Evaluating the Systematic Validity of a Medical Subspecialty Examination

Mikaela M. Raddatz, Kenneth D. Royal
American Board of Family Medicine

Jessica Pennington
University of Kentucky

Date: May 2012
Conference: Midwestern Educational Research Association
November 8, 2012
Evanston, IL

Abstract

The purpose of this study is to determine if the construct of a medical subspecialty examination, as defined by the hierarchy of item difficulties, is stable across physicians who completed a fellowship and recertifiers as compared to non-fellows. Three comparisons of groups are made: 1) Practice pathway board candidates compared to members of all other boards taking the subspecialty examination, 2) Practice pathway board candidates who did not complete a fellowship compared to members of all other boards, and 3) Practice pathway board candidates who completed a fellowship compared to new candidates who had not completed a fellowship. All group comparisons showed significant positive correlations. As expected, the study did not find evidence of DIF between subgroups. However, non-fellowship examinees do score systematically lower than their fellowship taking counterparts. This suggests the value of a fellowship program. The study demonstrates the stability of the construct, therefore the reason behind the difference in passing rate lies elsewhere and should be examined.

Evaluating the Systematic Validity of a Medical Subspecialty Examination

Physicians often seek certification in order to provide evidence from an independent and unbiased organization to the public that they possess sufficient education, medical knowledge, clinical decision-making skills and professional standing. Certification provides information to the public that physicians are qualified to give quality care to individuals, families, and the community. Certification in a medical specialty may also be a condition of employment by employers. Typically, physicians complete medical school, select a medical specialty (e.g. Family Medicine, Pediatrics, Emergency Medicine, Surgery, etc.) in which they wish to practice and then obtain additional training in that specialty via a residency program. After completing their residency, earning licensure, passing the certification examination, and meeting the other prerequisites established by the corresponding certification board, the physician becomes board certified in that specialty. As long as the physician remains board certified, the certification board usually refers to the physician as a diplomate of their board.

Some diplomates also choose to become certified in a subspecialty. Certification in a subspecialty allows diplomates to demonstrate an independent evaluation of their quality in a more specific medical area. For example, physicians can attain subspecialty certification in areas such as Adolescent Medicine, Geriatric Medicine, Hospice and Palliative Medicine, Sleep Medicine, and Sports Medicine. Among the certification boards that are members of the American Board of Medical Specialties (ABMS), there is often cooperation both with regard to developing the subspecialty examinations and with regard to establishing the prerequisites to become certified. As a result, there is substantial comparability of subspecialty certification even across primary specialties. Although the subspecialty certification is granted by the different respective specialty boards, the development of the examination is guided by a consortium of the certification boards that offer the subspecialty certificate. Across the certification boards, there are three prerequisites to earning the subspecialty certification. First, the physician must be certified by their respective specialty board and must be a diplomate in good standing. Second, the physician must have satisfactorily completed a one year ACGME-accredited fellowship in the subspecialty in which they are seeking certification. Finally, the diplomate must achieve a passing score on the subspecialty examination. Only after meeting these requirements is the subspecialty certificate awarded to the candidate.

When an ABMS medical specialty board decides to join an existing consortium of ABMS medical boards that sponsor a subspecialty examination, there are relatively few, if any, ACGME accredited fellowship programs available to them in that subspecialty. Traditionally, fellowship programs are affiliated with an ACGME accredited residency program for a particular medical specialty and there is a strong preference to accept into the fellowship those diplomates who are board certified in the same specialty as the residency program. When a new certification board joins an existing consortium, it is difficult for the diplomates to find slots in the limited fellowship programs available to them. The fellowship requirement also disadvantages existing diplomates from the new board because they will not be at a stage in their career where they can take a year off to complete a fellowship. Some of these diplomates may have been practicing in this subspecialty for many years without certification. To remedy this, the consortium typically authorizes the new consortium member to permit a “practice pathway” as an alternative to the fellowship eligibility requirement for a period of approximately five

years. Normally, the practice pathway makes eligible those diplomates who have completed five years of post-residency practice in that subspecialty or who have completed a non-accredited fellowship program that is affiliated with an ACGME-accredited residency training program; however the fellowship program must be consistent with the ACGME training requirements for that subspecialty. Although there is some variability in the fellowship program curricula among the different boards, much of it is standardized by the ACGME requirements.

