct being measured on the assessment instrument. If the assessment is valid, the test scores should demonstrate those differences. In our case, if the examination was a valid measure of what was taught in the core curriculum, one would expect students who had gone through the core curriculum to perform better on the examination than comparable students who did not go through it. That is, instructed students should perform better than un instructed students. To assure that the first and second cohorts were comparable, we compared the pretest means on the examination from the first and second cohorts using the independent samples t-test. To assess construct validity, we used a paired samples (also known as repeated measures) t-test to compare the mean scores on the examination pre- and post-test for the first cohort. We used an independent samples t-test to compare mean test scores between the instructed students (Cohort 1 post-test) and a different group of un instructed students (Cohort 2 pretest).

Criterion-related validity

We examined the relationship between relevant background characteristics (i.e., undergraduate GPA, years of health care experience) on performance on the examination. If the exam is valid, one would expect these variables to relate to the examination score. We correlated exam scores with undergraduate GPA and years of health care experience. We also included these variables as independent variables in a multiple regression analysis with exam score as the dependent variable. We also included these variables as independent variables in a multiple regression analysis with exam score as the dependent variable.

In addition, if the exam is a valid assessment of core course content, comprehensive exam scores should predict GPA for the core courses. To assess this relationship, we correlated exam scores with first-year GPA for the first cohort.

RESULTS

Students

Cohort 1 began the masters’ program in 2014 with 14 students. Two students dropped out and 1 additional student did not take the post-test, leaving 11 students who took the post-test. Cohort 2 entered in 2015 with 15 students. Table 1 lists the characteristics of the students.

Examination performance

Table 2 shows the descriptive statistics of the examination. There was not a statistically significant difference in the pretest scores between Cohort 1 (M = 55.00, SD = 10.32) and Cohort 2 (M = 57.87, SD = 4.81; P > .05). Upon inspection of the standard deviations, it appeared that, although the means were not different, the variances between the cohorts might be different, since Cohort 2 had a standard deviation approximately half that of Cohort 1 on the pretest. In further support of potential differences in variance, Cohort 1 had a range of pretest scores from 32 to 69, whereas Cohort 2 had a range of 48 to 66. We tested whether this difference was statistically significant, and there was indeed a statistically significant difference in the standard deviations of the 2 cohorts (F = 4.60, P < .05). There was also a statistically significant difference in the expected direction for the exam scores for Cohort 1’s pretest (M = 55, SD = 10.32) and post-test (M = 65.36, SD = 8.11; P < .05). Because of the small sample sizes, we also performed a variety of nonparametric tests that can be used with small sample sizes or if the data do not satisfy the assumptions for the parametric tests (i.e., One-Sample Kolmogorov–Smirnov test, Independent Samples Mann-Whitney U-test), which showed the same results (data not shown).

Reliability

The reliability of the pretest given to Cohort 1 on entering was 0.91, and the post-test reliability was 0.82. The reliability of the pretest for Cohort 2 was 0.43.

CONTENT VALIDITY

Table 3 shows the comparison of the ratings given by the 2 reviewers for the test questions compared to the CAHIIM curricular elements.

Inter-rater agreement was 0.8, with a kappa of 0.21 (P < .05), indicating that for 80% of the items (118/147), the reviewers agreed that the item either did or did not match the CAHIIM content. The content validity index of 0.76 (112/147) indicates both reviewers thought that over three quarters of the items were a match to the CAHIIM content and both raters agreed that only 4% of the items (6/147) did not match the CAHIIM content. Table 3 also shows most of the disagreement comes from Reviewer 2, who indicated that an additional 17% of the items did not match the CAHIIM content, while Reviewer 1 judged those same items as matching the content. There were only 4 items that Reviewer 1 judged as not relevant that Reviewer 2 thought were relevant. From these data, the overall content validity of the examination appears good.

