DOCUMENT RESUME

ED 429 119 TM 029 662

AUTHOR Wise, Steven L.
TITLE Comparison of Stratum Scored and Maximum-Likelihood Scored CATs.
PUB DATE 1999-04-00
NOTE 32p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Item Response Theory; Scores; *Scoring; *Test Construction

ABSTRACT

Outside of large-scale testing programs, the computerized adaptive test (CAT) has thus far had only limited impact on measurement practice. In smaller-scale testing contexts, limited data are often available, which precludes the establishment of calibrated item pools for use by traditional (i.e., item response theory (IRT) based) CATs. This paper introduces an alternative adaptive testing procedure--termed a stratum CAT--that requires no IRT methods for either item selection or proficiency estimation. In two simulation studies comparing stratum CATs to conventional tests and traditional CATs, stratum CATs were found to be substantially more efficient than conventional tests. In addition, with 100 or fewer examinee responses available per item, the efficiency of stratum CATs could match or exceed that of traditional CATs. The stratum CAT may prove to be a useful adaptive testing procedure in situations where limited item calibration data are available. (Contains 3 tables and 12 references.) (Author/SLD)

Comparison of Stratum Scored and Maximum-Likelihood Scored CATs

Steven L. Wise

James Madison University

Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, April, 1999.

Abstract

Outside of large-scale testing programs, the computerized adaptive test (CAT) has thus far had only limited impact on measurement practice. In smaller-scale testing contexts, limited data are often available, which precludes the establishment of calibrated item pools for use by traditional (i.e., IRT-based) CATs. This paper introduces an alternative adaptive testing procedure—termed a stratum CAT—which requires no IRT methods for either item selection or proficiency estimation. In two simulation studies comparing stratum CATs to conventional tests and traditional CATs, stratum CATs were found to be substantially more efficient than conventional tests. In addition, with 100 or fewer examinee responses available per item, the efficiency of stratum CATs could match or exceed that of traditional CATs. The stratum CAT may prove to be a useful adaptive testing procedure in situations where limited item calibration data are available.

The Stratum CAT: Administration and Scoring of Adaptive Tests Without Item Response Theory

The computerized adaptive test (CAT) has become highly familiar to the measurement community, with operational CATs currently being used in numerous large-scale testing programs. During the past quarter century, a large amount of psychometric research has been devoted to adaptive testing, with numerous studies focusing on an array of theoretical and practical CAT issues. The attractiveness of CATs is clear—along with the desirable features of computer-based testing (such as immediate test scoring and feedback of results to examinees) there is a dramatic increase in testing efficiency. Because the CAT algorithm matches the difficulty levels of the administered items to the proficiency levels of each examinee, far fewer items need to be administered for a CAT to yield score reliability equivalent to that of a conventional (i.e., nonadaptive fixed-length test).

Outside of large-scale testing programs, however, adaptive testing has thus far had far less impact on measurement practice. In smaller-scale testing situations—such as in classroom or school-based settings in which less than 200 examinees per year are tested—the use of conventional tests scored using classical measurement methods still predominates.

There are several factors that have contributed to this state of affairs. First, a major challenge in developing a CAT testing program is the development of an adequate item pool, which typically contains several hundred items. When IRT item selection and scoring methods are used, a substantial amount of data must be collected for each item; depending on the IRT model used, the recommended minimum number of examinee responses for each item ranges from 200 up to 1000 or more (Hambleton, 1989). Moreover, because it is usually infeasible to have an examinee receive every item in the pool, complex data-collection designs are often needed to gather sufficient data for item

calibration. In small-scale testing contexts, measurement practitioners may be discouraged by the time and effort required to establish IRT-calibrated item pools of adequate size to support adaptive testing. A second reason for the limited presence of CATs in small-scale testing situations is that it is generally assumed that CATs require IRT-based measurement methods to allow score comparisons across examinees who have each taken unique tests. Many practitioners who do not have a strong background in IRT may feel that they do not have the psychometric skills needed to implement and maintain a CAT testing program. Third, IRT-based CATs require sophisticated computer software for item banking, test administration, and scoring. Such software is typically expensive to purchase or would require substantial resources (in time and expertise) to develop. Collectively, these factors preclude many measurement practitioners from seriously considering using a CAT format.

The primary purpose of the present research was to develop a simpler method for adaptive testing—one that does not require IRT and large calibration data sets while yielding efficiently-obtained scores that are better estimates of examinee proficiency than those obtained from conventional tests. The ultimate goal is to develop a CAT procedure that could be useful to practitioners who both do not have a great deal of data and wish to use classical test theory-based scoring methods.

In addition to the general goal of developing a CAT procedure that did not require IRT, four criteria were established that guided the development of a new procedure. These included:

1. The scoring method could be easily understood by examinees.
2. The item selection method would provide effective item exposure control.
3. The scores yielded by the new CAT method would correlate substantially higher with true examinee proficiency than those from a conventional test of the same length.
4. The amount of data needed, per item, to establish a useful item pool would be less than 200 examinees per item.

The development of the new procedure was considered in terms of the two basic components of the CAT process: proficiency estimation and test administration. Proficiency estimation was based on a new method, while test administration was based on a modification of an older, established method.

Proficiency Estimation

In a CAT, an examinee receives a series of test items. After each test item, the examinee's proficiency estimate is updated, which is then used by the testing algorithm to select the item that would be most informative to administer next. In an IRT-based CAT, an examinee's proficiency estimate is usually based on the maximum value of a likelihood function (possibly incorporating additional Bayesian prior information regarding the population proficiency distribution). This likelihood function combines the characteristics of the administered items with the examinee's performance on those items to identify the proficiency level that was most likely to have resulted in his or her performance on those items.

Although this method for estimating proficiency is mathematically straightforward, it is not easily understood by examinees (or many practitioners) who are used to conventional number-correct scoring. In most testing contexts, relatively few examinees understand the principles of maximum likelihood estimation, and thus do not really understand how their proficiency estimates are calculated. Moreover, even if examinees did understand maximum likelihood, they might have difficulty accepting some of its features. Stocking (1996) noted that, when a three-parameter IRT model is used in a CAT, the credit given an examinee for an item response is dependent on that examinee's proficiency level. For example, if a lower-proficiency examinee and a higher-proficiency examinee both pass a difficult item, the lower-proficiency examinee will receive less credit for the correct response. Although such differential credit makes sense from the standpoint of IRT, it may appear unfair to examinees.