Aside from candidates who participate in the practice pathway, there is another group of examinees who are eligible to take the exam without having completed a fellowship. These candidates are members of any board who are taking the examination as recertification candidates. The exam must be re-taken every 10 years for certification to remain valid. Thus, candidates who are taking the exam in order to recertify and participated in the "practice pathway" prior to taking the initial exam would not have completed a fellowship. These candidates also represent a different group than their non-fellowship taking counterparts as they have more extensive experience in the field. For the purpose of this study, recertification candidates will be included in the "fellowship" category because they are more experienced than non-fellowship candidates.

On one particular subspecialty examination, physicians who have not completed a fellowship have had about a 60% pass rate compared to their counterparts who have completed a fellowship with an 80% pass rate. Based on this difference in pass rates, it is of interest to investigate whether this subspecialty examination exhibits systematic validity. In other words, is the construct of ability roughly the same for fellowship trained physicians as compared to non-fellowship physicians? This study aims to examine the construct as defined by the relative difficulty of the items on the examination.

In order for a test to demonstrate construct validity, it is important that the hierarchy of item difficulties shows the same pattern across different subsets of examines. Construct validity refers to the extent to which a theorized construct measures what it intends to measure. An aspect of construct validity is systematic validity. Systematic validity is the base of measurement set on a working definition of what the construct encompasses. The construct of the knowledge of the subspecialty can be defined as the difficulty of the items relative to each other. Therefore, the item hierarchy should remain stable across subgroups if the exam is measuring the same construct.

Evaluation of systematic validity in this study will be done by using a Differential Item Functioning (DIF) analysis. DIF occurs when there is a significant difference in the probability of a correct answer across subpopulations after controlling for differences in ability. DIF identifies items that are significantly easier or harder for a particular subgroup compared to another subgroup within the same sample. In this study, the subspecialty scores for candidates who have not completed a fellowship will be compared to candidates who have completed a fellowship (as well as recertifiers). This will determine if candidates who have not taken the fellowship are systematically scoring lower than both their fellowship and recertification counterparts. If the items are functioning differently for the candidates who do not complete the fellowship, the subpopulation could potentially be altering the calibration of the test items and thus should not be included in the calibration.

The purpose of this study is to determine if the hierarchy of item difficulties is stable across fellows and recertifiers as compared to non-fellows on the subspecialty examination. Three comparisons of groups will be made:

1. Practice pathway board candidates seeking the subspecialty certification compared to all other candidates taking the examination.
2. Practice pathway board candidates who have not completed a fellowship seeking the subspecialty certification compared to all other candidates taking the examination.
3. Practice pathway board candidates who have not completed a fellowship seeking the subspecialty certification compared to practice pathway board candidates who did complete a fellowship.

Method

Measure. This study used archival data from one of the subspecialty examinations. This examination consists of 200 items given in a non-adaptive computer format. Items are administered in a random order within forms. The exam has two forms and is administered twice per year, winter and summer. The items are in multiple choice format with each item having four to five distracters. Examinees have four hours to complete the exam and reported scaled scores on the exam range from 200 to 800. The scaled scores are derived from the logit metric that is produced from the scoring process which uses a dichotomous Rasch model (1960). There were two forms of the test administered consisting of 200 items, 49 of these items were common items across test forms leaving 151 items unique to that test form. This means that there was a total of 351 items being evaluated across both forms.

Participants. Participants were 391 physicians who took the subspecialty examination in the summer of 2009. Participants were separated into two comparison subpopulations: 1) physicians taking the exam for initial certification who have completed a fellowship program as well as physicians who have been board certified for at least nine years and are taking the exam for recertification (a minority of these physicians may not have completed a fellowship program) and 2) physicians who are taking the exam for initial certification and have not completed a fellowship program.

Of the 381 physicians who took the subspecialty exam, 89 were practice pathway board candidates. Of the 89 practice pathway board candidates, 66 had not completed a fellowship program and 23 had completed a fellowship program prior to taking the exam. Demographic data such as age, gender and number of years licensed was collected but for the purpose of this analysis only board association and completion of a fellowship was used.

Table 1
Study Participants by Board Membership

Board	Fellowship Candidates and/or Recertifiers	Non-Fellowship Candidates Attempting Certification	Total
Board 1	231		
Board 2	25		
Board 3	24		
Board 4	12		292
Board 5	23	66	89
Total	315	66	381

Results

Strong positive correlations were expected for item calibrations between all compared subpopulations. Group comparisons included: 1) Practice pathway board candidates compared to members of all other boards taking the subspecialty examination, 2) Practice pathway board candidates who had not completed a fellowship compared to members of all other boards, and 3) Practice pathway board candidates who had completed a fellowship compared to new candidates who had not completed a fellowship. All group comparisons showed significant positive correlations (as seen in Table 2). All correlations fell within the category of a large effect size of meeting or exceeding r = .50, as defined by Cohen (1992).