Construct validity

Comparison of the same group (Cohort 1) on their pretest and post-test scores (Table 2) shows that the mean percent correct scores on the post-test increased by 10 points (55–65), which was a statistically significant increase (P < .001). In addition, each student showed some increase in scores (data not shown), with a range of 3 to 18
The hypothesis that undergraduate GPA and health care experience would relate to test scores was not supported. There were no statistically significant correlations between these background variables and student test scores. The correlation of the pre- and post-tests was 0.86 ($P < .01$), which is consistent with the high reliability of both tests.

The comparison between the instructed group (Cohort 1 post-test) and a second uninstructed group (Cohort 2 pretest) also showed a statistically significant difference ($P < .01$). These data lend support to the construct validity of the exam in that the students who went through the curriculum performed significantly better on an examination designed to measure the content of the curriculum than did students who did not go through the curriculum.

**DISCUSSION**

The data provide some evidence that the examination performed as intended. The examination showed good reliability for Cohort 1 during both administrations, despite the cohort being relatively small. Although the reliability was lower for Cohort 2, it is likely to be because Cohort 2 was not only small, but also because the scores were statistically significantly more homogeneous than Cohort 1. The reliability is affected by the variance among the scores, and the smaller variance in Cohort 2 may have made it more difficult to demonstrate high reliability. Ironically, since our intent is to admit students with strong academic credentials and relevant work experience, as we improve our selectivity of students, we are likely to increase the homogeneity of student backgrounds, which is likely to make the groups’ test scores more homogeneous as well, and that can affect the psychometric properties of the examination.

The content validity of the examination was good, with 76% of the items judged by two independent raters as matching the intended content. However, it is the remaining 24% of the items that can be useful for reviewing the curriculum and/or revising the examination in the future. These items that 1 or both raters felt were not measuring the relevant content need to be discussed by a broader group of faculty to determine if they really should be included in the exam or if they can be modified so that they more closely match the desired content. In this way, the examination can be improved to be a better match to the desired content. Partly because of time pressure, we did not do the content validation before the examination was finalized for the first cohort of students, except informally, by the exam development team reaching consensus on the items. However, for formative evaluation to help design a better assessment tool, it would be valuable to do a formal content validity assessment before the exam is finalized and use the data to improve the exam before it is administered.

As the accreditation standards for health informatics education programs change, as they are expected to do, examination content will also need to be modified, and the content validation process will have to be repeated. If the informatics program is a National Library of Medicine program designed to train informatics researchers or an American Council for Graduate Medical Education–accredited program whose aim is to train clinical informaticians, the accreditation standards, the curricula, and the examination will differ, but the process of validating that exam and using the results of the content validation to assess the curriculum can still be done.

The data show statistically significant differences between instructed and uninstructed students, which speaks to the construct validity of the examination as well as the curriculum itself. Students on average knew between 55% and 60% of the material on the exam prior to entering, which is not surprising since we look for students with relevant work experience. After a year of coursework, their average score was approximately 10 points higher. It is gratifying for the instructors to know that the students improved on the exam after taking their core courses, but if the match was good between the content domain, the exam, and the curriculum, and the curriculum was well taught, a greater increase in performance would have been expected. Students’ scores clearly did not demonstrate mastery, but it is likely that if it were a high-stakes exam for which students actually prepared and studied, their scores would have been higher. In addition, since students are rarely given multiple-choice examinations in our classes, these questions may not have measured what students actually learned. On the other hand, it is possible that the exam may not match the curriculum as well as was intended. A review of the examination and the curriculum can help us assess where the problems lie. As we review the items on the examination for their match to the content of the domain, we plan to eliminate or modify items that clearly do not match, which can affect both pretest and post-test scores. We will also review the match between the domain content, the curriculum content, and the examination in terms of the relative emphases of different topics and may have to modify the curriculum and/or the examination to improve the match between them.