Both Stocking (1996) and Green (1997) discussed methods for making scores from CATs more understandable. In Stocking's (1996) method—which she call equated number-correct scoring—the item responses to the CAT are used to estimate the number of items that an examinee would have passed if he or she had taken the entire calibrated item pool. Green (1997) explored the effects of using a fixed (i.e., not dependent on examinee proficiency) scoring weight for each item equal to the theta value for which the item's information was maximum. Because both Stocking's and Green's methods involved IRT item parameters, however, they were not investigated in this study.

Conceptually, IRT-based scoring over an accumulated sequence of administered items (as in a CAT) can be largely characterized by two relatively basic principles:

1. Whenever an item is passed, an examinee's score is increased by an amount that is largely dependent on the difficulty level of that item. The more difficult the passed item, the larger the score increase.
2. Whenever an item is failed, an examinee's score is decreased by an amount that is largely dependent on the difficulty level of that item. The less difficult the failed item, the larger the score decrease.

Simply put, CAT examinees get rewarded more for passing harder items than easier items, and they are penalized more for failing easier items than harder items.

Admittedly, these principles oversimplify the process actually used in estimating examinee proficiency during IRT scoring, particularly when two- or three-parameter IRT models are used. The principles ignore the information inherent in the discrimination and guessing parameters of the administered items. It can be argued, however, that item difficulty is the characteristic that plays a dominant role in IRT-based proficiency estimation. Green (1997) implied these two principles, when he concluded that maximum likelihood scores, Bayesian scores, and equated number-correct scores,

> "are all refinements of the simple-to-understand fixed-weight scoring methods. That is, a CAT score is essentially a system for giving more credit for more difficult

questions, coupled with some credit for the difficulty of items answered incorrectly." (p. 5).

Additionally, it is the estimation of discrimination and guessing parameters that tend to require larger sample sizes for accurate item calibration under IRT parameter estimation procedures. In accordance with the criterion of using procedures that require only limited data per item, a scoring method was developed that differentiated the items only in their difficulties. This scoring method was termed stratum scoring.

The Stratum Scoring Method

In stratum scoring, an item pool is subdivided into a number of ordered equivalence classes, or strata, that correspond to different levels of item difficulty. The score assigned to an examinee's item response is a function of both the correctness of the response and that item's stratum. The higher the stratum, the higher the reward for a correct answer and the lower the penalty for an incorrect answer. Table 1 illustrates a stratum scoring model in which evenly spaced unit scoring weights are used with six difficulty strata. The scoring method is consistent with the two IRT scoring principles described earlier. At each step in a sequence of administered items, the more difficult a passed item, the more an examinee's score is increased. The less difficult a failed item, the more an examinee's score is decreased. Note, however, that IRT difficulty parameters are not necessary for stratum scoring as classical item difficulty (i.e., predetermined proportions of examinees passing the items) could be used to group items into strata.

An examinee's total stratum score on a test is equal to the sum of his or her item stratum scores. This score could be readily understood by practitioners and explained to examinees—thus allowing them to understand exactly how the test scores were calculated.

To explore the strength of concordance between stratum scoring and maximum likelihood proficiency estimation, stratum scoring was used to re-score examinee data from two operational CATs and compare the stratum scores to the corresponding

Table 1

An example of stratum scoring using six item difficulty strata and unit scoring weights

Stratum	Scoring Weight for Correct Answer	Scoring Weight for Incorrect Answer
1	+1	-6
2	+2	-5
3	+3	-4
4	+4	-3
5	+5	-2
6	+6	-1

IRT-based proficiency estimates. In the first data set, a 20-item CAT (using maximum information item selection) drawn from a small pool of 90 algebra items had been administered to 243 introductory statistics students. The items were ranked with respect to item difficulty parameters and divided into 6 strata each containing 15 items. A stratum score was then obtained for each examinee based on the stratum membership of the items that had been administered during the IRT-based CAT and the correctness of the examinee's response to each item. A very high correlation between the stratum scores and the maximum-likelihood proficiency estimates was found (r =0.98).

This procedure was repeated for a data set from a school district-wide CAT in which 287 examinees received a 20-item CAT drawn from a pool of approximately 1500 multiple-choice items. For this data set, the correlation between the stratum scores and the IRT-based proficiency estimates was found to be 0.99.

The high correlations in these two analyses suggested that the information provided by stratum scores can be highly congruent with that provided by the IRT-based proficiency estimates. In both data sets, however, the item pools (and corresponding CATs) were based on IRT models in which only difficulty parameters were estimated. Thus, although these results were encouraging, their generalizability to CATs that use item pools calibrated using multi-parameter IRT models remained unclear. This issue was investigated more thoroughly using simulation studies, the results of which are presented in this paper.

Test Administration

During the mid-1970's, Weiss and his associates conducted many of the early studies on adaptive testing. Much of this research was directed toward non-IRT methods. One particular method that they developed and studied was the stratified adaptive, or stradaptive, test (Weiss, 1973). Derived from Binet's strategy of ability testing, a stradaptive test uses an item pool in which the items have been grouped into strata based on IRT difficulty parameters. The items in each stratum are ordered according to decreasing values of their discrimination parameters, and items are drawn from each stratum in that order.

Although many test administration strategies could be used in a stradaptive test, the most common is the simple one-up, one-down branching rule in which an examinee is branched to the next higher stratum following a correct response, and branched to the next lower stratum following an incorrect response. If a correct (incorrect) response is given to an item from the highest (lowest) stratum, the next item administered is drawn from the same stratum.

A number of methods have been explored for scoring stradaptive tests (Thompson & Weiss, 1980; Vale & Weiss, 1975a, 1975b). Of these scoring methods, the mean IRT difficulty parameters of the administered items was found to possess the best criterion-related validity (Thompson & Weiss, 1980). In fact, Thompson and Weiss found that

mean difficulty scoring of stradaptive tests may have higher criterion-related validity than maximum likelihood or Bayesian scoring of the same item response data.

Stradaptive CATs have been found to be superior to conventional tests in precision of measurement (Bejar, Weiss, & Gialluca, 1977; Vale & Weiss, 1975b; Waters, 1977). Despite these encouraging results found with stradaptive CATs, they have fallen into disuse. CATs that use maximum information methods to select items, coupled with either maximum likelihood or Bayesian scoring methods, characterize virtually all current operational CATs—and throughout the remainder of this article will be referred to as traditional CATs. Given a well-calibrated item pool, maximum information item selection is more precise and efficient than the item selection method used in a stradaptive test.