Table 2
Subpopulation Correlations

Comparisons	Correlation	*p*-value
Practice Pathway Board Candidates vs. All Boards	.722	$p < .001$
Practice Pathway Board Candidate Non-Fellows vs. All Boards	.708	$p < .001$
Practice Pathway Board Candidate Non-Fellows vs. Practice Pathway Board Candidate Fellows	.798	$p < .001$

A Rasch based DIF analysis was utilized to evaluate each group comparison. For each comparison, the items were calibrated separately. For example, for the first comparison the items were calibrated for practice pathway board candidates only and then calibrated separately for candidates from all other boards. The item calibrations were then compared using *t*-tests in order to determine if there was a significant difference between the item difficulties. If a significant difference was shown, the item was identified as showing evidence of DIF. Thus, an item that showed evidence of DIF could potentially be functioning differently between subpopulations.

The first group comparison was practice pathway board candidates versus members of all other boards taking the subspecialty examination. Between these subpopulations, zero items showed evidence of DIF. These results are illustrated in Figure 1. If an item functioned differently between subpopulations, it would lie outside of the confidence intervals on the graph. For the second comparison, practice pathway board candidates who had not completed a fellowship were compared to members of all other boards. Once again, none of the items showed evidence of DIF. These results are shown in Figure 2. Finally, practice pathway board candidates who had completed a fellowship were compared to practice pathway board candidates who had

not completed a fellowship. As with the first two comparisons, zero items showed evidence of DIF. These results are shown in Figure 3.

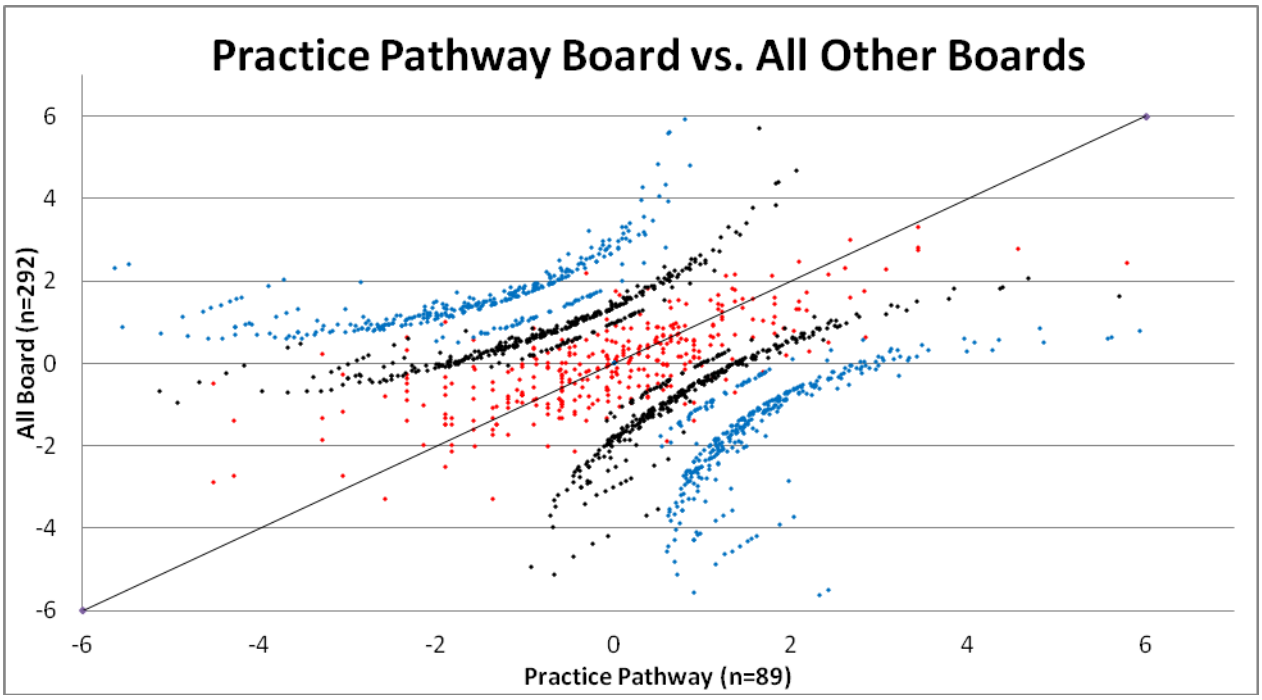

Figure 1. Scatterplot of calibrations for practice pathway board candidates and all other board members with 95% confidence intervals and confidence intervals after a Bonferroni adjustment.

