DOCUMENT RESUME

ED 380 498 TM 022 861

AUTHOR	Stocking, Martha L.
TITLE	An Alternative Method for Scoring Adaptive Tests. Research Report RR-94-48.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	Oct 94
NOTE	40p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Difficulty Level; *Equated Scores; *Item Response Theory; Psychometrics; *Scoring; Test Interpretation; Test Use
IDENTIFIERS	*Number Right Scoring

ABSTRACT

Modern applications of computerized adaptive testing (CAT) are typically grounded in item response theory (IRT; Lord, 1980). While the IRT foundations of adaptive testing provide a number of approaches to adaptive test scoring that may seem natural and efficient to psychometricians, these approaches may be more demanding for test takers, test score users, interested regulatory institutions, and so forth, to comprehend. An alternative method, based on more familiar equated number-correct scores and identical to that used to score and equate many conventional tests, is explored and compared with one that relies more directly on IRT. The conclusion is reached that scoring adaptive tests using the familiar number-correct score, accompanied by the necessary equating to adjust for the intentional differences in adaptive test difficulty, is a statistically viable, although slightly less efficient, method of adaptive test scoring. To enhance the prospects for enlightened public debate about adaptive testing, it may be preferable to use this more familiar approach. Public attention would then likely be focused on issues more central to adaptive testing, namely the adaptive nature of the test. (Contains 35 references, 2 tables, and 3 figures.) (Author)

RR-94-48

RESEARCH REPORT

AN ALTERNATIVE METHOD FOR SCORING ADAPTIVE TESTS

Martha L. Stocking

Educational Testing Service
Princeton, New Jersey
October 1994

AN ALTERNATIVE METHOD FOR SCORING ADAPTIVE TESTS[1]

Martha L. Stocking
Educational Testing Service
Princeton, New Jersey 08541

Phone: (609) 734-5985
Internet: mstocking@rosedale.org

[1] This research was supported in part by the Program Research Planning Council of Educational Testing Service.

AN ALTERNATIVE METHOD FOR SCORING ADAPTIVE TESTS

Abstract

Modern applications of computerized adaptive testing (CAT) are typically grounded in Item Response Theory (IRT; Lord, 1980). While the IRT foundations of adaptive testing provide a number of approaches to adaptive test scoring that may seem natural and efficient to psychometricians, these approaches may be more demanding for test-takers, test score users, interested regulatory institutions, and so forth, to comprehend. An alternative method, based on more familiar equated number-correct scores and identical to that used to score and equate many conventional tests, is explored and compared with one that relies more directly on IRT. The conclusion is reached that scoring adaptive tests using the familiar number-correct score, accompanied by the necessary equating to adjust for the intentional differences in adaptive test difficulty, is a statistically viable, although slightly less efficient, method of adaptive test scoring. To enhance the prospects for enlightened public debate about adaptive testing, it may be preferable to use this more familiar approach. Public attention would then likely be focussed on issues more central to adaptive testing, namely the adaptive nature of the test.

Key words: adaptive testing, adaptive test scores, IRT scoring, IRT equating, test scores.

AN ALTERNATIVE METHOD FOR SCORING ADAPTIVE TESTS

Introduction

Recent advances in psychometrics and computing technology have led to the development of a testing paradigm that is very different from linear paper-and-pencil testing -- computerized adaptive testing (CAT; see, for example, Eignor, Way, Stocking, & Steffen, 1993; Lord, 1977; Schaeffer, Steffen & Golub-Smith, 1993; Stocking & Swanson, 1993; Wainer, Dorans, Flaugher, Green, & Mislevy, 1990; and Weiss, 1982). As interest in large-scale implementation of modern adaptive testing has increased, particularly for high-stakes testing programs (Jacobson, 1993), test sponsors are, of course, obliged to ensure that professional standards are met for this new testing paradigm. Perhaps less obvious, but equally important, test sponsors are also obliged to ensure that the understanding of this testing paradigm by test-takers, test score users, legislative and/or regulatory institutions and other interested parties is as complete as their current understanding of conventional testing. This paper discusses one aspect of adaptive testing for which such understanding seems essential -- that of adaptive test scoring.

Test-takers are accustomed to a system of scoring in which a 'raw' test score is based on the number of questions answered correctly -- either a simple sum of the number of right answers as in number-correct scoring, or a sum of the number of right answers with a penalty for the number of wrong answers, as in formula scoring. For large scale standardized paper-and-pencil tests, these raw scores are converted to some arbitrary metric for score reporting purposes. This transformation, which involves the statistical process of test score equating, is also reasonably familiar to test-takers who

are provided with tables that give the correspondence between raw scores and reported (scaled) scores. Test score equating is performed in order to statistically adjust raw scores for inevitable and unintentional form-to-form variation in test difficulty, so that test-takers are neither advantaged or disadvantaged by the actual test form they were administered.

Proposed legislation to regulate the testing industry reflects and reinforces this familiar concept of test scoring. For example, bill S. 8063-A was passed by the New York State Senate in June, 1994, to extend existing regulations to include computer-based testing. (This bill has not yet been introduced or passed by the New York State Assembly so it is not yet a law or regulation, although legislation introduced during the coming year is anticipated to be similar.) In this proposed legislation, testing companies are required, among other things, to release the rules used for deriving test scores, and to allow candidates to derive their test scores using these rules, as they currently do for conventional tests.

