Discussant: Tim Wood, PhD

Gathering Evidence of External Validity for the Foundations of Medicine Examination: A Collaboration Between the National Board of Medical Examiners and the University of Minho

Marcia L. Winward, André F. De Champlain, Irina Grabovsky, Peter V. Scoles, David B. Swanson, Kathleen Z. Holtzman, Lorena Pannizzo, Nuno Sousa, and Manuel J. Costa

Abstract

Background
To gather evidence of external validity for the Foundations of Medicine (FOM) examination by assessing the relationship between its subscores and local grades for a sample of Portuguese medical students.

Method
Correlations were computed between six FOM subscores and nine Minho University grades for a sample of 90 medical students. A canonical correlation analysis was run between FOM and Minho measures.

Results
Moderate correlations were noted between FOM subscores and Minho grades, ranging from −0.02 to 0.53. One canonical correlation was statistically significant. The FOM variate accounted for 44% of variance in FOM subscores and 22% of variance in Minho end-of-year grades. The Minho canonical variate accounted for 34% of variance in Minho grades and 17% of the FOM subscore variances.

Conclusions
The FOM examination seems to supplement local assessments by targeting constructs not currently measured. Therefore, it may contribute to a more comprehensive assessment of basic and clinical sciences knowledge.

Acad Med. 2009;84(10 Suppl):S116–S119.

Globalization is now firmly interwoven into the fabric of higher education, including that of medical education. Internationalization in medical education is reflected in a host of areas including the use of distance learning technologies as well as the migration of practitioners and students across borders.[1–3]

The migration of both students and clinicians across borders is supported by the numerous training experiences available abroad.[3] Furthermore, in Europe, the desire to promote international mobility of practitioners and students is also reflected in a number of government-sponsored initiatives, including the European Union's Lifelong Learning Programme: 2007–2013 and the Bologna Process.[4,5]

The latter programs and processes underscore the need to develop common educational standards that can serve as quality-improvement tools and as an accreditation mechanism for medical schools around the world. Standards proposed by the World Federation of Medical Education are useful, because they allow medical schools to voluntarily measure themselves against a number of guidelines for self-improvement purposes and to prepare for external program reviews.[6] Similarly, the Global Minimum Essential Requirements have been successfully implemented in several pilot examinations to assess the extent to which both students and institutions meet minimal standards of competence set out by an international panel of faculty.[7]

Although these global standards serve as useful frameworks to support cross-border educational programs, comparatively little effort has been devoted to developing examinations and tools that can be used to assess related outcomes. Collaborative efforts aimed at developing such assessment tools have been reported in the literature.[8] However, the scope of these studies has been limited to local contexts with little intention of generalizing beyond the participating institutions. Clearly, more effort should be placed in developing measurement tools for evaluating performance both in terms of common standards and local requirements and needs.

Recently, the National Board of Medical Examiners (NBME) embarked on a collaborative effort with a consortium of medical schools from Italy, Portugal, and Belgium to develop a multiple-choice assessment tool, the Foundations of Medicine (FOM) examination. The FOM exam would be used by these institutions to gauge the proficiency level of their students in a number of basic and clinical science disciplines of common interest. The blueprint for the 2008 200-item FOM form was developed collaboratively. The FOM form was structured primarily by organ system and physician task with approximately 35% of test items focused on basic sciences and 65% on clinical sciences. On approval of the blueprint, NBME test development staff constructed a draft examination in English from items recently retired from the United States Medical Licensing Examination (USMLE). Concurrently, an Italian version of the examination was prepared.

Correspondence: Marcia L. Winward, MPhil, National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104; e-mail: (mwinward@nbme.org).

Examinees from Portugal and Belgium completed the English version of the exam, whereas Italian candidates completed the test in their native language.