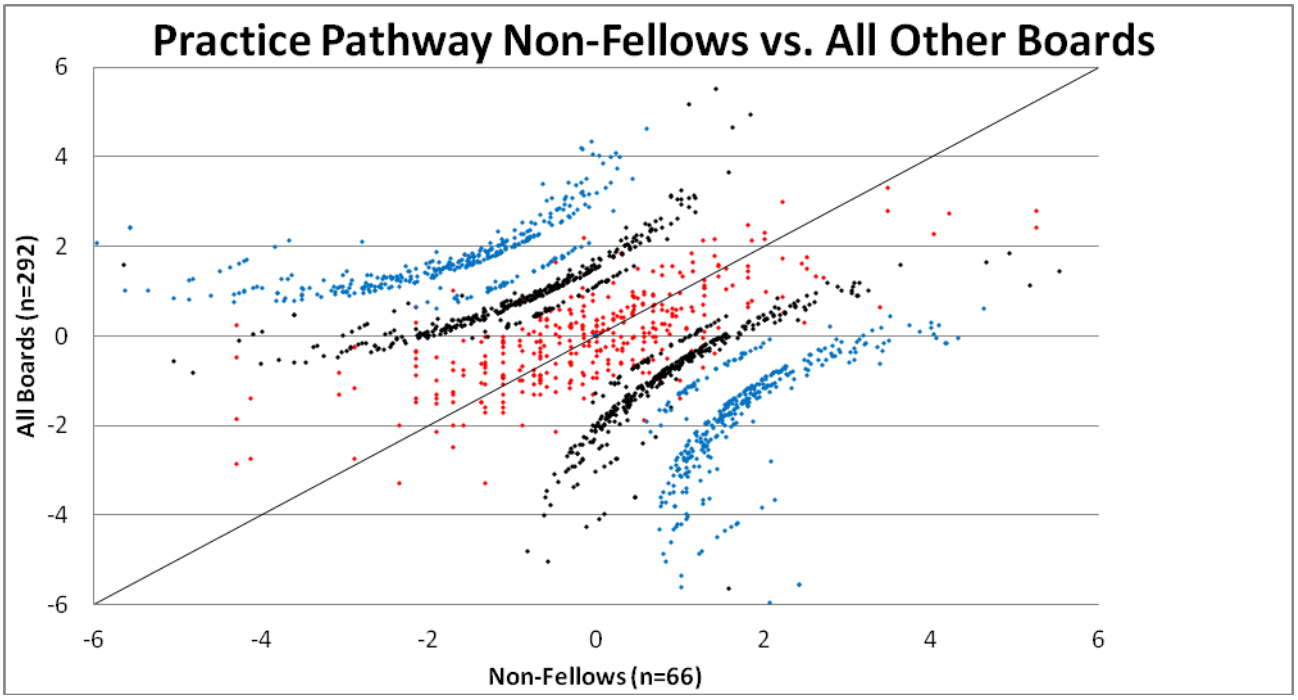

Figure 2. Scatterplot of calibrations for practice pathway board candidates who had not completed a fellowship and all other board members with 95% confidence intervals and confidence intervals after a Bonferroni adjustment.