The challenge of making adaptive test scoring understandable to test-takers may be difficult to meet because modern adaptive testing is grounded in Item Response Theory (IRT; Lord, 1980), thus making the psychometric underpinnings of adaptive testing more difficult to explain. It is hard to envision developing a system of rules that would allow "candidates to derive their test scores using these rules" in this context. This paper explores whether or not it is possible to score adaptive tests in a way that is more familiar to interested parties without undue sacrifices of the other efficiencies gained from adaptive testing. We focus in particular on the scoring of adaptive tests with the number of correct answers, which is then

equated to some arbitrary reporting metric, as is currently done for conventional standardized tests.

The goal of equated number-correct scoring of adaptive tests is made more difficult by two fundamental aspects of adaptive testing. The first challenge comes from the fact that in adaptive testing every test-taker can, in theory, be administered a completely different test and equating is required for each of them. The second challenge comes from the fact that, with perfect item pools and an item selection algorithm that considers only item information (Lord, 1980, equation 5-9), every test-taker can be expected to answer correctly about 60% of the items presented to him/her. This is because items provide the most information about test-taker proficiency if the probability of a correct answer is about half way between chance level and 1, and a typical chance level for the kinds of items characteristically seen in adaptive tests is around .20. In this context, number-correct scoring seems impossible.

The IRT foundations of adaptive testing provide many different inherent scoring approaches which seem natural and efficient for psychometricians. The nature and characteristics of two IRT-based approaches to adaptive test scoring are discussed in this paper. In addition, an equated number-correct approach to adaptive test scoring with less direct reliance on IRT is presented. This number-correct approach is compared to one of the more common approaches for six high-stakes adaptive tests in monte-carlo simulations.

Adaptive Testing With the Weighted Deviations Model

The psychometrics underlying the six tests studied in this paper are based on the three parameter logistic (3PL) IRT model (Lord, 1980). Items in

the item pools are calibrated and placed on the same metric using the computer program LOGIST (Wingersky, 1983) or BILOG (Mislevy & Bock, 1983). The item selection in the adaptive test employs the methodology of the weighted deviations model (WDM) (Stocking & Swanson, 1993; Swanson & Stocking, 1993) with the extended Sympson and Hetter (1985) exposure control methodology (Stocking, 1992) to increase item security. (For details of the test design process, see Eignor et al., 1993; and O'Neill, Folk, & Li, 1993.) For these tests, the goal of the test design process is to have (fixed length) adaptive test scores that are interchangeable with those from companion linear paper-and-pencil tests both in terms of their psychometric properties and the constructs being measured. This is necessary since it is envisioned that both modes of testing must co-exist for some indefinite period of time into the future.

In the WDM approach to adaptive testing, item properties or features are taken into account along with statistical properties in the selection of items. This is to insure that each adaptive test produced from the pool matches a set of test specifications and is therefore as parallel as possible to any other test in terms of content and types of items, while being tailored to an individual examinee in terms of difficulty. The WDM approach also allows specification of overlapping items that may not be administered in the same adaptive test. In addition, it is possible to restrict item selection to blocks of items, either because they are associated with a common stimulus or common directions or any other feature that test specialists deem important.

In summary, in the weighted deviations model, the next item selected for administration is the item that simultaneously

1) is as informative an item as possible at a test-taker's estimated ability level, while

2) contributing as much as possible to the satisfaction of all other constraints in addition to the constraints on item information.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if the previous item was in a block), starts a new block, or is in no block.

The Sympson and Hetter exposure control methodology further restricts item selection by determining if the selected item is likely to be overexposed if administered, based on exposure control parameters developed over a series of simulations with a (simulated) typical group of test-takers. If so, this methodology forces the administration of an item that has been administered less frequently.

Two Common IRT-Based Scoring Methods

Maximum Likelihood Estimate of Ability

In IRT, test-takers are characterized by a parameter, denoted by θ, that represents the trait, ability, proficiency, or skill that underlies responses to test questions. This conceptualization of a latent or unobservable trait is more fundamental than that of a test score, which depends upon the characteristics of items that compose a particular test and is therefore test specific. If we knew a test-taker's θ and the item characteristics or parameters for items in a test, we could estimate what a test-taker's number-correct score would be on that test in a probabilistic fashion.

In practice, of course, we cannot know a test-taker's θ, but we estimate it and this estimate is usually represented by $\hat{\theta}$. Many methods exist for obtaining such an estimate from available response data and the characteristics of items (see, for example, Lord, 1980, p58; Mislevy, 1986; Wainer et al., 1990, pp72-79). Much of the early theoretical research into adaptive testing paradigms used $\hat{\theta}$ as an adaptive test score (for example, Killcross, 1976; Lord, 1977; Mcbride & Martin, 1983; Owen, 1975; Vale, 1981; Weiss, 1974). A number of recent operational implementations use $\hat{\theta}$ directly or indirectly as a test score. For example, the adaptive version of the Armed Services Vocational Aptitude Battery reports test scores derived from equating $\hat{\theta}$ to scores on a paper-and-pencil test (D. Segall, personal communication, December 10, 1993). Also, NCLEX/CAT, a certification and licensing program for nurses, uses $\hat{\theta}$ to determine pass/fail status on adaptive tests (W. Way, personal communication, December 15, 1993).

Assuming that item parameters are known from pretesting the items, the maximum likelihood estimate of θ is the solution to the likelihood equation (Lord, 1980, equation 4-31)

$$\sum_{i=1}^{n} (u_i - P_i(\theta)) \frac{(P_i(\theta))'}{P_i(\theta)Q_i(\theta)} = 0. \tag{1}$$

In this equation, u_i is the scored response to an item (0 if incorrect and 1 if correct), $P_i(\theta)$ is the item response function (in this case the three parameter logistic item response function), $(P_i(\theta))'$ is the derivative of the item response function with respect to θ, $Q_i(\theta)$ is equal to $(1 - P_i(\theta))$, and n is the number of items that have been administered. Equations such as these are typically solved by iterative numerical methods (Wingersky, 1983).