The FOM examination holds a great deal of promise for use in a global context, but it must nonetheless be evaluated according to the same rigorous psychometric standards that are commonplace with other high-stakes assessments. Given that one of the central aims of the FOM program is to supplement local assessments, gathering evidence to support the *external* aspect of validity is important because it can inform users on the relationships that exist between its scores and school-based measures.[9]

The primary objective of this investigation was to gather evidence of external validity for the FOM examination by assessing the relationships between FOM subscores and local end-of-year measures for a sample of fourth- to sixth-year Portuguese medical students who participated in the 2008 pilot. The relationships between FOM and local Portuguese subscores were examined univariately, with simple correlation coefficients, as well as multivariately, via a canonical correlation analysis. This research is particularly important because it addresses a central aspect of validity, that is, how the FOM relates to local assessments, and contributes potentially unique information not captured by end-of-year grades. Additionally, this preliminary study is useful in assessing whether the FOM testing framework might generalize outside the U.S. medical setting. Finally, this investigation should be viewed as one of several studies currently underway that are aimed at providing various sources of validity evidence to support the use of the FOM abroad.

Method

Participants

A total of 128 students from the University of Minho's Integrated Masters in Medicine program completed the FOM examination on April 24, 2008. The masters program at Minho University is six years in length, divided into four phases that address Biological-Psychological-Sociological Aspects of Health, Diseases and Patients, Diseases at the Clinic, and Supervised Professional Practice. The majority of candidates who completed the FOM were in Years 4 to 6 (90/128 or 70.3%). Because the examination was primarily targeted to end-of-degree students, analyses were restricted to this fourth- to sixth-year cohort. Participants consented to having their scores used anonymously for research purposes. All data were deidentified for analyses and stored securely. Because the study results pertain to the overall population of examinees, not to any individual or student class year, the risk of harm attaching to any individual is negligible. The University of Minho approved the use of deidentified student data for this study.

Examination and end-of-year grades

The six-hour FOM examination was composed of 200 retired USMLE multiple-choice items that required the examinee's single best answer. The number of options for a given item ranged from 4 to 13. The FOM examination targeted a variety of content and skill areas, including (1) Physician Tasks, such as Normal Structure and Function, (2) Normal Conditions and Disease categories, such as Cardiovascular Diseases, and (3) Disciplines, such as Medicine. For the purposes of the present analyses, the following six discipline subscores were retained: Medicine, Obstetrics–Gynecology, Pediatrics, Psychiatry, Surgery, and Clinical Pharmacology.

Given that our study was restricted to students in the fourth to sixth years of their degrees, candidates were tracked with respect to the following nine nonelective end-of-year Minho grades: Functional and Organic Systems II and III (Year 2), Biopathology and Introduction to Therapeutics, Introduction to Community Health, Introduction to Clinical Medicine, and Maternal and Child Health Clerkship (Year 3), and Medicine I Clerkship, Mental Health Clerkship, and Health Centre Clerkship I (Year 4). To ensure complete data for our entire cohort, electives and Year 5 and Year 6 grades were not retained for analytic purposes. End-of-year grades reported to candidates were on a scale of 1 to 20 and derived from a host of assessments including multiple-choice examinations, clinical vignettes, clinical skills assessments with real patients, and ratings of professionalism and clinical competence by faculty observers.

Analyses

First, Pearson product-moment correlations were computed between the 15 measures (six FOM discipline subscores and nine Minho grades) to assess relationships at the univariate level. Then, a canonical correlation was run between the FOM subscores and Minho grades. The goal of canonical correlation is to assess the relationships between two sets of variables. Specifically, canonical correlation attempts to address the following question: Along how many dimensions are the variables in one set related to the variables in the other? To illustrate, imagine a scenario in which the researcher is interested in looking at the relationship between three job characteristics and three measures of employee satisfaction. Although it is possible to compute simple correlations between the six variables, canonical correlation goes one step further by generating pairs of linear combinations of these variables, that is, canonical variates. The first pair of canonical variates is produced to maximize the correlation between a linear combination of one set (e.g., job characteristic measures) and a linear combination of the other (e.g., employee satisfaction measures). The process continues to extract orthogonal pairs of variates until no significant linkages remain. Our study, in addition to looking at univariate relationships (i.e., simple correlations between the 15 variables), explored the correlation between each set of measures (FOM subscores and Minho end-of-year grades) taken as two distinct aggregates.