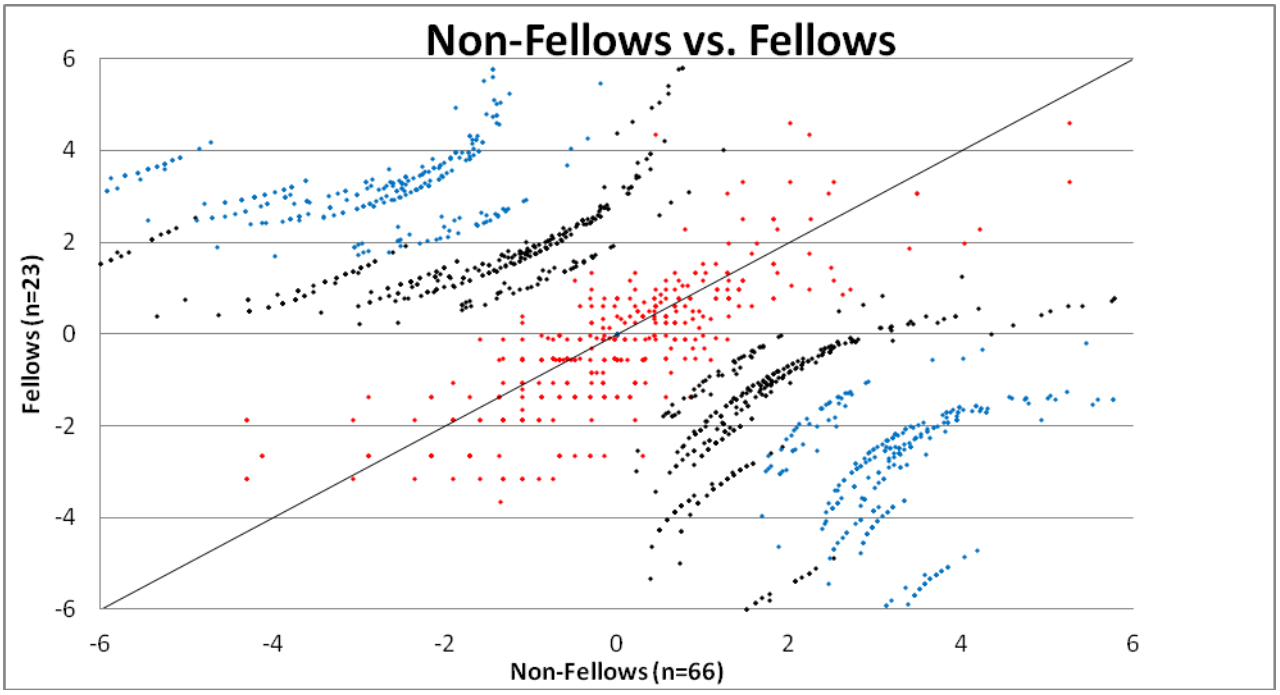

Figure 3. Scatterplot of calibrations for practice pathway board candidates who had not completed a fellowship and practice pathway board candidates who had completed a fellowship with 95% confidence intervals and confidence intervals after a Bonferroni adjustment.

This study used a multiple comparison procedure in the analysis in order to control for the inflation of alpha error. A Bonferroni adjustment was used in the analysis and is considered conservative under most conditions. Bonferroni holds experimentwise alpha constant at or below what it should be for multiple comparisons. This multiple comparison procedure is important to use for the purpose of this study because the study is seeking evidence to support the stability of the construct as a whole. If specific items were evaluated as a means for quality control, the use of multiple comparison procedures would not be appropriate. The Bonferroni correction is the most flexible multiple comparison procedure and ensures that the probability of showing significance by chance will not be greater than 5%.

Discussion

The purpose of this study was to evaluate the systematic validity of a subspecialty examination by comparing subpopulations within the candidates who completed the exam. The groups examined included practice pathway board candidates attempting to earn certification compared to members of all other boards taking the subspecialty examination, practice pathway board candidates who had not completed a fellowship compared to members of all other boards, and practice pathway board candidates who had not completed a fellowship compared to practice pathway board candidates who had completed a fellowship. A DIF analysis was used to identify items that were functioning differently for different groups of test takers. Correlations between group calibrations were significant and strong in a positive direction. This strong positive correlation shows that the test is generally operating in the same way for all groups of test takers, showing the test as a whole demonstrates systematic validity.

The construct in question, the knowledge of the subspecialty, should be equivalent for all groups given that all candidates who pass the examination receive the same certification. This construct is defined by the hierarchy of item difficulties. Results from the three comparison groups showed that no items showed evidence of DIF. This suggests that the construct of the subspecialty exam is stable across all subpopulations and that the items are functioning the same for all examinees. This does not suggest that the ability of all test takers is the same, but that the test itself is presenting an item hierarchy that looks the same for all test takers.

If items had shown evidence of DIF in this analysis, these items would be operating differently for different groups of test takers. Such items would need to be examined further to determine why they are functioning differently. This would not suggest the item is necessarily biased or needs to be removed from the exam, but that the item needs to be explored substantively for why it showed evidence of DIF. However, these items could show evidence of bias or they could perhaps show evidence of real differences between subpopulations. Such information could aid in training content for residency directors, or preparation procedures for examinees.

As expected, the study did not find evidence of DIF between subgroups. However, non-fellowship examinees do score systematically lower than their fellowship taking counterparts. This suggests the value of a fellowship program. The study demonstrates the stability of the construct, therefore the reason behind the difference in passing rate lies elsewhere. When examining a trend such as this, it is important to determine whether there is something wrong

with the examination or if there is something else going on that affects test scores. In this case, it is evident that completing a fellowship increases the probability that an examinee will pass the subspecialty exam.