DOCUMENT RESUME

ED 346 119 TM 018 366

AUTHOR Reshetar, Rosemary A.; And Others
TITLE An Adaptive Testing Simulation for a Certifying Examination.
PUB DATE Apr 92
NOTE 11p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992). Research supported by the American Board of Internal Medicine.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Classification; Computer Assisted Testing; *Computer Simulation; *Cutting Scores; Estimation (Mathematics); Evaluation Methods; Higher Education; *Licensing Examinations (Professions); Pass Fail Grading; *Physicians; Test Construction; Test Format; *Test Length; Test Use
IDENTIFIERS Ability Estimates; Two Parameter Model

ABSTRACT

This study examined performance of a simulated computerized adaptive test that was designed to help direct the development of a medical recertification examination. The item pool consisted of 229 single-best-answer items from a random sample of 3,000 examinees, calibrated using the two-parameter logistic model. Examinees' responses were known. For tests of 60, 120, and 180 items, estimation error and the accuracy of pass/fail classification decisions were studied. Ability estimates were stable across test length changes, and accurate estimates were obtained with all three test lengths. However, it is recommended that overall pass/fail decisions be based on longer tests, especially when the cutscore is close to the mean. This initial application suggests that computerized adaptive testing has promise in professional evaluation settings. Six tables present study data, and there is a 10-item list of references. (Author/SLD)

An Adaptive Testing Simulation for a Certifying Examination

Rosemary A. Reshetar*, Judy A Shea**, and John J. Norcini*

* American Board of Internal Medicine

** University of Pennsylvania

This research was supported by the ABIM but does not necessarily reflect its opinions.

Paper presented at the Annual Meeting of the American Education Research Association, 1992, San Francisco, CA.

An Adaptive Testing Simulation for a Certifying Examination

Rosemary A. Reshetar*, Judy A Shea**, and John J. Norcini*

* American Board of Internal Medicine

** University of Pennsylvania

Abstract

This study examined performance of a simulated computerized adaptive test that was designed to help direct development of a medical recertification examination. The item pool consisted of single-best-answer items (n=229) calibrated using the 2-parameter logistic model. Examinees' responses were known. For tests of 60, 120 and 180 items, estimation error and accuracy of pass/fail classification decisions were studied. Ability estimates were stable across test length changes and accurate estimates were obtained with all three test lengths. However, it is recommended that overall pass/fail decisions be based on longer tests, especially when the cutscore is close to the mean.

An Adaptive Testing Simulation for a Certifying Examination

Rosemary A. Reshetar, American Board of Internal Medicine
Judy A. Shea, University of Pennsylvania
John J. Norcini, American Board of Internal Medicine

Introduction

Recently the American Board of Internal Medicine (ABIM) introduced a plan for recertification (Benson, 1991). One component of the program will require examinees to take a final, secure examination. It will consist of 60-item modules, and examinees will need to pass each module in aggregate. In the first years of the program, the examination will be administered in the typical paper-and-pencil format. However, one possibility being considered for future administration is computerized adaptive testing (CAT). The primary advantage of CAT is that is increases the accuracy and efficiency of testing by selecting items for presentation that are matched to the ability level of the examinee, thus reducing both estimation error and test length (Green, 1983; Wainer, 1990).

Item response theory (IRT) (see, for example, Lord, 1980) is currently used in most applications and discussions of CAT, and is applied in this study. Optimally, the item pool for an adaptive test would include a large number of items that fit the IRT model of choice in the appropriate range(s) of difficulty (Green, Bock, Humphreys, Linn & Reckase, 1984). Even though examinees may be administered different numbers of items, as long as the item pool meets the requirement of unidimensionality, ability estimates obtained via adaptive testing are directly comparable from one examinee to the next (see, for example, Wainer, 1990).

During the administration of an adaptive test to each examinee, an individual ability estimate and standard error of the ability estimate on the theta scale may be calculated. Items are administered one at a time until the established confidence interval around the individual examinee's ability estimate no longer includes the cutting score. Thus a pass/fail decision is made with a specified level of accuracy (e.g., ± 3 standard errors). Even though CAT's of variable lengths may yield equally precise ability estimates for examinees, as one makes the transition from conventional testing to CAT it may be necessary to administer tests of fixed length in order to gain acceptance of the presentation mode. The purpose of the present study is to explore some of the implications related to using adaptive mastery tests of different fixed lengths for a recertification examination. Although other professional testing agencies have explored similar issues (Bergstrom & Lunz, 1991; Bosma & Dvorak, 1987), most investigators have used the 1-parameter model. Past work shows that the 2-parameter model, which allows items to vary in discrimination is more appropriate for ABIM data (Shea & Norcini, 1988).

The first interest of the study is in the relationship between test length and error of estimation. Three fixed-length CATs will be examined: 60, 120, and 180 items. A second interest is in the relationship between test length and classification decisions. It is anticipated that the ability level of most examinees for recertification would be fairly high, as they will have passed an initial certification examination and been in practice for several years. Thus, for most examinees a passing decision could be made with a high degree of certainty as their ability estimate would be well above the cutting score. This study will look at the percentages of examinees for initial certification who pass at six different hypothetical cutscores, who have "uncertain" decisions, and for whom the pass/fail outcome differs depending on the length of the test. The findings will help direct future CAT development for the recertification program.