Results

Univariate analyses

Pearson product-moment correlations were computed between the FOM subscores and Minho grades. Correlations between FOM subscores ranged from .25 (between Pediatrics and Psychiatry) to .72 (between Medicine and Surgery), with a mean of .49. Correlations between Minho end-of-year grades varied from .31 (between Introduction to Clinical Medicine and Introduction to Community Health) to .85 (between Functional Organic Systems

II and Functional Organic Systems III), with a mean of .51. Finally, correlations between FOM subscores and Minho grades ranged from −.02 (between Pediatrics scores and Mental Health Clerkship grades) to .53 (between Obstetrics–Gynecology and Biopathology and Introduction to Therapeutics), with a mean of .25. Note that reliability of FOM scores for the 2008 form was .90.

Canonical correlation

Results from the canonical correlation are provided in Table 1. Findings show that only the first canonical correlation was statistically significant, $F(54, 336) = 1.78$, $P = .0012$. In other words, the composite of FOM subscores and Minho end-of-year grades differed along one dimension only. The actual canonical correlation value was .70, indicating that there was nearly 50% (i.e., $.70^2$) overlapping variance between the first pair of canonical variates. In this study, the FOM canonical variate accounted for 44% of the variance contained in FOM subscores. It also explained 22% of the variance in Minho end-of-year grades. This is referred to as *redundancy* in canonical correlation parlance. Redundancy corresponds to the proportion of variance that one canonical variate extracts from the variables contained in the other canonical variate. The Minho canonical variate accounted for 34% of the variance contained in Minho grades and explained 17% of the FOM subscore variances.

To assess the relative contribution of each subscore to its canonical variate, a rough cutoff value of .30 was used. The latter values are analogous to regression coefficients. Using this criterion, Medicine (.68) and Obstetrics–Gynecology (.34) were most highly associated with the FOM canonical variate. Although slightly below our cutoff, Psychiatry also seemed to be inversely related to the FOM variate (−.29). Findings for the Minho canonical variate were similar. Standardized coefficients for Medicine I Clerkship (.62), Health Centre Clerkship I (.29), and Functional and Organic Systems II (.54) seemed to be associated with the Minho variate. Similarly, the standardized canonical coefficient computed for the Mental Health Clerkship score (−.50) was inversely related to its canonical variate.

Our findings suggest that examinees with higher FOM subscores in Medicine and Obstetrics–Gynecology tend to have higher end-of-year grades in Medicine I Clerkship, Health Center Clerkship I, and Functional and Organic Systems II. FOM Psychiatry scores and Minho Mental Health Clerkship scores tend to be related to one another, but they are inversely associated with each canonical variate.

Table 1
Canonical Correlation Results

Subscores	Standardized canonical coefficient*
FOM discipline subscores	
Medicine	.6810
Obstetrics–Gynecology	.3377
Pediatrics	.1560
Psychiatry	−.2873
Surgery	−.0231
Clinical Pharmacology	.1742
Percent of variance[†]	.4386
Redundancy[‡]	.2172
Minho subscores	
Biopathology and Introduction to Therapeutics	−.1447
Introduction to Community Health	.1520
Introduction to Clinical Medicine	−.2032
Medicine I Clerkship	.6245
Mental Health Clerkship	−.4997
Health Centre Clerkship I	.2945
Maternal and Child Health Clerkship	.0319
Functional and Organic Systems II	.5388
Functional and Organic Systems III	.0364
Percent of variance[†]	.3408
Redundancy[‡]	.1687
Canonical correlation	.7037

* The standardized canonical coefficient can be thought of as a standardized regression coefficient. Higher positive values suggest a stronger association between a given variable and the linear composite, whereas negative values indicate an inverse relationship.

[†] This value indicates the proportion of variance in individual variable scores accounted for by the linear composite or canonical variate.

[‡] Redundancy corresponds to the proportion of variance in individual variable scores accounted for by the other canonical variate.

Discussion

Results from our study suggest that moderate relationships exist between subscores from the FOM examination and Minho end-of-year grades. This finding was noted not only in the univariate analyses undertaken (the simple correlations) but also in the canonical correlation analysis.

