

Annotated Bibliography of ASCP Board of Registry Publications on CAT

Sekula-Wacura, R, Brito, C (2000) **A Review of CAT Review**, *Laboratory Medicine*, 31, 8, 442-444.

ASCP Board of Registry examinations are administered in a computerized adaptive format. This article discusses the impact of the examination review component on the final candidate score. Overall, the review process benefits the candidate.

Stahl, J.A. and Lunz, M.E. (1993). **Assessing the Extent of Overlap of Items Among Computerized Adaptive Tests**. Presented at the Annual Meeting of the National Council of Measurement in Education in Atlanta, Georgia in 1993.

The purpose of this paper is to assess the degree of item overlap among computerized adaptive tests. Overlap is defined as the percent of common items among adaptive tests. Three hypotheses are investigated. (1) Examinees who are close in ability will have a higher percentage of overlapping items than mean percentage of overlapping items regardless of the examinee ability distribution. (3) The order of item presentation will differ regardless of the proximity of examinee abilities or the size of the item bank. Data from five medical certification examinations administered in 1991 are used to explore these hypotheses. In general, all three hypotheses are supported. To control item overlap, the minimum item bank size must be in the range of 400-500 items and an adequate bank size be in the 600-800 item range or larger.

Lunz, M.E., Stahl, J.A. and Bergstrom, B.A. (1993), **Test Targeting and Precision Before and After Review on Computerized Adaptive Tests**. Paper presented at the annual meeting of the National Council of Measurement in Education in Atlanta, Georgia.

The purpose of this study was to assess the effect of allowing candidates to review and alter responses on test targeting and precision of measurement on computerized adaptive tests (CAT). After the computer adaptive test was complete, candidates were allowed to review and alter their responses. For each candidate, estimated ability, accuracy of targeting and test precision were calculated before and after review.

Candidate ability estimates after review improved slightly, but are correlated at .99 with before review estimates. Accuracy of targeting as affected by review depending upon the number of items altered, the direction of change and the candidate's ability. Test precision was unaffected.

Stone, G.E. and Lunz, M.E. (1993). **The Effect of Review on the Psychometric Characteristics of Computerized Adaptive Tests**. Paper presented at the Annual Meeting of the American Educational Research Association in Atlanta, Georgia.

The effect of reviewing items and altering responses on examinee ability estimates, test precision, test information, decision confidence, and pass/fail decision accuracy for computerized adaptive tests is explored. Subjects were two different populations of examinees taking different computerized certification examinations. For purposes of analysis each population was divided into three ability levels (high, medium and low). Ability measures before and after review were highly correlated, but slightly lower examinee pass/fail decision confidence levels were found after review. Decision accuracy is most affected for examinees with estimates close to the pass point. Decisions remained the same for 94% of the examinees. Test precision was not affected by review, and the information lost due to review can be recovered by the addition of one item.

Bergstrom, B.A. & Lunz, M.E. (1992). **Confidence in Pass/Fail Decisions for Computer Adaptive and Paper and Pencil Examinations**. *Evaluation and the Health Professions*, 15, 4, 453-464.

This study compared the level of confidence in pass/fail decisions obtained with CAT and pencil and paper (p&p) tests. Subjects were 645 medical technology students from across the country. Examinees took a variable length CAT and two fixed length p&p tests. Results show that greater confidence in the accuracy of the pass/fail decisions is obtained for more examinees when the CAT implements a 90% confidence stopping rule than with p&p tests of comparable length.

Bergstrom, B.A. & Lunz, M.E. **Equivalence of Rasch Item Calibrations and Ability Estimates Across Modes of Administration**. In Mark Wilson (Ed.). *Objective Measurement*. NJ: Ablex.

This paper explores two related issues to determine whether item precalibrations from p&p tests are appropriate for use in CATs. The first addressed the equivalence of item calibrations from p&p and CAT administrations. The second addressed the equivalence of examinee ability estimates when item precalibrations from p&p tests versus item calibrations from CATs are used for the tailoring algorithm. Three hundred and twenty-one medical technology students provided data for the precalibration of 726 items. The correlation for examinee ability estimates was .99 and for item calibrations was

.90. Some item calibrations shifted but most remained consistent within the limits of error. Item shift did not affect the ordering of examinee ability estimates.

Bergstrom, B.A. & Lunz, M.E. **Computer Adaptive Testing: A National Pilot Study.** In Mark Wilson (ed.). *Objective Measurement*, NJ: Ablex.

A national pilot study was undertaken to ascertain the psychometric, psychological and social attributes of CATs. Students in medical technology programs took a p&p test and an individualized CAT. Students were randomly placed in experimental conditions to ascertain the effect of altering: 1) the difficulty of the start, 2) the targeted level of difficulty of the test, 3) minimum test length, and 4) the opportunity to skip, defer or review items on the CAT. Students who took the p&p test first performed significantly better on the CAT than students who took the CAT first. Students who were allowed to skip items on the CAT performed significantly better than students who had no opportunity to skip, defer, or review answers. Altering the difficulty of the start, the targeted level of difficulty or the minimum test length did not affect examinee ability estimation.

Bergstrom, B.A., Lunz, M.E. & Gershon, R. (1992). **Altering the Level of Difficulty in Computer Adaptive Testing.** *Applied Measurement in Education*, 5, 2, 137-149.

This study examines the effect of altering test difficulty on examinee ability measures and test length in a CAT. Examinees were randomly assigned to three test difficulty conditions and administered a variable length CAT. Examinees in the hard, medium, and easy conditions had, respectively, a 50, 60 or 70% probability of getting each item presented correct. The results show that altering the probability of a correct response does not affect estimation of examinee ability and that taking an easier CAT only slightly increases the number of items necessary to reach specified levels of precision. These results indicate that, with an item pool of sufficient depth and breadth, acceptable targeting to varying levels of test difficulty is possible.