Methods

An item bank was created that consisted of all 229 single-best-answer items used in one year's certification examination in internal medicine. All of the items were pretested before being selected for use in the certification examination. As a result, all items were reviewed for their statistical properties using p-values and r-biserial values, as well as for content considerations. Generally, extremely easy and extremely difficult items were not selected. For the total group of 8242 examinees, the mean p-value was .66 and the mean r-biserial value was .36. Item parameters were calibrated with PC-BILOG using the 2-parameter logistic model and marginal maximum likelihood estimation (Mislevy & Bock, 1990). The prior distributions on slope (discrimination) had a mean of 0 and a standard deviation of 1. Ability was assumed to have a standard normal distribution. Items were calibrated with the 5266 first-time takers of the examination; their ability estimates were centered on 0. Estimated item difficulties had a mean of -1.782 (SD=1.540) and item slopes had a mean of .323 (SD=.120)

A random sample of 3000 examinees was selected from the total test population of 8242. Sixty-four percent of the total group were first-time takers. The entire group was relatively homogeneous in that most had recently completed the end of a long training process. Their actual responses were used for the CAT. Three fixed-length adaptive tests were simulated for each examinee: 60, 120, and 180 items. For each of the tests, the first two items were selected randomly and the examinee's actual responses for those items were retrieved. Following that, a maximum likelihood ability estimate was calculated and maximum information was used for item selection; i.e., the item yielding the most information at the current ability estimate was next selected. This process was repeated until the desired test length was achieved.

The first part of the analyses provides an overview of the resulting ability estimates and standard errors given variations in test length. The second section of the analyses addresses issues pertinent to mastery testing by examining classification decisions with six hypothetical cutting scores; for these analyses, the 180-item test was considered the gold standard.

Results

Test Length and Error of Estimation

Summaries of the ability estimates are given in Table 1. Correlations between ability estimates were high: .9737 between the 60 and 120 item tests, .9633 between the 60 and 180 item tests, and .9907 between the 120 and 180 item tests. The mean ability estimates decreased slightly as test length increased.

Differences in ability estimates and root mean square errors (RMSE) for each pair of tests were calculated for each examinee and are summarized in Table 2. RMSEs were calculated comparing the shorter test to the longer test, and are a measure of the increase in estimation error that results from using the shorter test. Mean differences close to zero reveal that, on average, little shift in ability estimates is detected when the number of items in the test is changed. The RMSEs show that there are some individual differences in ability estimates. The smaller mean difference and RMSE found between the 120 and 180 item tests indicates little change is found between ability estimates for these two test lengths.

Table 3 summarizes the standard errors of the ability estimates. As expected, increases in test length corresponded to decreases in the standard errors of the individual ability estimates. Correlations between standard errors were also high: .9584 between the 60 and 120 item tests, .9423 between the 60 and 180 item tests, and .9854 between the 120 and 180 item tests. As with the ability estimates, decreases in standard errors were smallest, and the correlation of standard errors was highest, between the 120 and 180 item tests.

Test Length and Classification Decisions

In order to study classification decisions, six cutscores were evaluated: -3, -2.5, -2, -1.5, -1, and -0.5. For each test length, the percentage of examinees passing at each cutscore is given in Table 4. With all cutscores between -3 and -1, the percentages of examinees passing increased slightly as the test length increased. For example, at a cutscore of -2.5, 97.7% of the examinees passed with a 60 item test and 98.3% passed with a 180 item test. Thus, even though mean ability estimates were higher for the 60 item tests, the greater variability of these estimates served to fail more examinees with the lower cutting scores.

The certainty of the pass/fail decision was calculated by tallying the percentage of examinees whose pass/fail status was "uncertain", that is, those for whom the interval of their ability estimate plus or minus 2 and 3 standard errors included the cutscore. Results in Table 5 show that the percentage of "uncertain" classifications increase as the cutscore approaches the mean of the ability distribution, and there are always fewer "uncertain" decisions with the longer tests. These differences can be explained by the lower standard errors of the ability estimates with larger numbers of items, coupled with the group's distribution of ability centered between 0 and -0.5. With the stringent interval of ± 3 standard errors, relatively few examinees would be classified as "uncertain" at the lower cutscores, representative of many certifying and licensure examinations.

The inconsistencies in pass/fail decisions between the two shorter test lengths and the 180 item test are presented in Table 6. The percentages of examinees who change status in each direction are shown for the six cutscores. As the cutscore is moved closer to the mean ability level of the group, the percentages of classification changes increase. Also, at each cutscore there are fewer changes in outcome between the 120 and 180 item tests than between the 60 and 180 item tests.

Discussion and Conclusions

In summary, mean total group ability estimates were very similar and individual estimates were reasonably stable across test lengths. The slight lowering of ability estimates corresponding to increased test length may be traced to limitations of the item bank. Specifically, with a small item bank and a relatively homogeneous group of examinees whose ability estimates were centered about 1.5 standard deviations higher than the items' mean difficulty values, many of the items administered in the longer tests were not optimally matched to the examinees' ability levels.