As an example, the CAHIIM curricular elements include a heavier emphasis on computer science than does our curriculum. If the examination also includes a heavy emphasis on this area, it is not surprising that students did not show as great an improvement as we expected after taking the core courses. This discrepancy can lead us either to

---

**Table 2: Descriptive data on pretests and post-test**

<table>
<thead>
<tr>
<th>Test</th>
<th>Cohort</th>
<th>Mean (SD)</th>
<th>Median</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pretest</td>
<td>Cohort 1 ($n = 11$)</td>
<td>55.00 (10.32)</td>
<td>55</td>
<td>32</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>Cohort 2 ($n = 15$)</td>
<td>57.87 (4.81)</td>
<td>59</td>
<td>48</td>
<td>66</td>
</tr>
<tr>
<td>Post-test</td>
<td>Cohort 1 ($n = 11$)</td>
<td>65.36 (8.11)</td>
<td>66</td>
<td>46</td>
<td>76</td>
</tr>
</tbody>
</table>

**Table 3: Independent relevance review of questions on the Cohort 2 examination by 2 health informatics faculty members**

<table>
<thead>
<tr>
<th>Reviewer 1</th>
<th>Not relevant or slightly relevant</th>
<th>Relevant or very relevant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not relevant or slightly relevant</td>
<td>6</td>
<td>25</td>
<td>31</td>
</tr>
<tr>
<td>Relevant or very relevant</td>
<td>4</td>
<td>112</td>
<td>116</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>137</td>
<td>147</td>
</tr>
</tbody>
</table>
modify the curriculum if we feel we are underemphasizing an important area or modify the exam so that it reflects a better match to the curricular foci.

As CAHIIM will be revising its criteria and moving to a competency-based set of accreditation standards, these questions can be addressed as the curriculum is designed and as the examination is developed. If careful attention is given to these issues, it is likely that the high content and construct validity can be maintained, but the exam itself may then be able to identify particular areas of the curriculum where students did not appear to learn what was expected, and remediation of the problem areas can be done. Using the contrasted groups method and assessing the performance of instructed vs uninstructed students would be an appropriate approach to validate an assessment instrument for other types of informatics programs as well as those accredited by CAHIIM.

We had assumed that the background variables that we take into account in admission would relate to an exam that assesses the content that the curriculum emphasizes. This relationship was not demonstrated in the present data. We also hypothesized that the pretest exam would be predictive of students’ grades at the end of the first year, and this relationship also was not significant. There can be several reasons for the lack of relationships among the variables, all of which need to be explored.

We know we have a wide range of ages of students, and it is likely that their undergraduate grade point averages do not have the same meaning, given the trends toward grade inflation that have been observed over the last few years. Health care experience may be with or without health care IT experience. More explicit measures of health care, IT, and health care IT experience would provide more confidence in these measures.

As we discussed above, the exam may not match as closely to the curriculum as we intended. It may have different emphases from the curriculum, for instance, or some areas that were intended to be taught may not have been taught or taught well. A review of the emphases in the courses and the exam can determine whether the exam appropriately reflects the curricular content. A more in-depth review by program faculty of the item analysis results for each of the individual questions can determine if particular content areas that were supposed to be included in the curriculum were, in fact, not covered. This type of review can be used to modify the curriculum.

There are statistical reasons that can also explain the low correlation that may have nothing to do with the curriculum/exam match. Since a C is considered unacceptable graduate level performance, most students get A’s and B’s, making the variability among graduate GPAs small. This lack of variability can depress the correlations. The small sample size of the cohorts and its effect on the statistical data is likely to affect many informatics graduate programs, since the class sizes in many informatics graduate programs tend to be small.

Regardless of sample size or statistical significance, the examination performance and the item analyses of its content can highlight areas for educational programs to review as they work to develop curricula to teach, and assessment tools to measure, a common set of competencies. Although we used a multiple-choice format, similar analyses could be done with assessments using different formats.

Going forward, once the review that we described in this manuscript is completed and the exam and/or the curriculum are modified, we may require students to demonstrate mastery on the exam to advance to the specialty tracks. As informatics accreditation criteria are applied more widely, it is likely that informatics masters’ programs will become much more standardized than they are now. When that occurs, it may be possible to do cross-program collaborative research on student performance. With larger sample sizes and validated test instruments, the results of studies like this one may provide lessons that can be broadly applicable.