The stradaptive test, however, provides a useful model for developing an adaptive test that does not use IRT in either item selection or proficiency estimation. The new adaptive testing procedure explored in this investigation, which is termed a stratum CAT, differs in three ways from Weiss's (1973) stradaptive procedure. First, item difficulty strata are formed on the basis of classical item difficulty (i.e., proportion passing) rather than IRT difficulty parameters. Second, items are drawn randomly from each strata, rather than in order of decreasing item discrimination. Third, proficiency is estimated using stratum scoring rather than mean difficulty of administered items. Given these modifications of the stradaptive procedure, the stratum CAT procedure uses no IRT methods, either in item selection or proficiency estimation.

The decision to select items randomly from strata in a stratum CAT may seem puzzling. Certainly, if one could estimate classical item difficulty then one could likely also estimate classical item discrimination. This would allow the items in each stratum to be ordered in discrimination (as in a stradaptive test), and selection of the most discriminating item available from a given stratum would provide more efficient measurement than use of random selection. Psychometric efficiency, however, is not the

only issue. A consequence of any item selection strategy that favors the most discriminating items is that such items will be administered in far more tests than will the least discriminating items. This issue of item exposure has emerged as one of the most serious practical problems faced by measurement practitioners in large-scale CAT programs. In smaller-scale settings—in which items are likely of more limited supply—it might be more desirable to use all available items at a comparable rate, even at the cost of some psychometric efficiency. It was decided, therefore, that alleviating item exposure problems by a random selection of items within strata was more important than gains in efficiency that would result from prioritizing the most discriminating items for item selection.

Having established the basic stratum CAT procedure, the next step was to investigate its usefulness as an adaptive test. This was accomplished using two simulation studies in which stratum CATs were compared to both conventional tests and traditional CATs.

Simulation Study 1

The stratum CAT procedure is far simpler than that of a traditional CAT. Consequently, its process of randomly selecting items from a relatively small number of strata is imprecise in comparison to the more mathematically sophisticated IRT methods based on item information functions that characterize traditional CATs. And the stratum CAT's proficiency estimation comprised of a simple sum of difficulty-weighted item scores (that ignore item discrimination) appears crude in comparison to maximum likelihood proficiency estimates. A primary issue, therefore, concerns the extent to which the increased psychometric efficiency that has been observed with traditional CATs is also exhibited by stratum CATs

The first simulation study compared stratum CATs, traditional CATs, and conventional tests under a variety of conditions. It was expected that stratum CATs would yield scores that were more precise than scores from conventional tests, but less

precise than scores from traditional CATs. Moreover, it was expected that a majority of the adaptive gains in precision yielded by traditional CATs would also be observed with stratum CATs.

Method

Item Pool

For use with each of the three test types, IRT item parameters (discrimination, difficulty, and guessing) were generated for 300 items. The discrimination parameters (a) were generated by random sampling from a uniform distribution ranging from .80 to 1.70. The difficulty parameters (b) were drawn from a uniform distribution ranging from -3.0 to +3.0. The guessing parameters (c) were drawn from a normal distribution with a population mean of .20 and standard deviation of .07. This item pool was used in all test simulations in which the data were assumed to fit the three-parameter logistic IRT model. For test simulations in which the data were assumed to fit the one-parameter model, the three-parameter item pool was modified by setting all a parameters to 1.0 and all c parameters to 0.0.

Stratum Assignment

Stratum CATs with four different numbers of item difficulty strata (3, 5, 7, 9) were investigated. The item pool was ordered by difficulty parameters, and strata were formed for each CAT by dividing this distribution into the desired number of strata. Thus, the three-strata CAT used three 100-item strata, and the five-strata CAT used five 60-item strata. Because investigations of test lengths up to 50 items were planned, however, the stratum assignments for the seven- and nine-stratum CATs required placement of higher numbers of items in the highest and lowest strata. Because very high (low) proficiency examinees might remain at the highest (lowest) stratum during testing, additional items were needed in the highest and lowest strata to accommodate examinees who passed or failed nearly all of their items. Accordingly, the seven-strata CAT had 45 items in the highest and lowest strata, and 42 items in each of the remaining strata. The

nine-strata CAT had 40 items in strata one and nine, 35 items in strata two and eight, and 30 items in each of the remaining strata.

Computer Programs

Computer programs were written in Fortran 77 for simulating conventional tests, traditional CATs, and stratum CATs. In the conventional tests, a true proficiency value was generated for a given hypothetical examinee, from which the probability of passing each item was calculated using the three-parameter logistic model. Then, for each item, a uniform random number between 0 and 1 was generated. If the probability of passing an item exceeded its random number, the item was passed by the examinee; otherwise, it was failed. The test score for each examinee was the number of items passed.

The traditional CAT programs administered fixed-length adaptive tests based on the three-parameter IRT model. An initial proficiency estimate of 0.0 was used, and a maximum information criterion was used to select the item to be administered at each step of the CAT. Whether an examinee passed an item was determined by the same procedure used in the conventional test. Proficiency was estimated using maximum likelihood, and bounded at -5.0 and +5.0 to prevent nonconverging proficiency estimates.

The stratum CAT programs administered fixed-length adaptive tests using a one-up, one-down branching rule. Each item that was administered from a stratum was randomly selected without replacement. Beginning with the middle stratum (e.g., from the third stratum for a five-strata CAT), an examinee was branched to the next higher stratum following a correct response, and to the next lower stratum following an incorrect response. If a correct (incorrect) response was given to an item from the highest (lowest) stratum, the next item administered was selected from the same stratum. Whether an examinee passed an item was determined by the same procedure used in the conventional test, and the item score was the stratum score based on unit scoring weights. The sum of the item stratum scores comprised the overall proficiency estimate for an examinee.

Test Simulations

The conventional tests of a given length were obtained by selecting items that spanned the difficulty range of the item pool. This was accomplished by ordering the item pool by difficulty, randomly choosing a low-difficulty starting item, and skipping a fixed number of items between those selected. For example, the 30-item conventional test was constructed by randomly choosing the 7th item, and choosing every 10th item from the 300-item pool until 30 items had been selected.

In each simulation, a conventional test, a traditional CAT, and four types of stratum CATs (using 3, 5, 7, & 9 strata) were performed. In addition, test length was varied (15, 20, 25, 30, 35, & 40 items) for each test, as was type of item pool (one-parameter or three-parameter). Each of the 72 combinations of test type, test length, and item pool were administered to independent samples of 10,000 hypothetical examinees whose true proficiencies were randomly generated from a standard normal distribution. For each simulated test, the squared correlation between estimated and true proficiency was computed, and this statistic provided the basis of comparison among the simulated tests.

