False-Negative and False-Positive Rates Resulting from Measurement Error for a Mastery Examination of clinical Competence Based on Standardized-Patient Cases

COLLIVER, JA, et al.

COLLIVER, JA, et al. False-Negative and False-Positive Rates Resulting from Measurement Error for a Mastery Examination of clinical Competence Based on Standardized-Patient Cases. Teaching and Learning in Medicine, 1993, vol. 4, no. 4, p. 238-242

Available at: http://archive-ouverte.unige.ch/unige:26622

Disclaimer: layout of this document may differ from the published version.
False Negative and False Positive Rates Resulting from Measurement Error for a Mastery Examination of Clinical Competence Based on Standardized-Patient Cases

Jerry A. Colliver, Nu Viet Vu, Randall S. Robbs Steven J. Verhulst, Terry A. Travis, Howard S. Barrows

Southern Illinois University School of Medicine
ABSTRACT

Standardized-patient (SP) examinations are being used increasingly for the identification of examinees who have mastered the clinical competencies deemed necessary for the practice of clinical medicine. Hence, it is of critical importance to evaluate the psychometric properties of the mastery decisions, in particular, the false negative and false positive rates associated with these pass-fail decisions. In this paper, methods are described for estimating the false negative and false positive rates resulting from measurement error for a mastery examination. These rates as well as the reliabilities of the pass-fail decisions based on a recently adopted standard-setting procedure are reported for data for six classes of senior medical students. The methods for determining the false negative rate, in particular, provided valuable information for our faculty about the effectiveness of the new standard setting procedure for minimizing false negative decisions for our SP-based examination.
The use of standardized patients (SPs) in performance-based examinations of clinical competence has increased rapidly in recent years. SPs are currently being used to assess medical students and residents, and plans are underway for their use in national licensure and certification tests. A primary purpose of these examinations is the identification of examinees who have mastered the clinical competencies deemed necessary for the practice of clinical medicine. Hence, it is of critical importance to evaluate the psychometric properties of these mastery decisions, that is, of the pass-fail decisions. In particular, it is of importance to determine the reliability of the pass-fail decisions and to determine the false negative and false positive rates resulting from the unreliability of these decisions.

At Southern Illinois University (SIU) School of Medicine, an SP-based Post-Clerkship Examination of clinical competence is given to senior medical students upon completion of their clinical clerkship rotations. Students are expected to pass the examination to fulfill a part of their graduation requirements. Hence, faculty have been especially concerned about the psychometric properties of the pass-fail decisions, in particular, with the false negative rate (i.e., the probability of incorrectly failing an examinee who in fact has truly mastered the clinical competencies assessed by the examination). Because the examination is new and because students are expected to pass the examination as a part of their graduation requirements, faculty thought it was better to err on the side of the student at this stage of the development of the examination. Although there is a large body of research that supports the psychometric properties of the examination, faculty thought that there is still a lot to learn about the examination (e.g., scoring, standard setting, etc.). Hence, faculty felt that false negative decisions would constitute more serious errors than false positive
decisions (i.e., passing students who really should fail) and therefore were primarily concerned with minimizing the false negative rate at this stage of the development of the examination.

Starting with the class of 1991, the basis for the examination pass-fail decision was changed from number of cases passed to total examination score (i.e., the average of all of a student's case scores). Prior to the class of 1991, students were required to pass 60 percent of the cases in order to pass the examination as a whole. Under the new grading policy, to pass the full examination, a student's total examination score must exceed the Examination Pass Level, which is derived from the mean of the Case Pass Levels adjusted downward slightly to control for measurement error and minimize false negative decisions (i.e., failing students who really should pass). The Case Pass Level for each case is set by the case author and agreed to by faculty on the Post-Clerkship Examination committee and reflects the standards of performance expected of senior medical students upon completion of their clerkship rotations. Thus, the Examination Pass Level is an average mastery cut-off adjusted for measurement error. This new grading method provided us for the first time with a reasonable, objective basis for making mastery decisions concerning clinical competence.