Using $\hat{\theta}$ as an adaptive test score has a number of concomitant facets or features. Since IRT underlies adaptive testing, and $\hat{\theta}$ as an estimate of θ is a fundamental feature of IRT, its use as a test score conforms to the IRT model. Furthermore, $\hat{\theta}$ estimates the proficiency that underlies responses to items, and thus estimates a more fundamental aspect of test-taker behavior than a test score based on a particular collection of items. However, it could be argued that since test-takers are accustomed to test scores based on number-correct or formula scores, the concept of estimating ability rather than a test score is unfamiliar and of less utility to them. In addition, the process of obtaining a solution to equation (1) is not very intuitive and therefore would be difficult for test-takers to understand when compared to the simple process of adding up the number of correct answers.

The maximum likelihood estimate of ability, $\hat{\theta}$, is a fallible test score just like any other test score such as number-correct, and as such, it has certain asymptotic statistical properties that make it attractive if the number of items upon which it is based is large. For example, given item responses to a large set of items with known parameters, the maximum likelihood estimate of θ is a consistent estimator of ability (Lord, 1983), and it is also the most informative (Lord, 1980, Theorem 5.3.2). That is, there is no other estimator of ability that has a smaller sampling error. However, adaptive tests are frequently designed to be as short as possible. It may be questionable whether these desirable properties hold in adaptive testing since the number of items is small (Lord, 1980, p59).

A final feature or property of $\hat{\theta}$ as a test score comes from viewing this estimate as a weighted sum of item scores. Number-correct scores, of course, can be viewed as a weighted sum of item scores in which the weights are all

equal (usually 1) and the items are scored as 1 (correct) or 0 (incorrect). Likewise it is possible to express $\hat{\theta}$ as a function of weighted item scores. Regardless of the form of the item response function, if the item parameters are known from pretesting then $\hat{\theta}$ is obtained by solving (Lord, 1980, equation 5-19)

$$\sum_{i=1}^{n} \frac{(P_i(\theta))'}{Q_i(\theta)} = \sum_{i=1}^{n} w_i(\theta) u_i, \tag{2}$$

an alternative form of equation (1) in which the weights are defined as (Lord, 1980, equation 5-15):

$$w_i(\theta) = \frac{(P_i(\theta))'}{P_i(\theta) Q_i(\theta)}. \tag{3}$$

Under some IRT models, such as the one-parameter and two-parameter logistic models, the w_i do not depend on θ. Under others, and in particular the 3PL, they do. If we substitute $\hat{\theta}$ for θ under the 3PL model, we discover that the weights have the following properties (Lord, 1980, p75)

1) At high ability levels, the item weights become independent of ability. Although different items have different weights, since the weights are proportional to the item discrimination at these high ability levels, all high ability test-takers receive essentially the same weight for any particular item.
2) The weights for difficult items decrease as ability decreases.
3) At low levels of ability, the weights for difficult items are virtually zero.

The fact that $\hat{\theta}$ can be viewed as (a function of) a weighted sum of item scores and that these weights are functions of ability lead to desirable

features in the psychometric context. Moreover, if we had a perfect adaptive test item pool, and were able to administer items of exactly the right difficulty for all test-takers, the weights for all items would be independent of ability and depend only the properties of the items. However, in a practical context, with less than perfect pools and item selection algorithms that take into account nonstatistical features of items, the nature of these weights is problematic. It may be hard to explain to two test-takers who receive exactly the same (say, for example, difficult) item in an adaptive test, and who make exactly the same response to that item, that the lower ability test-taker receives 'less credit' for the response than the higher ability test-taker.

Estimated number-correct true score

As previously discussed, if we knew a test-taker's θ and the item characteristics or parameters for a set of items, we could estimate the test-taker's number-correct true score, ξ, on that set of items. This is accomplished using the test characteristic curve (Lord, 1980, equation 4-5):

$$\xi = \sum_{i=1}^{n} P_i(\theta). \tag{4}$$

In practice, we substitute $\hat{\theta}$ and estimates of item parameters for a set of items to obtain an estimated number-correct true score, $\hat{\xi}$, (or a formula-score equivalent) on the set of items. The use of $\hat{\xi}$ as an adaptive test score has become increasingly common because it is viewed as a mechanism for overcoming some of the aforementioned disadvantages $\hat{\theta}$ (Dorans, 1990; Eignor, et al., 1993; Schaeffer, et al., 1993; Stocking, 1987).

The set of items used to compute $\hat{\xi}$ (called the reference set or reference test) can be arbitrary, as long as the estimated item parameters and

the $\hat{\theta}$ are all on the same IRT metric. In some programs of adaptive testing, the reference set is the entire available item pool (Ward. 1988). In this case the interpretation is that $\hat{\xi}$ is the score a test-taker with estimated proficiency $\hat{\theta}$ would have obtained if the entire item pool had been administered to him/her as a conventional test and there were no fatigue or speededness effects. In other programs of adaptive testing, the reference set is composed of items comprising some intact (linear) test form that has already been equated to the score reporting scale (Eignor et al.. 1993; O'Neill et al., 1993). In this case the interpretation is that $\hat{\xi}$ is the score a test-taker with estimated proficiency $\hat{\theta}$ would have obtained if they had been administered the intact test form as a conventional test.