Results and Discussion

The results of the simulations for the one-parameter item pool is depicted in Figure 1. At all test lengths, the traditional CAT showed higher squared correlations than the conventional test, reflecting the increased measurement precision that is characteristic of traditional CATs. The stratum CATs generally showed squared correlations with values between those found for the conventional tests and the traditional CATs, with higher values associated with higher numbers of strata. If the increase in squared correlation of the traditional CAT over that found with the conventional test is considered as a reference for a given test length, one can gauge the relative degree of "adaptive benefit" yielded by various stratum CATs. Across test lengths, the nine-strata CAT performed best, recovering 70 - 78% of the traditional CAT increase, followed by the

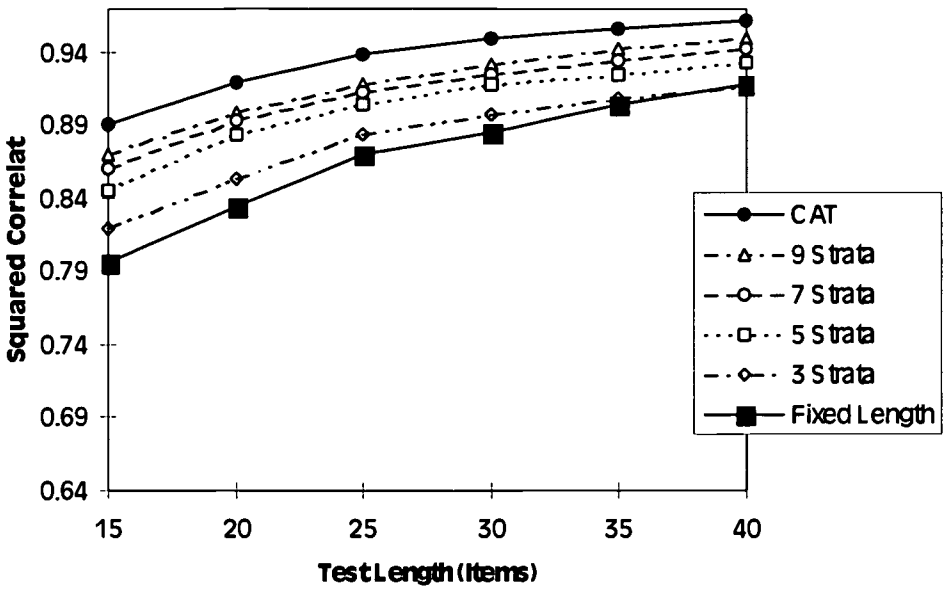

Figure 1. Squared correlations between actual and estimated proficiency for conventional and adaptive tests when data were generated from a 1PL model (true IRT parameters and stratum membership known).

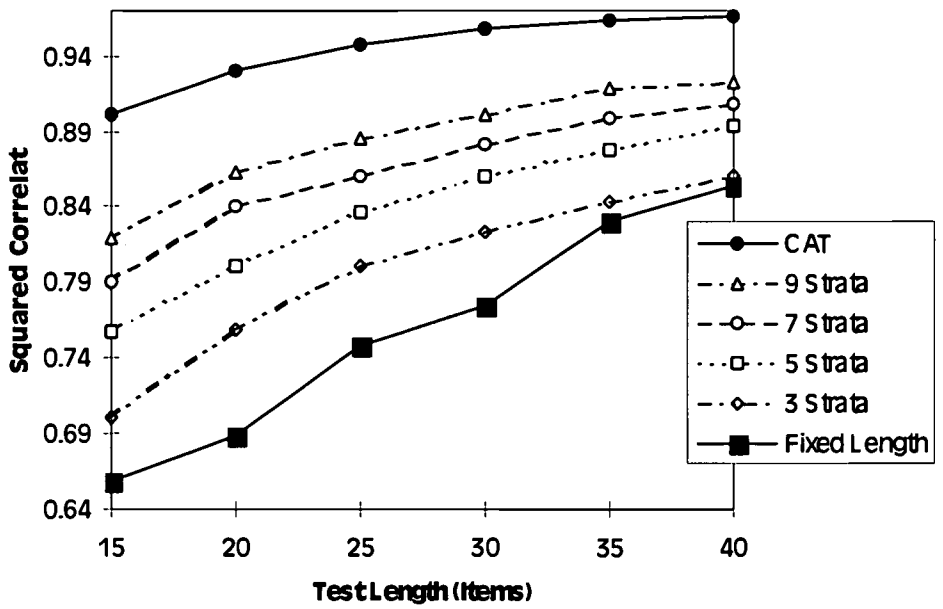

Figure 2. Squared correlations between actual and estimated proficiency for conventional and adaptive tests when data were generated from a 3PL model (true IRT parameters and stratum membership known).

seven-strata CAT (57 - 68%), the five-strata CAT (34 - 58%) and finally the three-strata CAT (0 - 24%).

A similar pattern of results was found for the three-parameter item pool (shown in Figure 2), though performance of the stratum CATs was not as strong as with the one-parameter pool. Across test lengths, the nine-strata CAT recovered 61 - 72% of the traditional CAT increase, followed by the seven-strata CAT (48 - 63%), the five-strata CAT (35 - 47%) and the three-strata CAT (6 - 29%).

Comparison of Figures 1 and 2 reveals the weaker performance of the stratum CATs in Figure 2. Moving from the one-parameter item pool to the three-parameter pool produced a decrease in the squared correlations for the conventional test and the stratum CATs, while it produced a small increase for the traditional CAT. These findings reflect the fact that the traditional CATs took advantage of the extra information provided by differences in item discrimination (in the three-parameter pool), while the conventional tests and stratum CATs did not.

The results clearly indicate, however, that the stratum CATs exhibited increased efficiency compared to conventional tests. For example, in Figure 1, the squared correlation for the 25-item, nine-strata CAT is equivalent to that from the 40-item conventional test. In Figure 2, the squared correlation for the 20-item, nine-strata CAT is slightly higher than that from the 40-item conventional test. Thus, it appears that stratum CATs are more efficient than conventional tests, as they can attain equivalent measurement precision with substantially fewer items.

The results of Study 1 showed that, when the population characteristics of the item parameters are known, stratum CATs—based on non-IRT methods—can recover most of the adaptive benefits of their IRT-based counterparts. This suggests that measurement practitioners who have neither the resources nor expertise to implement IRT-based CATs may still exploit the increased efficiency provided by stratum CATs.