Increases in test length resulted in decreases in estimation error as expected. The RMSEs were consistently smaller than the average standard error of an examinee's ability estimate, indicating that shifts in estimates due to test length changes were within an acceptable range. The small changes in precision noted between the 120 and the 180 item tests indicate little is gained

statistically from this test length increase. Classification uncertainties and inconsistencies were most notably affected when the cutscore was located closer to the mean ability level. It is reassuring that few differences attributable to test length were noted at the lower cutscores, those most representative of licensure and (re)certification examinations (e.g., -2.0 to -3.0).

Generalizations of these conclusions are limited by several factors. First, maximum likelihood methods were used for ability estimates. Properties of other types of ability estimation methods such as Bayesian methods should be explored (Hambleton & Swaminathan, 1985). Second, the only criterion for item selection was maximum information. In an actual application a more sophisticated selection strategy would probably be after a well-defined blueprint. Third, the item bank was small for the longer test length, requiring examinees to take items that were not well suited for their abilities.

In general, this initial application suggests that CAT has promise in professional evaluation settings where the examinees' abilities are homogeneous, and cutscores are relatively low. It is expected that the examinee group for recertification will be less homogeneous than the examinee group for initial certification as differences between practical experiences gained after initial certification are probably greater than differences between training experiences in accredited programs. However, it is also expected that the ability level of many recertification examinees will be well above the established cut-score. Thus results of this study can be applied to recertification as well as initial certification CAT development. Accuracy of estimation is achieved with relatively short test lengths. Ability estimates for 60-item modules are sufficiently precise for score reporting, given appropriate caveats. However, it is recommended that overall pass/fail decisions be based on somewhat longer tests, especially when the cutscore is close to the mean. Future research which addresses issues of content consideration, item bank development, and other statistical methods for administration and scoring is suggested.

Table 1
Means and Standard Deviations
of Ability Estimates by Test Length

Test Length	Mean	SD
60 Item	-.1913	1.1155
120 Item	-.2073	1.0714
180 Item	-.2171	1.0557

Table 2
Summary of Ability Estimate Changes
By Test Length

Comparison	Mean Difference	RMSE
$\theta_{60} - \theta_{120}$	.0160	.2550
$\theta_{60} - \theta_{180}$	.0257	.3012
$\theta_{120} - \theta_{180}$	.0098	.1466

Table 3
Means and Standard Deviations of
Standard Error Estimates by Test Length

Test Length	Mean	SD
60 Item	.3955	.0656
120 Item	.3145	.0485
180 Item	.2846	.0419

Table 4
Percentage of Candidates Passing at Each Cutscore by Test Length

Test Length	Cutscore -3.0	-2.5	-2.0	-1.5	-1.0	-0.5
60 Item	99.3%	97.7%	94.8%	88.6%	78.2%	60.8%
120 Item	99.3%	98.0%	95.5%	89.2%	78.6%	61.3%
180 Item	99.6%	98.3%	95.7%	89.5%	78.6%	60.9%

Table 5.
Percentages of "Uncertain" Pass/Fail decisions at ± 2 and ± 3 S.E.'s

Test Length	Cutscore -3.0	-2.5	-2.0	-1.5	-1.0	-0.5
± 2 Standard Errors:						
60 Items	2.9	6.4	13.1	24.5	39.5	51.9
120 Items	2.1	4.2	9.7	18.6	32.0	44.0
180 Items	1.8	3.9	8.6	17.2	29.5	41.4
± 3 Standard Errors:						
60 Items	5.1	11.4	22.2	40.7	59.7	71.3
120 Items	3.2	7.7	16.1	30.7	49.0	62.2
180 Items	2.8	6.8	14.4	27.8	45.6	57.9

Table 6
Percentages of Classification Inconsistencies

	Cutscores					
	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5
Pass 60 and Fail 180	<0.1	0.1	0.4	1.1	2.5	4.3
Fail 60 and Pass 180	0.3	0.8	1.3	2.0	2.9	4.4
Pass 120 and Fail 180	<0.1	0.1	0.2	0.6	1.4	2.1
Fail 120 and Pass 180	0.3	0.4	0.4	1.0	1.5	1.6

References

Benson, J.A. Jr. (1991). Certification and recertification: One approach to professional accountability. Annals of Internal Medicine, 111(3), 238-242

Bergstrom, B. & Lunz, M. (1991). Confidence in pass/fail decisions for computer-adaptive and paper-and pencil examinations. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Bosma, J. & Dvorak, E. M. (1987). Simulating adaptive administration of a nursing licensure examination. Paper presented at the annual meeting of the American Educational Research Association. Washington, D.C.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21(4), 347-360.

Green, B.F., Jr. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement. (pp.69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Mislevy, R.J. & Bock, R.D. (1990). PC-BILOG: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Shea, J. A. & Norcini, J. J. (1988). The validity of ability estimates in item response theory. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

Wainer, H. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.