There are a number of concomitant features of using $\hat{\xi}$ as an adaptive test score. Test-takers are more familiar with a test score than with an estimated proficiency. The (raw) score reporting metric ranges from chance level to a perfect score on the reference set of items and is similar to raw score metrics typically encountered by test-takers. However, although these two aspects may provide some advantage to using $\hat{\xi}$, in terms of explaining test scoring to test-takers, they do not overcome the fact that $\hat{\xi}$ is still based on the unfamiliar concepts and theory required to obtain $\hat{\theta}$ in the first place. In fact, one could argue that the use of $\hat{\xi}$ might actually be more confusing since it involves additional computations to obtain.

A second feature in using $\hat{\xi}$ obtained on a reference test as an adaptive test score is that it can be transformed to the scaled score reporting metric using an equating transformation previously developed for the reference test. The mathematics are trivial and this easily ties adaptive test scores to

conventional test scores that may have been in existence for some period of time and may continue to exist after the introduction of adaptive testing.

A final property in using $\hat{\xi}$ as an adaptive test score concerns issues of test score use. If a conventional test were scored with $\hat{\theta}$, the standard errors for extreme values of θ would be large compared to the standard errors for θs near the level aimed at by the test (Lord, 1983). That is, extreme values of $\hat{\theta}$ have large standard errors on the θ scale because the θs are not well estimated. However, these large standard errors do not imply that test takers with very high or very low abilities are likely to score very differently on a test similar to the one under consideration; almost all of the items in such tests are too hard or too easy for such test-takers. Thus the size of the standard error on the θ scale may not adequately reflect the importance of that standard error to test score users.

If it were possible to obtain a perfect adaptive test item pool, with as many items as desired from all content areas at all ability levels, these differences in the sizes of standard errors would disappear. However, with realistic item pools, these differences in standard errors persist. The discrepancy between the size of the standard error and its importance for test score users can be reduced by using $\hat{\xi}$ as the adaptive test score (Lord, 1983).

A Number-Correct Scoring Method

The use of either $\hat{\theta}$ or $\hat{\xi}$ as adaptive test scores does not overcome, in any obvious fashion, the challenge of making adaptive test scoring easily understandable to test-takers. Such a challenge could possibly be met by an approach that uses concepts already familiar to test-takers, such as number-correct or formula scoring. (For the remainder of this paper, our focus will

be on a number-correct approach, with the understanding that formula score analogues exist.)

The fundamental concept underlying a number-correct approach to adaptive test scoring is IRT true score equating (Lord, 1980, chapter 13). Test score equating is used in many large scale conventional testing programs to adjust for unintentional differences in difficulty across different editions of the same conventional test so that test scores are reported and may be compared on a single score reporting metric. IRT true score equating is an accepted equating methodology (see, for example, Bejar & Wingersky, 1982; Cook & Eignor, 1983; Cook & Petersen, 1987; Cook, Petersen & Stocking, 1983; Eignor, Cook & Stocking, 1990; Lawrence & Dorans, 1990). In applying this approach to adaptive testing, the number-correct score on each adaptive test is individually equated to a score on the reference test.

In IRT equating, observed scores on two different tests that measure the same construct are considered to be equated if they correspond to the same value of θ, as determined by the test characteristic curves of the two tests. Given two tests, both measuring θ, their number-correct true scores, ξ and η, are related to θ by the parametric equations (Lord, 1980, equation 13-12)

$$\xi = \sum_{i=1}^{n} P_i(\theta), \qquad \eta = \sum_{j=1}^{m} P_j(\theta). \tag{5}$$

In practice, for a number-correct score on an adaptive test, represented by the sum of item scores, we solve

$$\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} P_i(\theta) \tag{6}$$

or

$$\sum_{i=1}^{n} (u_i - P_i(\theta)) = 0 \tag{7}$$

for $\hat{\theta}$ using estimated item parameters for the items in the adaptive test and iterative numerical methods. Then, using this value of $\hat{\theta}$ and the estimated

item parameters for the reference test, we find the corresponding estimated number right true score on the reference test

$$\hat{\xi} = \sum_{j=1}^{m} P_j(\hat{\theta}). \tag{8}$$

If the reference test scores are to be transformed to a different score reporting metric, as is usually the case, the transformation is applied to the $\hat{\xi}$ from equation (8).

Equation (7) is similar to the 3PL likelihood equation for estimating θ, equation (1), except the fractional term is now taken to be 1. Indeed, if we used the one parameter logistic (1PL) item response function model, the likelihood equation for estimating θ would have a form identical to equation (7). However, equation (7) is not the 1PL likelihood equation for estimating θ because the $P_i(\theta)$ are 3PL item response functions, not 1PL item response functions. Because we are ignoring information available for estimating θ in the 3PL model, equation (7) can be considered a 'reduced information' approach to estimating $\hat{\theta}$ for the 3PL model.

From the perspective of the test-taker, this approach to scoring adaptive tests looks identical to that for conventional tests. The raw score on an adaptive test is simply the number of correct answers, and all items count the same amount towards this score. The reported score is the result of an equating that adjusts for (now intentional) form-to-form variation in test difficulty, just as equating transformations do for conventional testing. The only practical difference is that each adaptive test is separately equated to the reference test, although this may not be of much interest to test-takers. If test-takers understand how conventional test scores are derived, they will also understand how adaptive test scores are derived, since the mechanisms are the same.

From the perspective of the psychometrician, this approach is very different from using $\hat{\theta}$ or $\hat{\xi}$ as adaptive test scores. Some of the information available about the items administered in an adaptive test and the responses to these items when using the 3PL model to estimate θ is ignored. We use the information about the items administered in the adaptive test to construct the test characteristic curves, but we do not exploit this information as fully as we could when we do not use the association between specific item parameters and specific item responses. A major theoretical issue arises as to the impact of the loss of information on the psychometric properties of an adaptive test. A second, more practical issue also arises. We are accustomed to the theoretical notion that if we had a perfect item pool and item selection algorithm, everyone would obtain the same (raw) number-correct score on the adaptive test administered to him or her. It remains to be shown whether, in practice, IRT equating is sensitive enough to provide appropriately different scaled scores when all adaptive test raw scores are the same or nearly the same. The monte-carlo experiment described in subsequent sections is designed to explore these issues.