The results of Study 1 should be interpreted with caution, however, as they are based on the unrealistic assumption that the true difficulty parameters of the items are known. In practice, item difficulty will typically be estimated empirically using the proportion of a group of examinees who pass the item (i.e., the item's p-value). In situations where only limited amounts of data are available per item, these estimated item difficulties will not be very stable, and some items consequently will not be assigned to their true strata. Such misclassifications should be expected to degrade the relationship between true proficiencies and those estimated via stratum scores, because incorrect scoring weights will be applied for those items. In contrast, conventional tests require no information regarding item difficulty, and thus would be unaffected by testing situations in which limited numbers of examinees are available. Therefore, although the results of Study 1 were encouraging, it remained unclear how efficiently stratum CATs would perform when stratum membership is based on limited data. This research question was investigated in Study 2.

Simulation Study 2

This study was designed to assess the efficiency of stratum CATs and traditional CATs when item characteristics are estimated from small samples of examinees. It was expected that—due to item misclassifications—the stratum CAT would not exhibit the same degree of increased measurement precision as found in Study 1. The traditional CAT, however, would also not be expected to fare as well as in Study 1 when items were calibrated on small samples. Hence, the objective of Study 2 was to assess the relative performance of stratum CATs and traditional CATs when limited data are available to estimate item characteristics.

Method

Calibration of Item Characteristics

Study 2 used the same 300-item pools as in Study 1. In Study 2, however, independent samples of 50 and 100 hypothetical examinees (whose proficiencies were

randomly selected from a standard normal distribution) were generated for each item pool to serve as calibration samples. For each examinee in a sample, item responses (0, 1) were obtained for each of the items in the pool by first computing the probability that the examinee would pass the item, then generating a uniform random number between 0 and 1 and scoring the item as passed if the probability exceeded the random number.

In each of the four calibration samples, p-values were computed for each item; these p-values were used as estimates of item difficulty in assigning items to difficulty strata. Assignment of items to strata followed the same procedure described in Study 1, except that p-values were used rather than IRT difficulty parameters. Because the p-values were computed from relatively small sample sizes—and thus were considerably vulnerable to sampling error—substantial numbers of items were assigned to different strata than in Study 1.

The data from each of the calibration samples were also used to estimate the IRT item parameters. These item calibrations were performed using BILOG (Mislevy & Bock, 1982). Because the sample sizes were so far below the minimum of 1000 examinees typically recommended for three-parameter model calibration (Hambleton, 1989), calibrations of the items from three-parameter pool were instead based on the one-parameter model.

Some of the items in each sample had p-values of either zero or one. For these items, BILOG was unable to estimate item difficulty parameters; consequently, the items were excluded from item pool used by the traditional CAT for that item pool/sample size combination. For the one-parameter pool, the numbers of excluded items for the samples of size 50 and 100 were 14 and 5, respectively. For the three-parameter pool, the respective numbers of excluded items were 23 and 12. Thus, IRT calibrations based on these small samples resulted in the loss of substantial numbers of items. However, items with p-values of zero or one were not excluded from use by the stratum CATs, because their stratum membership was clearly indicated. Items with p-values of zero were simply

assigned to the highest difficulty stratum, and those with p-values of one were assigned to the lowest stratum.

Proportional Stratum Scoring Weights

When the strata were formed on the basis of p-values, it became apparent that some strata covered a wider range of difficulty than others. Inspection of the midpoints of the ranges of p-values across the strata revealed an uneven spacing between the midpoints. Such uneven spacing—which was largely due to the calibration sample's proficiencies being drawn from a normal distribution—is analogous to the uneven spacing observed among decile points in normal distributions. The uneven spacing suggested that it might be more useful to scale the differences among the stratum scoring weights to be proportional to the differences among the stratum difficulty midpoints. Table 2 shows proportional scoring weights for nine strata from the three-parameter item pool. The lowest and highest weights for correct answers were set to +1 and +9, respectively, while the other weights were proportional to the differences among the p-value midpoints. Trial analyses showed that stratum scoring based on proportional weights yielded proficiency estimates exhibiting slightly higher correlations with true proficiency than those yielded by unit scoring weights. Because of these findings, and because unevenly spaced strata would typically be expected in practical applications of stratum scoring, proportional scoring weights were computed for each of the stratum CATs and used in Study 2.[1]

Test Simulations

Aside from the use of empirically-based IRT parameter estimates, strata based on empirical p-values, and the use of proportional stratum scoring weights, the Study 2 simulations were replications of those performed in Study 1. The conventional tests were constructed in the same way. Four types of stratum CATs and six test lengths were again simulated. It should be noted that, although the traditional CAT used estimated item

Table 2

Proportional scoring weights by the stratum CAT with nine strata (three-parameter item pool; sample size = 100)

Stratum	Range of p-values	p-value Midpoint	Scoring Weights for Correct Answer	Scoring Weights for Incorrect Answer
1	.98 - 1.0	.99	+1.00	-9.00
2	.95 - .97	.96	+1.28	-8.72
3	.79 - .94	.86	+2.20	-7.80
4	.69 - .78	.74	+3.30	-6.70
5	.56 - .68	.62	+4.40	-5.60
6	.42 - .55	.48	+5.69	-4.31
7	.30 - .41	.36	+6.79	-3.21
8	.22 - .29	.26	+7.71	-2.29
9	.03 - .21	.12	+9.00	-1.00

parameters in item selection and proficiency estimation, the probability that a given examinee passed a particular item was computed using the true item parameters.

Results and Discussion

Figures 3 and 4 show the results of the simulations for the one-parameter item pool. As was found in Study 1, the stratum CATs showed squared correlation values between those found for the conventional tests and the traditional CATs, with higher values associated with higher numbers of strata. Relative to the increase in squared