Bergstrom, B.A. & Stahl, J.A. (1992). **Assessing Existing Item Bank Depth for Computer Adaptive Testing.** Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

This paper reports on a method for assessing the adequacy of existing item banks for CATs. The method takes into account content specifications, test length and stopping rules and can be used to determine if an existing item bank is adequate to administer a CAT efficiently across differing levels of examinee ability. Simulated data show that the adequacy of the bank can depend upon the stopping rule implemented. For example, when a specified standard error of measurement is used as the stopping rule, test length is variable, since all examinees are tested to the same level of precision regardless of the number of items required to reach the specified SEM.

Bergstrom, B.A. & Gershon, R. (1992). **Comparison of Item Targeting Strategies for Pass/Fail Computer Adaptive Tests.** Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

This paper demonstrates that the most useful method of item selection for making pass/fail decisions with a CAT depends on the distribution of the examinee population, the stopping rule implemented, and the length of the test. Eighty-six medical technology students took a CAT in which the items were targeted to the ability of the examinee. Overall, if test length is sufficient, targeting items at the ability of the examinee and using a confidence level stopping rule results in the most efficient CAT for making a pass/fail decision.

Bergstrom, B.A. (1992). **Ability Measure Equivalence of Computer Adaptive and Pencil and Paper Tests: A Research Synthesis.** Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

This paper reports on existing studies which compare the results of actual administrations of p&p tests and comparable CATs, and uses meta-analysis to compare and synthesize the results of twenty studies from eight research reports. The effect size computed was the standardized mean difference between the ability measure estimated by the CAT and the ability measure estimated by the p&p test. Most studies, despite differences in test content, age of examinees, IRT model used and study design, show comparable mean ability measures on the CAT and p&p test versions. In all cases where mean differences are statistically different, mean measures are higher for a pre-existing p&p test.

Lunz, M.E., Bergstrom, B.A. & Wright, B.D. (1992). **The Effect of Review on Student Ability and Test Efficiency for Computerized Adaptive Tests.** *Applied Psychological Measurement*, 16, 1, 33-40.

The effect of reviewing items and altering responses on the efficiency of CATs and the resultant ability estimates of examinees were explored. One sample of students was randomly assigned to a review condition; their test instructions indicated that each item must be answered when presented, but that the responses could be reviewed and altered at the end of the test. The other student sample did not have the opportunity to review and alter responses. Within the review condition, examinee ability estimates before and after review were correlated .98. The average efficiency of the test was

decreased by 1% after review (based on using the test information function). Approximately 32% of the examinees improved their ability estimates after review but did not change their pass/fail status. The review group scored slightly higher than the non-review group. Disallowing review on adaptive tests administered under these rules is not supported by these data.

Lunz, M.E. & Bergstrom, B.A. (1991). **Comparability of Decisions for Computer Adaptive and Written Examinations.** *Journal of Allied Health*, 20, 1, 15-22.

The purpose of this study was to determine the comparability of examinee ability measures, pass/fail decisions, and confidence in the accuracy of pass/fail decisions on written fixed-length and CATs. Medical technology students took a written test of 109 questions and a CAT that included 50 to 100 items tailored to the ability of each student. Results indicated that ability measures correlated .84. Decisions were made with 90% confidence in their accuracy for 72% of the examinees on the computer adaptive test, and for 58% of the examinees on the written test. The results support previous findings that CATs can provide the same or higher level of confidence in pass/fail decisions with substantially fewer questions than fixed-length written tests.

Lunz, M.E. & Wright, B.D. (1990). **Criterion Standards from Benchmark Performances for Judge Intermediated Examinations.** Paper presented at the annual meeting of the American Educational Research Association, Boston.

This paper explains a technique for determining a criterion-referenced standard for oral, essay, or practical examinations requiring assessment by judges. The data from a histotechnology practical examination are analyzed with the Rasch FACETS model. A criterion standard is constructed from detailed gradings and global classifications of examinee performance chosen to “benchmark” the probable pass/fail region. The final criterion standard is chosen so that misclassifications of benchmark performances rated “competent” and “incompetent” is minimized.

Lunz, M.E., Bergstrom, B.A. & Gershon, R. (1990). **Test-retest Consistency of Computer Adaptive Tests.** Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

This paper reports on the test-retest reliability when alternate parallel forms of CATs are administered sequentially to the same examinee. One hundred and sixty-two students took two contiguous tests with the same test specifications but different items. The ability measures from the test and retest were found to correlate .95, demonstrating that differentiation among examinee measures is comparable regardless of the length of the test or the particular subset of items. For most examinees the same pass/fail decision is made.

Stahl, J.A., Lunz, M.E. & Snyder, J.R. (1990). **Validity of Statistical Detection of Bias in Test Items.** Paper presented at the annual meeting of the American Educational Research Association, Boston.

The purpose of this study was to explore whether or not judges can validate item bias detected from statistical analysis. The statistical analysis was based on Rasch model item calibration estimates. Two subgroups of training program graduates were analyzed: hospital based (n=1507), and university based (n=1017). The results indicate that 1) when consensus methods are used the judges demonstrate high agreement in correctly identifying the examinee subgroup identified by the item, 2) judge characteristics related to job responsibility and education do not enhance the ability of the judge to correctly classify statistically biased items, and 3) there is a significant difference in the ease with which items can be classified correctly by judges.