The Monte Carlo Experiment

The Tests

Results from test design simulations were obtained for adaptive tests in six different areas: Verbal Reasoning, Quantitative Reasoning, Analytical Reasoning, Mathematics, Reading, and Writing. The test design simulations for the first three measures are described in Eignor et al. (1993); those for the last three measures are described in O'Neill et al., (1993).

All six tests used the WDM adaptive testing paradigm described earlier. The item parameters for the items in the Verbal, Quantitative, and Analytical Reasoning item pools were estimated from large samples of test-takers (2000+) using the 3PL item response model and the computer program LOGIST (Wingersky, 1983); those for the Mathematics, Reading, and Writing item pools were estimated from smaller samples of test-takers (500+) using the 3PL model and the computer program BILOG (Mislevy & Bock, 1983). All six measures used $\hat{\xi}$ on a corresponding reference test from equation (4) as an adaptive test score. The test design simulations were conducted to establish the test lengths, exposure rates, constraint weights, and item pool sizes required to meet minimum desirable levels of reliability (computed using Green, Bock, Humphreys, Linn & Reckase (1984), equation 6) and desirable conditional standard error of measurement (CSEM) curves. The simulations were conducted with reference to estimated distributions of true ability for the intended population, computed by the method of Mislevy (1984).

Table 1 contains specific information about each measure. The number of items in the (fixed length) adaptive test and the reference test used for scoring purposes are shown in columns 1 and 3. The adaptive tests are approximately 1/2 to 3/4 the length of the reference tests. The number of elements in the pool shown in column 2 reflects the number of stimuli such as passages or graphs that are associated with multiple items as well as the number of discrete items. This is because some of the constraints on item selection, shown in column 4, apply to stimuli rather than items themselves. The fifth column shows the desired expected maximum exposure rate for items and stimuli in the item pool required by the extended Sympson and Hetter procedure. A value of .20 means that no more than 20% of a typical population

should see any individual item or stimulus. Attained values for this maximum exposure rate for the final iteration of the simulations are typically slightly higher than the pre-specified expected maxima. The final column in Table 1 shows the number of simulated examinees (simulees) used for each adaptive test design simulation.

Insert Table 1 about here

The Method

The results from the final test design simulations consist of

1) Correct and incorrect responses generated for each simulated examinee for items in adaptive tests constructed as previously described. These responses were generated using the estimated item parameters from items in the appropriate pool, and a value of true ability (θ) for each simulee.

2) The adaptive test score, i. e., $\hat{\xi}$, on the appropriate reference test computed from equations (1) and (4) for each simulee.

The adaptive tests (1300 for the Verbal measure, 1650 for the Quantitative measure, and so forth) are then rescored number-correct, and the corresponding equated scores on the reference test are computed using equations (7) and (8). No further simulations are required.

Three aspects of equated number-correct scoring are evaluated and compared to scoring adaptive tests with $\hat{\xi}$. First, the total mean squared error for each scoring method is decomposed into the variance and the residual (bias) and the results compared. Second, the sensitivity of the equating required for number-correct scoring in a context in which all simulees are

expected (in theory) to obtain the same number-correct score is examined for reasonableness. Finally both scoring methods are compared to criterion values of reliability and conditional standard errors of measurement used to establish the original test design when using $\hat{\xi}$ as the adaptive test score.

The Results

Residuals of adaptive test scores from true values are shown for each test in Figure 1. Each test is represented by a small plot, with the horizontal axis the appropriate true score scale for the generating values used in the simulation. The estimated distribution of true ability is represented by the histogram of proportional frequencies (that is, the proportional frequencies sum to unity) on each plot. Residuals (conditional on true ability) from true values for scoring based on the full information (solid line, equation (1) and (4)) and the reduced information (dashed line, equations (7) and (8)) are to be read from the left vertical scale. This scale is constant for all panels in the figure. Proportional frequencies of the estimated distribution of true ability are to be read from the right vertical scale. These scales are not the same throughout the figure, but remain constant within each row. That is, both the Verbal and the Quantitative measure have the same right-hand vertical scale, as do the Analytical and Mathematics measures, and the Reading and Writing measures.

Insert Figure 1 about here

For each measure, scoring adaptive tests as number-correct tends to produce residual values that lie below those obtained from scoring adaptive tests $\hat{\xi}$, although the two curves 'track' each other closely. This tracking is

not surprising given that we have simply rescored the same item responses. Columns 5 and 6 of Table 2 display the residuals (bias) for a typical distribution of ability for the two scoring methods for each test, where the weights are the appropriate estimated distribution of true ability.

Insert Table 2 about here

For four of the six tests, the absolute value of the residual using number-correct scoring is smaller than that for scoring adaptive tests with $\hat{\xi}$. This means that for these four tests, number-correct scoring recovered true values better (had less bias) than $\hat{\xi}$. The corresponding (weighted) variances are shown in columns 7 and 8 of Table 2. (The variance and the square of the bias sum to the total mean squared error for each row in the table.) For all tests, the variance using number-correct scoring is larger than that obtained when using $\hat{\xi}$. Thus, while number-correct scoring may be less biased for some (but not all) adaptive tests, it is more variable.