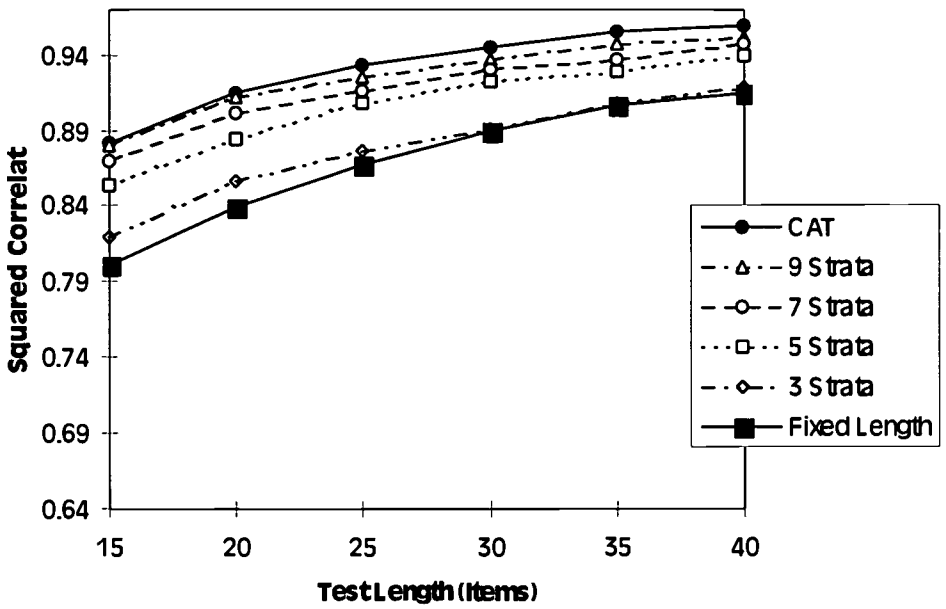

Figure 3. Squared correlations between actual and estimated proficiency for conventional and adaptive tests when data were generated from a 1PL model (IRT parameters and stratum membership based on a sample of 100 examinees).

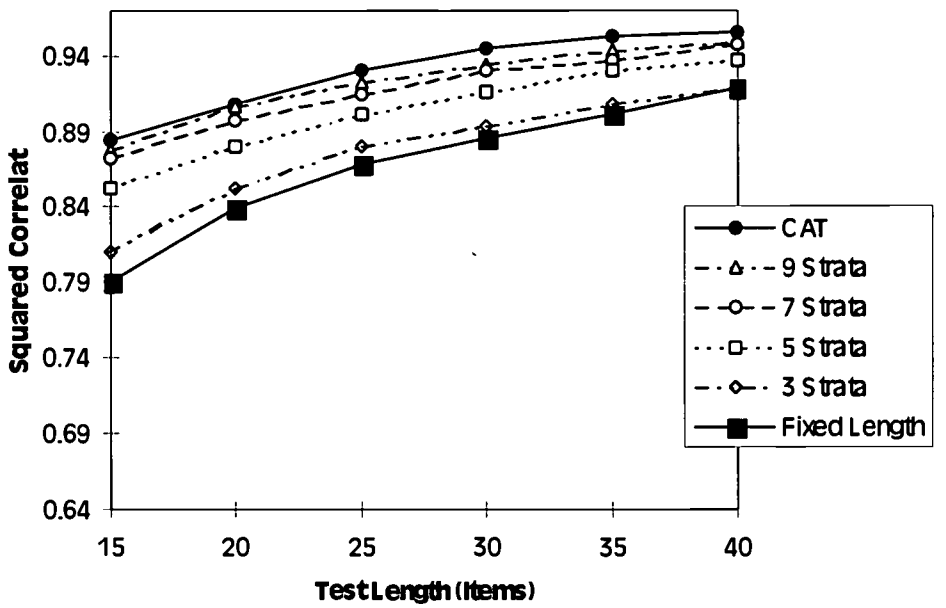

Figure 4. Squared correlations between actual and estimated proficiency for conventional and adaptive tests when data were generated from a 1PL model (IRT parameters and stratum membership based on a sample of 50 examinees).

correlation of the traditional CAT over that found with the conventional test, the nine-strata CAT based on a sample size of 100 performed best, recovering 80 - 98% of the traditional CAT increase, followed by the seven-strata CAT (72 - 85%), the five-strata CAT (54 - 65%) and the three-strata CAT (4 - 23%). For the sample size of 50, the nine-strata CAT recovered 80 - 97% of the traditional CAT increase, followed by the seven-strata CAT (69 - 87%), the five-strata CAT (50 - 66%) and the three-strata CAT (0 - 21%). Overall, the results for the one-parameter item pool were similar across sample sizes, and the stratum CATs performed better relative to the traditional CATs than they did in Study 1.

Somewhat different results were found for the three-parameter item pool. For both sample sizes (shown in Figures 5 and 6), the nine-strata CAT and the traditional CAT performed similarly. For the sample of size 100, the nine-strata CAT recovered 91 - 103% of the traditional CAT increase. For the sample size of 50, the nine-strata CAT recovered 100 - 109%—matching or exceeding the squared correlation of the traditional CAT at every test length. Compared to the results of Study 1, the traditional CATs clearly decreased in measurement precision to a much greater degree than did the stratum CATs. It is unclear to what extent the traditional CATs' decreased performance was due poorly estimated item parameters, as opposed to using one-parameter item calibrations with the three-parameter item pool. Both problems, however, would likely be present if a traditional CAT were applied in an actual measurement situation with limited available data.

Figures 5 and 6 clearly indicate that the stratum CATs were substantially more efficient than conventional tests. Measurement precision equivalent to that of conventional tests were obtained with much shorter stratum CATs. In particular, the results suggest that nine-strata CATs can perform as well as IRT-based traditional CATs.

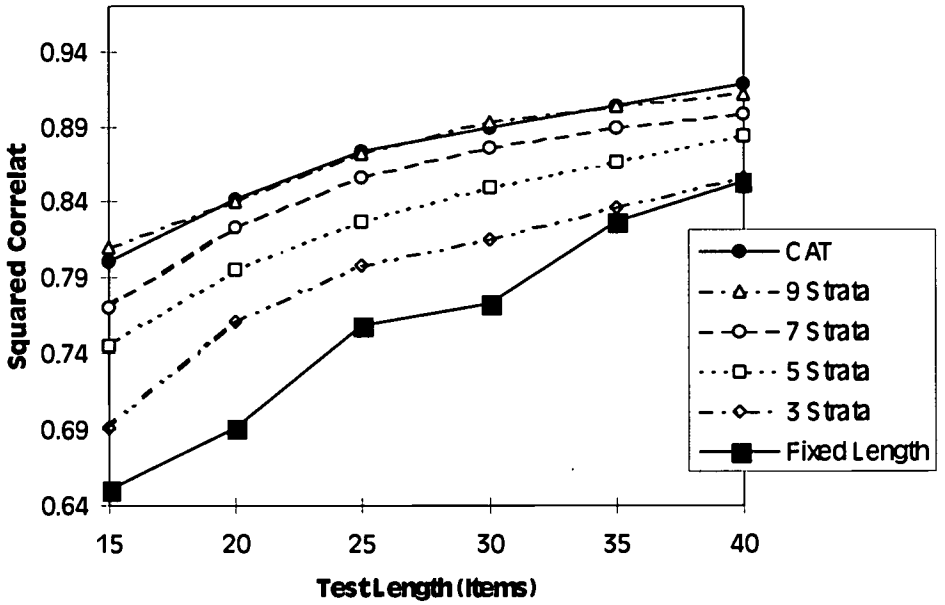

Figure 5. Squared correlations between actual and estimated proficiency for conventional and adaptive tests when data were generated from a 3PL model (IRT parameters and stratum membership based on a sample of 100 examinees).