LIMITATIONS

There are some limitations to our study. The number of students is small for a research study, which means that results must be looked at with caution. However, our purpose was to present a method for using the examination data to target areas for further review, and they can still be useful for that purpose even with the small sample size.

There were some differences in the questions on the exams given to the 2 cohorts, but it is unlikely that they had a strong influence on the results. First of all, very few questions changed from the pretest to the post-test for Cohort 1. Although somewhat more questions had wording to improve clarity for the Cohort 2 pretest, if their performance had improved significantly due to these questions, it would have maximized the differences between the 2 pretest scores and minimized the differences between the Cohort 1 post-test and the Cohort 2 pretest. On the contrary, the results showed that there were no significant differences on the pretest between the cohorts, and there were significant differences between the Cohort 2 pretest and the post-test of Cohort 1.

Bigger issues are that in this initial stage of examination development, although the questions reflected the CAHIIM curricular elements, we did not explicitly ensure that the number of questions for content areas was representative of the distribution of topics in either the CAHIIM requirements or the curriculum. However, overall, the major content areas in both the CAHIIM elements and the curriculum did have more questions than areas that were less emphasized.

Finally, our criteria for selection of items as well as the instructions for the content validation process focused on the match of topics to CAHIIM curricular elements. For the content validation, we did not explicitly tell our reviewers to also focus on the cognitive level of the question. If one reviewer was focusing on that and the other was not, that might be a reason for the discrepancies identified. In the future, as we examine the areas where there was lack of agreement as to the match to the CAHIIM content, we will also assess the cognitive level of the questions to assure that the distribution of items on the examination is not only representative of the appropriate content but also that it is at the right cognitive level.

The limitations we identified provide additional lessons in test development:

1. Pilot-test the questions and use item analysis data to refine and improve questions as part of the test development process prior to its official use.
2. Make sure the test items not only match the content of the curriculum and the domain, but also that they are a representative sample in terms of the major content areas and the cognitive level that is desired.

CONCLUSION

As the profession of informatics evolves, there will be more emphasis on accreditation of educational programs, which in turn will demand a rigorous approach to curriculum design, curriculum improvement, and evaluation. Good assessment instruments can be used to inform curriculum development, and systematic assessment of the instruments themselves can be used to improve the assessment instruments as well as the curriculum and student selection process.

We described methods of assessing the psychometric properties of a comprehensive exam designed to assess students’ core knowledge in informatics. There are statistical reasons that can also explain the low correlations that may have nothing to do with the curriculum/exam match. Since a C is considered unacceptable graduate level performance, most students get A’s and B’s, making the variability among graduate GPAs small. This lack of variability can depress the correlations. The small sample size of the cohorts and its effect on the statistical data is likely to affect many informatics graduate programs, since the class sizes in many informatics graduate programs tend to be small.

Regardless of sample size or statistical significance, the examination performance and the item analyses of its content can highlight areas for educational programs to review as they work to develop curricula to teach, and assessment tools to measure, a common set of competencies. Although we used a multiple-choice format, similar analyses could be done with assessments using different formats.

Going forward, once the review that we described in this manuscript is completed and the exam and/or the curriculum are modified, we may require students to demonstrate mastery on the exam to advance to the specialty tracks. As informatics accreditation criteria are applied more widely, it is likely that informatics masters’ programs will become much more standardized than they are now. When that occurs, it may be possible to do cross-program collaborative research...
knowledge in an informatics masters’ program. We found that the examination that was developed had generally good psychometric properties, and we suggested ways that the examination results could be used for curriculum improvement.

CONFLICT OF INTEREST
None of the authors have any conflicts of interest.

ACKNOWLEDGEMENTS
The authors appreciate Drs Shannon Houser and Burynamin Ozaydin for their assistance with rating the relevance of the examination questions, and Dr Dorthea Juul from the American Board of Psychiatry and Neurology for her helpful comments on an earlier draft of this manuscript.

REFERENCES

AUTHOR AFFILIATIONS
1Professor and Director, Center for Health Informatics for Patient Safety/Quality, Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, USA
2Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, USA