Figure 2 shows the mean number-correct scores, conditional on ability, for the adaptive test (solid line) and corresponding equated scores (dashed line) for each test, both converted to proportions correct so that they can be compared. Values for these curves are to be read from the left vertical scale. For each test, the horizontal axis and right-hand vertical axis have the same meaning as in Figure 1, as does the histogram representing the estimated distribution of true ability. The (raw) percent correct scores, averaged over the typical distributions of ability are given in column 9 of Table 2 for each test. These tend to be around 60%, as we would expect for adaptive testing with these types of items.

Insert Figure 2 about here

The plots of the mean conditional raw scores form an S-shaped curve that tends to be flatter throughout the middle of the ability range and steeper at both extremes. As mentioned earlier, with perfect item pools and an item selection algorithm based solely on the information property of items, this curve would be a horizontal line indicating that all simulees received the same number-correct (or proportion-correct if the adaptive test is fixed length) score on their adaptive tests. The fact that this curve is not a horizontal line can be attributed both to deficiencies in the item pool and to the fact that WDM adaptive testing algorithm considers more than statistical properties of items in item selection.

Most of these conditional curves have a relatively flat region in the middle where the proportions correct tend to be around .60, indicating that the item pool was more nearly adequate in terms of items appropriate for these ability levels. For lower ability levels, the curves decline below .60, indicating that lower ability simulees get lower (raw) number-correct scores. The item pools do not contain items that are sufficiently easy for these simulees. The reverse is true at higher levels of ability, where the curve increases above .60, indicating that higher bility simulees get higher number-correct scores. This is because the pool does not contain a sufficient number of hard items appropriate for these simulees.

The dashed line represents the raw number-correct score after it has been equated to the reference test (and transformed to the proportion correct metric for ease of comparison). As appropriate, lower ability simulees tend

to have equated scores lower than their raw scores, a consequence of the fact that these simulees were administered easy items. Higher ability simulees tend to have equated scores that are higher than their raw scores, a consequence of the harder items administered to these simulees. The IRT equating seems to be functioning as intended.

An interesting artifact appears in Figure 2, most noticeable on the plot for the Verbal measure. For very low ability simulees, equated scores are higher than raw scores (as they are for simulees of middle to high ability), in contrast to simulees of slightly higher ability where the reverse is true. This was investigated extensively and was found to be a consequence of adaptive testing per se and known sampling correlations in the estimation of item parameters. Very low ability simulees are responding at chance level. The chance level on the reference test is higher than the chance level on the adaptive tests administered to these simulees. This occurs because in the adaptive test, we are intentionally selecting the easiest informative items, which will be those easy items with the lowest guessing parameters. In addition, there is a strong positive sampling correlation between estimates of item difficulty and estimates of the pseudo-guessing parameter (Wingersky and Lord, 1984). Thus if item difficulty is underestimated, as it undoubtedly is for some of these easy items, the pseudo-guessing parameter will tend to be underestimated also. The chance level on the adaptive test, then, is lower than one might expect because of these two factors. This level is equated to chance level on the reference test, which has a more typical value because it has the properties that we are accustomed to observing for conventional tests.

Figure 3 and the first four columns of Table 2 compare number-correct scoring of adaptive tests with the results obtained from scoring adaptive

tests with $\hat{\xi}$, in terms of reliabilities and CSEMs. The reliabilities for the reference test scored number-correct and $\hat{\xi}$ are computed from the appropriately weighted versions of equations 4-3 and 6-5 in Lord (1980). Scoring adaptive tests with number-correct reduces the reliability slightly over scoring adaptive tests with $\hat{\xi}$. However, only for the Analytical measure is the reliability reduced below that of the reference test scored as number-correct, the current scoring method. If the criterion reliability is considered to be the reliability of the current test, number-correct scoring of adaptive tests performs adequately.

Insert Figure 3 about here

The horizontal and right-hand vertical axes in Figure 3 are the same as in the other two Figures. The left-hand vertical axis is the CSEM, with the same scale used for all plots. Four CSEM curves are plotted: CAT with the full information scoring (solid line), CAT with the reduced information scoring (heavy dashed line), the reference test scored number-correct as is the current practice (thin dashed line), and the reference test scored with $\hat{\xi}$ (dotted line). The CSEM curves for CATs scored $\hat{\xi}$ have already been judged acceptable at the end of the test design simulations when compared to those for the two methods of scoring the linear reference tests. It seems likely that the CSEM curve for CATs scored with number-correct could reasonably be judged acceptable using the same criteria since the two methods of scoring adaptive tests produce very similar CSEM curves.

Conclusions and Discussion

Adaptive testing represents a new advance in testing technology that appropriately utilizes modern computing equipment and modern psychometrics. To enhance the prospects for public acceptance of adaptive testing, test sponsors must insure that adaptive testing is as understandable to test-takers, test users, and interested public institutions as conventional paper-and pencil testing. Scoring adaptive tests using the more familiar number-correct score (or the analogous formula score), accompanied by the necessary equating to adjust for the intentional differences in adaptive test difficulty, may be an important alternative.

Using equated number-correct as an adaptive test score can be viewed as a reduced information approach that ignores information available in the full information approach for the 3PL item response function model. Ignoring available information has an impact on the psychometric quality of adaptive tests. For some of the adaptive tests studied in this paper, equated number-correct scoring appears to be slightly less biased than $\hat{\xi}$ scoring, while for all tests it was more variable. In terms of the criterion quantities used to make comparisons of properties of the adaptive test and the (parent) reference test, the reliability of the equated number-correct score is decreased slightly when compared to that of $\hat{\xi}$ and the CSEMs are increased slightly, but the differences appear unimportant for these tests. In addition, the required IRT equating of individual adaptive tests to the reference test appears to function well.