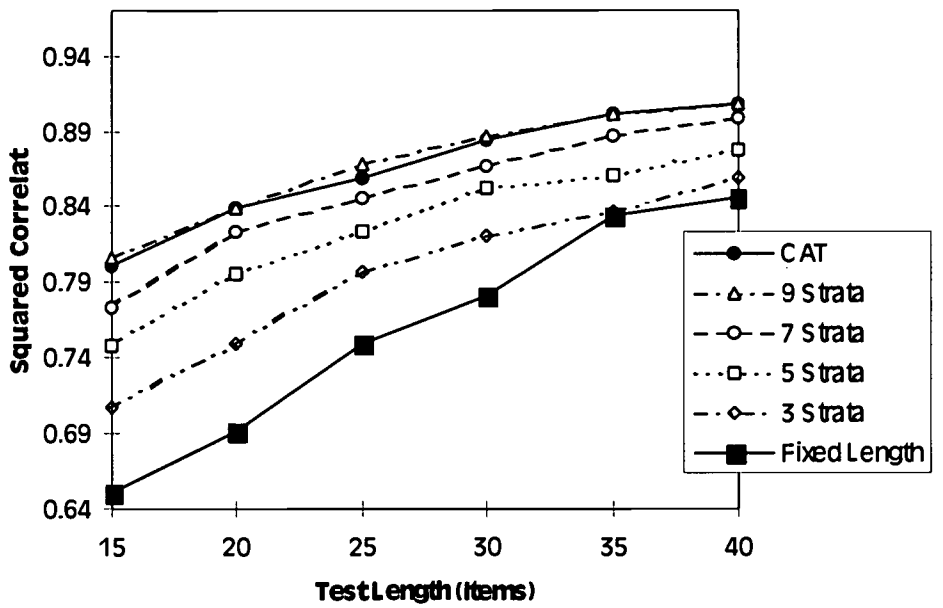

Figure 6. Squared correlations between actual and estimated proficiency for conventional and adaptive tests when data were generated from a 3PL model (IRT parameters and stratum membership based on a sample of 500 examinees).

Additional Analyses

An additional research question concerned the effectiveness of stratum scores as compared to scoring schemes for stradaptive tests that had been previously studied. In other words, does stratum scoring provide greater measurement precision than earlier methods of scoring stradaptive tests? Previous research had indicated that scores based on average IRT difficulty parameter of administered items had been found to have the best psychometric characteristics of those studied by Weiss and his associates (Thompson and Weiss, 1980). Because the purpose of the present study was to investigate non-IRT adaptive testing, it was decided that the most appropriate comparison would be between stratum scoring and the average p-values of administered items (i.e., a score not based on IRT). Such a comparison was performed for a subset of the Study 2 data; specifically, stratum scores and average p-values were compared for the nine-strata CATs based on the three-parameter item pool and a sample of size 100. For all six test lengths, the stratum scores exhibited the higher squared correlation with true proficiency. The size of the differences decreased with test length; for the 15-item test, the squared correlation for stratum scoring was .027 higher, and for the 40-item length it was .015 higher. Hence, this preliminary evidence suggests that stratum scoring represents an improvement over previously studied methods for scoring stradaptive tests.

A second additional analysis focused on the relative measurement precision of the test types at specific levels of examinee proficiency. An exploratory comparison of the root mean square errors (RMSEs) for standardized values of the different test scores was conducted for the same subset of the Study 2 data used in the average p-value analyses. These data (sample size = 100; three-parameter item pool) were chosen for the RMSE analysis because it was for this combination of sample size and item pool that the stratum and traditional CATs had shown the most similar squared correlations with true proficiency.

For each test length, z-scores were calculated for true proficiency, the stratum CAT's stratum scores, the traditional CAT's maximum likelihood scores, and the conventional test's number correct scores. Next, each distribution of true proficiency was divided into quintiles, and RMSEs were computed at each quintile for each of the three test's z-scores. The results are shown in Table 3. The stratum CAT's RMSE values were consistently lower than the other test types for examinees from the first and fifth quintiles. For the third and fourth quintiles, the traditional CAT showed the lowest RMSE at all test lengths, as well at four of the six test lengths for the second quintile. Overall, these results indicate that the stratum CAT estimated proficiency relatively best in the tails of the proficiency distribution, whereas the traditional CAT performed best in the middle of the distribution.

The RMSE values in Table 3 also revealed another interesting difference among the three test types. At all test lengths, the range of the RMSE values across quintiles was markedly smallest for the stratum CAT, indicating that the stratum CATs provided the most consistent measurement across levels of proficiency. This suggests that the stratum CATs in the present analysis best fulfilled a primary advantage of adaptive tests over conventional tests—equiprecision of measurement across the scale of proficiency.

General Discussion

It is useful to begin a discussion of the usefulness of the stratum CAT procedure with a review of criteria that guided its development. Clearly, a stratum CAT in which the strata are formed on the basis of ranked p-values does not require IRT. There were four additional guiding criteria—all of which were met by the stratum CAT procedure. First, stratum scoring appears likely to be easily understood by examinees—to a much higher degree than maximum likelihood or Bayesian methods. Given only the strata of each administered item and whether that item was passed or failed, an examinee could readily verify his or her score. Second, stratum CATs provide effective item exposure control. No item within a stratum would be exposed at a higher rate than others. Such

Table 3

RMSEs for conventional tests, stratum CATs, and traditional CATs, by proficiency quintile