As such, it seems that the FOM examination in part targets constructs that are not currently measured by the various assessments that contribute to Minho end-of-year grades. This implies that the combination of FOM and Minho grades provides a more comprehensive student assessment than either stand-alone measure. The goal of the FOM was to supplement local assessments (not replace them) by providing measures of content domains deemed important by participating medical schools but not necessarily targeted by their own exams. Therefore, our findings provide evidence to support that the FOM is contributing information not present in Minho assessments.

With respect to the contributions of individual measures, it is not surprising that Medicine subscores were highly related across and within canonical variates. The FOM Medicine subscore contained the largest amount of items (91) and thus contributed a greater amount of information to its canonical variate. Similarly, the Minho Medicine I Clerkship grade was computed from a number of assessments including knowledge-based tests, clinical ratings, patient interactions, and clinical vignettes. One finding that may seem to differ from this trend pertains to the strong associations noted between the Functional and Organic Systems II grade with its own canonical variate and the FOM linear combination of variables. This result is likely ascribed to the similarities of the contents addressed in these courses; both Functional and Organic Systems II and Medicine I Clerkship cover the topics of Cardiovascular and Respiratory Systems.

It is interesting that both the FOM Psychiatry subscore and the Mental Health Clerkship end-of-year grade were positively correlated with other variables both within and between sets but negatively related to each canonical variate. What might seem like a counterintuitive outcome is actually a classic illustration of a negative suppressor variable. Both measures are related to other FOM subscores, as evidenced by positive correlations between FOM Psychiatry scores, Minho Mental Health grades, and additional variables. However, these two measures are inversely related to both linear composites of FOM and Minho variables. By including them in the model, we actually improve its fit by partialing out the variance shared between the suppressor variables and other measures; that is, the latter variance resides in the suppressor variable only.

It is important to stress that the results of this study must be interpreted with a few caveats. First, the sample size was small, which impacts the power of the analysis and the extent to which findings can be generalized to other cohorts. Second, the analysis, because of the small sample size, was restricted to a portion of all subscores computed. Nonetheless, there was consensus that the latter variables were the most germane because they were consistent for all examinees. Also, it is possible that the modest relationships noted between FOM subscores and Minho end-of-year grades reflect some decay in knowledge, especially for those content areas applied less frequently. Finally, canonical correlation, although useful, must be viewed as a descriptive analysis which may lead to solutions that are difficult to interpret. As such, this study needs to be replicated with other participating cohorts before making more definitive conclusions about the merits of the FOM examination program.

Despite these limitations, our findings suggest that the FOM examination provides useful additional measures of constructs not targeted by end-of-year grades. Future plans call for replicating these analyses not only with further cohorts of Portuguese students but also with other participants, once more local data become available.

References

1 Harden RM, Hart IR. An international virtual medical school (IVIMEDS): The future for medical education? Med Teach. 2002;24: 261–267.

2 Vermund SH, Sahasrabuddhe VV, Khedkar S, Jia Y, Etherington C, Vergara A. Building global health through a center-without-walls: The Vanderbilt Institute for Global Health. Acad Med. 2008;83:154–164.

3 McKinley DW, Williams SR, Norcini JJ, Anderson MB. International exchange programs and US medical schools. Acad Med. 2008;83:s53–s57.

4 European Union Lifelong Learning Programme: 2007–2013. Available at: (http://ec.europa.eu/education/programmes/newprog/index_en.html). Accessed June 15, 2009.

5 Council of Europe. Bologna for pedestrians. Available at: (http://www.coe.int/t/dg4/highereducation/EHEA2010/BolognaPedestrians_en.asp). Accessed June 15, 2009.

6 Karle H. Global standards and accreditation in medical education: A view from the WFME. Acad Med. 2006;81:s43–s48.

7 Stern DT, Friedman Ben-David M, De Champlain A, Hodges B, Wojtczak A, Schwarz MR. Ensuring global standards for medical graduates: A pilot study of international standard setting. Med Teach. 2005;27:207–213.

8 De Champlain AF, Melnick D, Scoles P, et al. Assessing medical students' clinical sciences knowledge in France: A collaboration between the NBME and a consortium of French medical schools. Acad Med. 2003;78:509–517.

9 Messick S. Standards of validity and the validity of standards in performance assessment. Educ Meas Issues Pract 1995;14:5–8.