The present study was conducted for the purpose of evaluating the psychometric properties of the mastery decisions on the SIU SP-based examination of clinical competence. In order to broaden the data base for this study, test results from the classes of 1986 through 1990 were rescored in accordance with the new grading policy instituted with the class of 1991 and analyses were performed on these data as well. The analyses for these six classes were performed to determine the reliabilities of the pass-fail decisions and to determine the false negative and false positive rates.
associated with the decisions. In particular, the purpose of the analyses was to provide faculty with information about the false negative rate associated with the new standard setting procedure, a matter of central concern to the faculty. In addition, the paper describes an approach for determining false negative and false positive rates resulting from measurement error that can be adapted to any performance-based examination that is used to make mastery decisions.

It should be emphasized at the outset that the false negative and false positive rates reported here are different from the usual false negative and false positive rates reported with screening and diagnostic tests. The usual false negative and false positive rates are due to a lack of agreement between an examination and a "gold standard" criterion. Unfortunately, there is no generally accepted "gold standard" against which an SP-based examination can be compared. Consequently, the false negative and false positive rates reported here are those that are due to unreliability (i.e., measurement error) in the examination. This unreliability can result in failing scores for true masters (i.e., examinees with true scores above the mastery level, but observed scores below the pass-fail cut-off) and passing scores for true nonmasters (i.e., examinees with true scores below the mastery level, but observed scores above the pass-fail cut-off). In brief, the methods described here are for a mastery examination, for determining the false negative and false positive rates resulting from measurement error.

Methods

The Examination. A thorough discussion of the SIU Post-Clerkship Examination, including details of the development, administration and
scoring, is presented elsewhere and should be consulted for a full description of the examination.\textsuperscript{5,6} In brief, the examination is a performance-based examination that uses about 18 forty minute SP cases (20 minutes for the student-SP encounter and another 20 minutes immediately following the encounter for students to answer written questions about the case). The examination is objectively scored and is administered to all senior medical students (about 70 per class) upon completion of their clinical clerkship rotations. Cases for the examination are chosen by the faculty Post-Clerkship Examination Committee and represent the most frequently encountered patient problems as well as the most important patient problems that students are expected to evaluate and manage competently. Each case selected is developed by a faculty physician, who determines the competencies to be assessed on the case and, with the faculty committee, develops instruments for collecting the necessary data about student performance. These data consist of checklists completed by SPs who record actions performed by students on history and physical, and written responses by students following the patient encounters to questions concerning findings, tentative diagnostic conclusions, and plans for treatment and management. For each case, a passing level is established by the case author and agreed to by the faculty committee. This Case Pass Level reflects the standards of performance expected of senior medical students upon completion of their clerkship rotations as expressed by minimal scores on the clinical competencies being assessed by the case.

Pass-Fail Decisions. In order to pass the Examination, a student's total score, which is the average of all of the student's case scores rounded to the nearest integer, must exceed the Examination Pass Level. The Examination Pass Level is the adjusted mean of the Case Pass Levels, where
the adjustment involves subtracting 1.29 times the standard error of measurement (SEM) from the mean Case Pass Level. This adjustment was employed to minimize the false negative rate (i.e., failing a student who really should pass). The adjustment was designed so that a student with true score exactly at the mean Case Pass Level would have only a 10% chance of scoring below the adjusted mean Case Pass Level, simply due to measurement error. Thus, the probability of a false negative decision is .10 for a student with true score exactly at the mean Case Pass Level and decreases rapidly for students with true scores farther and farther above the mean Case Pass Level.

Reliability of Pass-Fail Decisions. For each class, a dependability index with cut-off was used to assess the reliability of the pass-fail decisions. The dependability index with cut-off measures the consistency with which examinees' scores fall above or below the cut-off and thus is the recommended measure of reproducibility for a mastery examination. A dependability index with cut-off was also computed for all classes combined, using variance components pooled across classes as well as means of examination scores and means of Examination Pass Levels, both computed across classes.