Given the wide variety of measures represented by the six tests, these results suggest that equated number-correct scoring is a feasible alternative to scoring methodologies that rely more heavily on IRT. Comparable

investigations should, of course, be carried out on real adaptive test data from real test-takers when such data become available.

If adaptive tests were scored equated number-correct, it seems likely that public concerns would focus less on scoring adaptive tests (since the scoring would be the same for as for conventional tests) and more on issues that are central to the very nature of adaptive testing itself. These issues focus on questions about how items are chosen for a particular test-taker in such a way as to intentionally introduce differences in test difficulty. Test sponsors must not only be able to justify adaptive test construction in terms of the domain sampled and test purpose, as they now must do for conventional tests, but must also be able to justify the deliberate creation of test forms that are as parallel as possible in all aspects except difficulty. This seems to be a more germane focus for test-takers, test score users, interested legislative and regulatory institutions and test sponsors than a focus on particular scoring methodologies.

References

Bejar, I., and Wingersky, M. S. (1982). A study of pre-equating based on item response theory. Applied Psychological Measurement, 6, 309-325.

Cook, L. L., and Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., and Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. Applied Psychological Measurement, 11, 225-244.

Cook, L. L., Petersen, N. S., and Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study in scale stability. Journal of Educational Statistics, 8, 136-156.

Dorans, N. J. (1990). Scaling and equating. In H. Wainer (Ed.) Computerized adaptive testing: a primer, 137-160. Hillsdale, NJ: Lawrence Erlbaum Associates.

Eignor, D. R., Cook, L. L., and Stocking, M. L. (1990). Simulation results of effects on linear and curvilinear observed and true score equating procedures of matching on a fallible criterion. Applied Measurement in Education, 3, 37-52.

Eignor, D. R., Way, W. D., Stocking, M. L., and Steffen, M. (1993). Case studies in computer adaptive test design through simulation (Research Report 93-56). Princeton, NJ: Educational Testing Service.

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, 347-360.

Jacobson, R. L. (1993, September 13). New computer technique seen producing a revolution in testing. The Chronicle of Higher Education, p A22.

Killcross, M. C. (1976). A review of research in tailored testing. (Report APRE No. 9/76). Franborough, Hants, England: Ministry of Defense, Army Personnel Research Establishment.

Lawrence, I., and Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. Applied Measurement in Education, 3, 9-36.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. Applied Psychological Measurement, 1, 95-100.

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance and of their parallel forms reliability. Psychometrika, 48, 233-243,

McBride, J. R. and Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New Horizons in Testing, (pp 223-236). New York: Academic Press.

Mislevy, R. J. (1984). Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. Psychometrika, 51, 177-195.

Mislevy, R. J. and Bock, R. D. (1983). BILOG: Item and test scoring with binary logistic models [computer program]. Mooresville, IN: Scientific Software.

O'Neill, K., Folk, V., and Li, M.-Y. (1993). Report on the pretest calibration study for the computer-based academic skills assessments of The Praxis Series: Professional Assessments for Beginning Teachers (TM). Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 70, 351-356.

Schaeffer, G., Steffen, M., Golub-Smith, M. (1993). Introduction of a computer adaptive GRE general test (Research Report XX-XX). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 3/4, 263-277.

Stocking, M. L. (1992). Controlling item exposure rates in a realistic adaptive testing paradigm (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 18, xxx-xxx.

Swanson, L., and Stocking, M. L. (1993, in press). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17, 151-166.

Sympson J. B., and Hetter, R. D. (1985, October) Controlling item-exposure rates in computerized adaptive testing, as described in Wainer, et al. (1990).

Vale, C. D. (1981). Implementing the computerized adaptive test: What the computer can do for you. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., and Thissen, D. (1990). Computerized adaptive testing: a primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. Machine-Mediated Learning, 2, 217-282.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Weiss, D. J. (1974). Strategies of adaptive ability measurement. Research Report 74-5, Minneapolis: University of Minnesota, Psychometric Methods Program.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Wingersky, M. S., and Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Yen, W. (1983). Use of the three-parameter model in the development of a standardized achievement test. In R. K. Hambleton (Ed.),

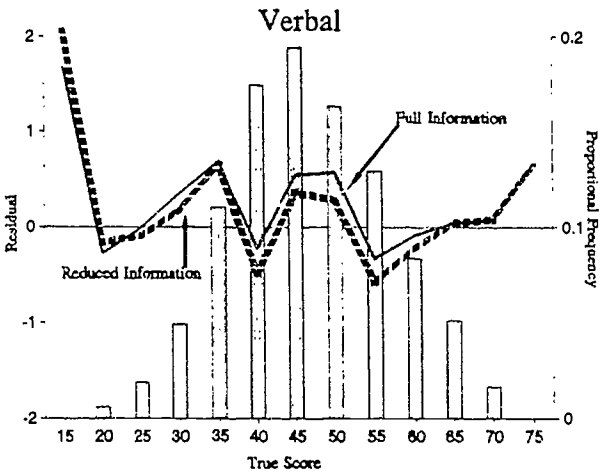
Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

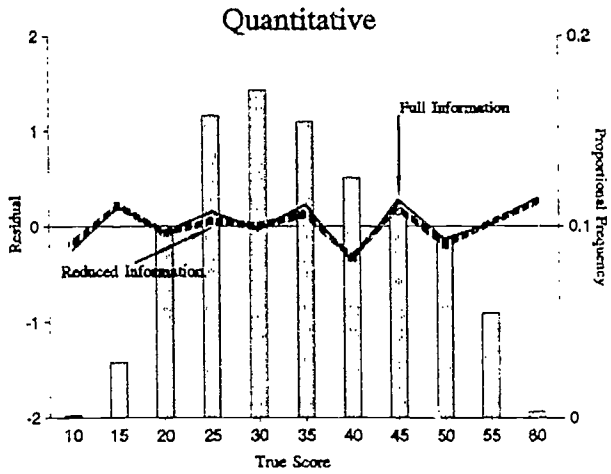
Table 1: The six adaptive tests.