	Test Length					
Quintile	15	20	25	30	35	40
First						
Stratum CAT	.494	.448	.394	.344	.336	.328
Traditional CAT	.556	.488	.434	.410	.406	.361
Conventional Test	.727	.667	.584	.561	.484	.446
Second						
Stratum CAT	.449	.422	.369	.323	.319	.306
Traditional CAT	.463	.405	.367	.337	.302	.271
Conventional Test	.630	.584	.524	.497	.443	.400
Third						
Stratum CAT	.448	.400	.359	.336	.310	.293
Traditional CAT	.426	.379	.324	.319	.278	.263
Conventional Test	.600	.561	.488	.477	.406	.384
Fourth						
Stratum CAT	.395	.361	.323	.297	.282	.258
Traditional CAT	.375	.312	.275	.251	.234	.214
Conventional Test	.574	.529	.465	.452	.388	.358
Fifth						
Stratum CAT	.457	.394	.372	.351	.321	.302
Traditional CAT	.458	.424	.379	.360	.333	.308
Conventional Test	.566	.557	.466	.457	.391	.353
Range of RMSE Across Quintiles						
Stratum CAT	.099	.087	.071	.054	.054	.070
Traditional CAT	.181	.176	.159	.159	.172	.147
Conventional Test	.161	.138	.119	.109	.096	.093

Note. Results reported for nine-strata CATs, with stratum membership estimated from the responses of 100 hypothetical examinees to the three-parameter item pool. These responses were also used to estimate the difficulty parameters used by the traditional CATs.

uniformity of exposure should extend the useful life of an item pool. Third, proficiency estimates from the stratum CATs with at least five strata were found to consistently outperform conventional tests by a substantial amount. In addition, the nine-strata CATs in Study 2 were found to perform equal to or better than traditional CATs. The fourth criterion stated that the amount of data needed to establish a useful item pool would be less than 200 examinees per item. The results of Study 2 indicated that item pools based on as few as 50 examinees per item could provide sufficient information to support effective adaptive testing. Thus, the stratum CAT procedure appears to satisfy all of the guiding criteria.

Based on the findings presented here, the stratum CAT procedure appears to be a promising adaptive testing method in measurement contexts for which item data are in limited supply. Because the two simulation studies represent the initial investigations of stratum CATs, however, their results should be interpreted with some caution. The generalizability of the present results to different item pools (either simulated or real) or different distributions of proficiency is not yet clear. Future research should provide important information regarding this and other relevant issues concerning the practical utility of the stratum CAT procedure.

Additional Measurement Issues

If a stratum CAT were to be used in practice, there are a number of accompanying practical issues that would need to be addressed. Several of these issues are discussed below.

Reliability. Vale and Weiss (1975b) discussed a method for estimating the reliability of an examinee's stradaptive test score that could be applied to stratum CATs. Given the variances and covariances of the items administered to an examinee, which could be estimated from the calibration sample used to calculate the p-values, coefficient alpha could readily be computed for the test of any examinee. The usefulness of this estimation method should be studied in future research.

Variable-Length Tests. A strategy commonly used with traditional CATs to help insure that the proficiency levels of different examinees will be measured with equivalent precision is to permit test length to vary across examinees. In such tests, items are administered to a given examinee until the standard error of proficiency estimation has decreased to some threshold value. In stratum CATs, the reliability of an examinee's test could be used as a criterion for terminating a test.

Content Balancing. A common concern in adaptive testing is whether a CAT maintains in the same proportions the content specifications used in a conventional test over the same content domain. The stratum CAT procedure could readily be modified to constrain item selection as to maintain balanced content. That is, instead of randomly selecting an item from a stratum, a slightly more complex algorithm could be used that randomly chooses an item from a target content area (which could rotate across items) within that stratum. Such a strategy would be very similar to that proposed by Kingsbury and Zara (1989).

Summary

The present research was motivated by a desire to make adaptive testing more accessible to measurement practitioners who lack the data, resources, or expertise to apply IRT-based methods in establishing an item pool and administering/scoring a traditional CAT. Although such CATs have become commonplace in large-scale testing programs, they have thus far had limited impact elsewhere.

In this investigation, the development of an adaptive testing procedure was described that could be practically useful in situations where limited examinee data are available. The resultant procedure—which was termed a stratum CAT—combines many of the aspects of stradaptive testing (Weiss, 1973) with a new scoring method (stratum scoring). The key feature of the stratum CAT is that it does not use data-intensive IRT methods for either item selection or proficiency estimation. Its scores would be easy for examinees to understand, it provides effective item exposure control, it can yield

proficiency estimates superior to those from conventional tests, and relatively little data per item would be needed to calibrate an item pool. The relatively simple principles of the stratum CAT method should be more comprehensible to practitioners than those of traditional CATs. Once that suitable, easy-to-use testing software is developed, stratum CATs promise to substantially broaden the impact of adaptive testing on measurement practice.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, DC: American Council on Education.

Bejar, I. I., Weiss, D. J., & Gialluca, K. A. (1977). An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Green, B. F. (1997, March). Alternate methods for scoring computer-based adaptive tests. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago. IL.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 147-200). New York: Macmillan.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 359-375.

Mislevy, R., & Bock, R. D. (1982). BILOG: Maximum likelihood item analysis and test scoring with logistic models. Mooresville, IN: Scientific Software.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. Journal of Educational and Behavioral Statistics, 21, 365-89.

Thompson, J. G., & Weiss, D. J. (1980). Criterion-related validity of adaptive testing strategies (Research Report 80-3). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Vale, C. D., & Weiss, D. J. (1975a). A study of computer-administered stradaptive ability testing (Research Report 75-4). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Vale, C. D., & Weiss, D. J. (1975b). A simulation study of stradaptive ability testing (Research Report 75-6). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Waters, B. K. (1977). An empirical investigation of the stratified adaptive computerized testing model. Applied Psychological Measurement, 1, 141-152.

Weiss, D. J. (1973). The stratified adaptive computerized ability test (Research Report 73-3). Minneapolis: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

Footnote

[1] A negative consequence of adopting proportional stratum scoring weights is that the scoring becomes more complicated to explain to examinees—which runs counter to the goal of establishing a scoring procedure that examinees can readily understand. One should therefore be cautious regarding use of proportional weights unless their increased precision outweighs the potential for examinee confusion.

Author Note

I wish to thank Christine DeMars, Gage Kingsbury, and Vicente Ponsoda for their insightful comments on an earlier version of this paper. In addition, I thank Rita Gentile for her kind assistance and advice in the running of the simulation programs.

TM029662

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Comparison of Stratum Scored and Maximum-Likelihood CATs

Author(s): Steven L. Wise

Corporate Source:

Publication Date: April, 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Steven L. Wise

Printed Name/Position/Title: Steven L. Wise/Professor

Organization/Address:
James Madison University
Center for Assessment & Research Studies
ISC 9001
Harrisonburg, VA 22807

Telephone: (540) 568-7022

Fax: (540) 568-7878

E-mail Address: wisesl@jmu.edu

Date: 4/6/99