False Negative Rate. The false negative rate was defined as the probability of incorrectly failing an examinee who in fact has truly mastered the competencies assessed by the examination. For purposes of these analyses, a student was said to have mastered the competencies if his or her true score exceeded the average mastery cut-off established by the faculty case authors, which was defined here as the mean of the Case Pass Levels. A student with true score above the mean of the Case Pass Levels, then, would incorrectly fail the examination if, due to measurement error, his or her
observed score fell below the Examination Pass Level. Two methods were used to estimate students' true scores, the direct method and the regression method. (See Lord and Novick, 9 pages 152-153.) With the direct method, the observed score for each student was used directly as an estimate of the student's true score (which is the expected value of the observed score). With the regression method, the true score for each student was estimated from the regression of all students' true scores on their observed scores. 9

The false negative rate was obtained by determining for each true score above the average mastery cut-off (1) the probability that an examinee with that true score would obtain an observed score below the Examination Pass Level, assuming a normal distribution of measurement error and (2) the proportion of examinees with that given true score, based on the empirical frequency distribution of students' scores. The products of (1) and (2) were summed across these true scores and this sum was divided by the proportion of students with these true scores, to obtain the false negative rate.

False Positive Rate. The false positive rate was defined as the probability of incorrectly passing an examinee who in fact has not truly mastered the competencies assessed by the examination. The computation of the false positive rate was conceptually similar to the computation of the false negative rate, except that the focus here was on true scores below the average mastery cut-off with observed scores above the Examination Pass Level.

Results

Pass-Fail Decisions. The failure rates for the six classes ranged from 2 of 70 (3%) for the class of 1986 to 7 of 65 (11%) for the class of 1990. (See Table 1.) For all six classes combined, the failure rate was 26 of 404 (6%).
Reliability of Pass-Fail Decisions. The dependability indices with cut-off for the six classes ranged from .55 to .93, with five of the six above .70. (See Table 1.) For all classes combined, the dependability index for an examination consisting of 18 SP cases was .84.

False Negative Rate. Using true scores estimated directly from observed scores, the false negative rates for the six classes ranged from .005 to .02, with a mean of .01. (See Table 1.) Similar results were obtained using estimates of true scores obtained with the regression method; for the six classes, the false negative rates ranged from .001 to .02, with a mean of .01.

False Positive Rate. Using the direct method to estimate true scores, the false positive rates for the six classes ranged from .24 to .66, with a mean of .53. (See Table 1.) Using true scores estimated with the regression method, the false positive rates were slightly larger; for the six classes, the rates ranged from .65 to .91, with a mean of .81.

Discussion

In general, the pass-fail decision reliabilities and the false negative rates were encouraging. The dependability coefficient with cut-off based on results for all classes combined was .84, and the false negative rate was .01 with both the direct and regression methods of estimating true scores, respectively.

Not surprisingly, though, the false positive rates were high: .53 with the direct method and .81 with the regression method. High false positive rates were expected given that the Examination Pass Level was adjusted downward from the average mastery level set by faculty case authors in order to guard against false negative errors. At this stage of the development of the Post-Clerkship Examination, it was felt that any incorrect pass/fail
decisions should err in favor of the student. By adjusting the Examination Pass Level downward by 1.29 SEM from the average mastery level, the number of students failing across all six classes decreased from 25% to 6%, so naturally the number of false positives would be expected to increase considerably. (See Table 2.) If there had been no adjustment and the Examination Pass Level had been set exactly at the average mastery level, the false positive rates would have been .18 with the direct method and .40 with the regression method. However, these lower false positive rates with the unadjusted passing level would have been obtained at the expense of higher false negative rates, namely .10 and .12 with the two methods respectively. In addition, the reliability of the pass-fail decisions would have dropped from .84 with the adjusted passing level to .68 with the unadjusted. The results for two intermediate-level adjustments for measurement error are also presented in Table 2.

Clearly, as shown in Table 2 and as would be expected, the false negative and false positive rates are dependent upon standards set by faculty, plus any adjustment deemed necessary for measurement error. Our faculty were very concerned about making false negative decisions (i.e., failing students who really should pass) and, consequently, adjusted the passing level downward by 1.29 SEM below the average mastery level, so that a borderline passing student with true score exactly at the mastery level would have only a 10% chance of incorrectly failing the examination. Unfortunately, this led to a misunderstanding among our faculty that the false negative rate for the examination as a whole was 10%, rather than that it was 10% for this hypothetical borderline passing student. It was this misunderstanding that demonstrated a need for these analyses, in particular,