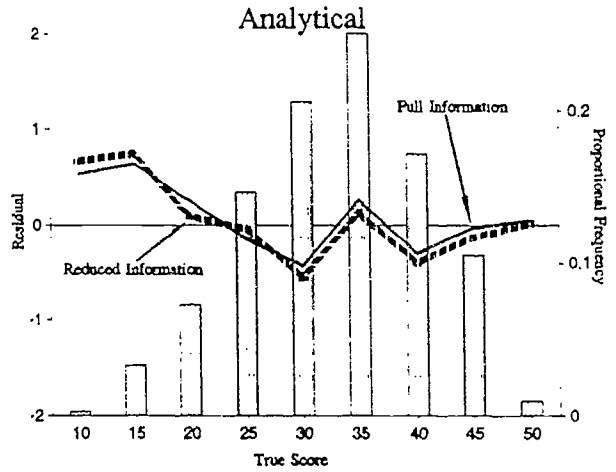
Test	CAT Length (items)	Number of Elements in Pool	Reference Test Length	Number of Constraints	Maximum Exposure Rate	Number of Simulees
Verbal	30	381	76	38	.20	1300
Quantitative	28	348	60	27	.20	1650
Analytical	35	512	50	43	.20	1170
Mathematics	24	287	40	30	.20	1600
Reading	31	443	40	27	.25	1400
Writing	30	373	40	34	.20	1400

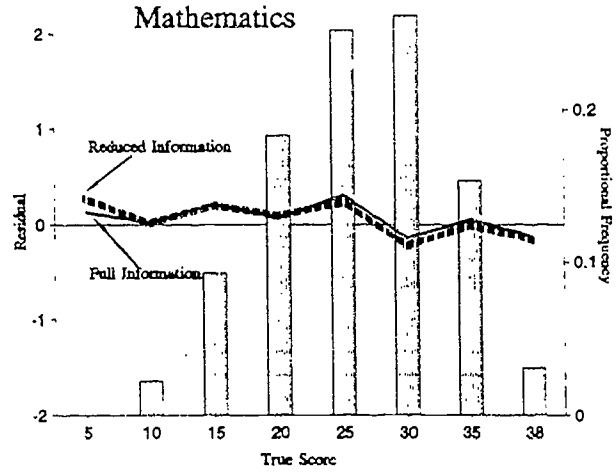
Table 2: Comparisons of full information (FI) and reduced information (RI) scoring of adaptive tests.

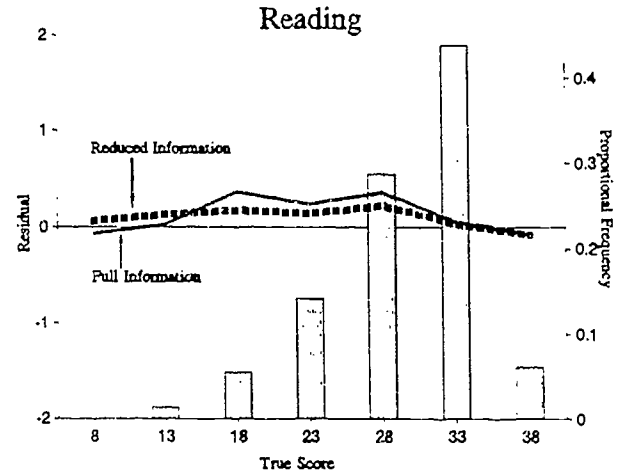
Test	Reliabilities				Weighted				Weighted
	CAT, FI	CAT, RI	Reference nr	Reference $\hat{\xi}$	Residual FI	Residual RI	Variance FI	Variance RI	Percent Correct
Verbal	.902	.897	.890	.910	.21	.02	10.87	11.59	63
Quantitative	.927	.922	.922	.933	.03	.00	8.86	9.55	54
Analytical	.894	.886	.889	.902	-.05	-.13	7.39	7.90	58
Mathematics	.883	.881	.874	.886	.08	.04	5.84	5.91	59
Reading	.846	.837	.816	.833	.18	.11	4.83	5.23	68
Writing	.836	.818	.794	.816	-.02	-.19	4.58	5.17	69











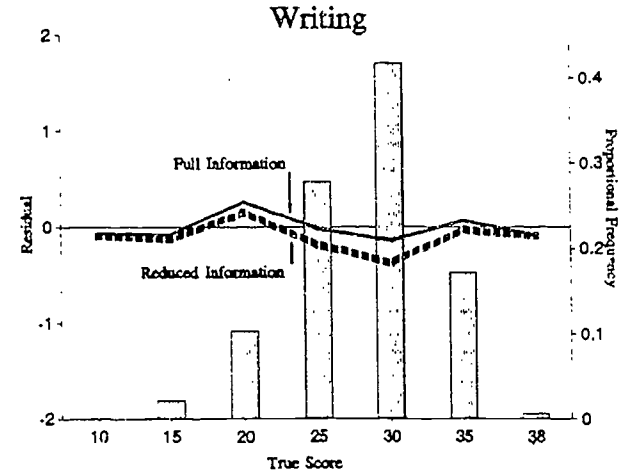


Figure 1: Residuals from true values for adaptive test scores derived from the full information method and the reduced information method.

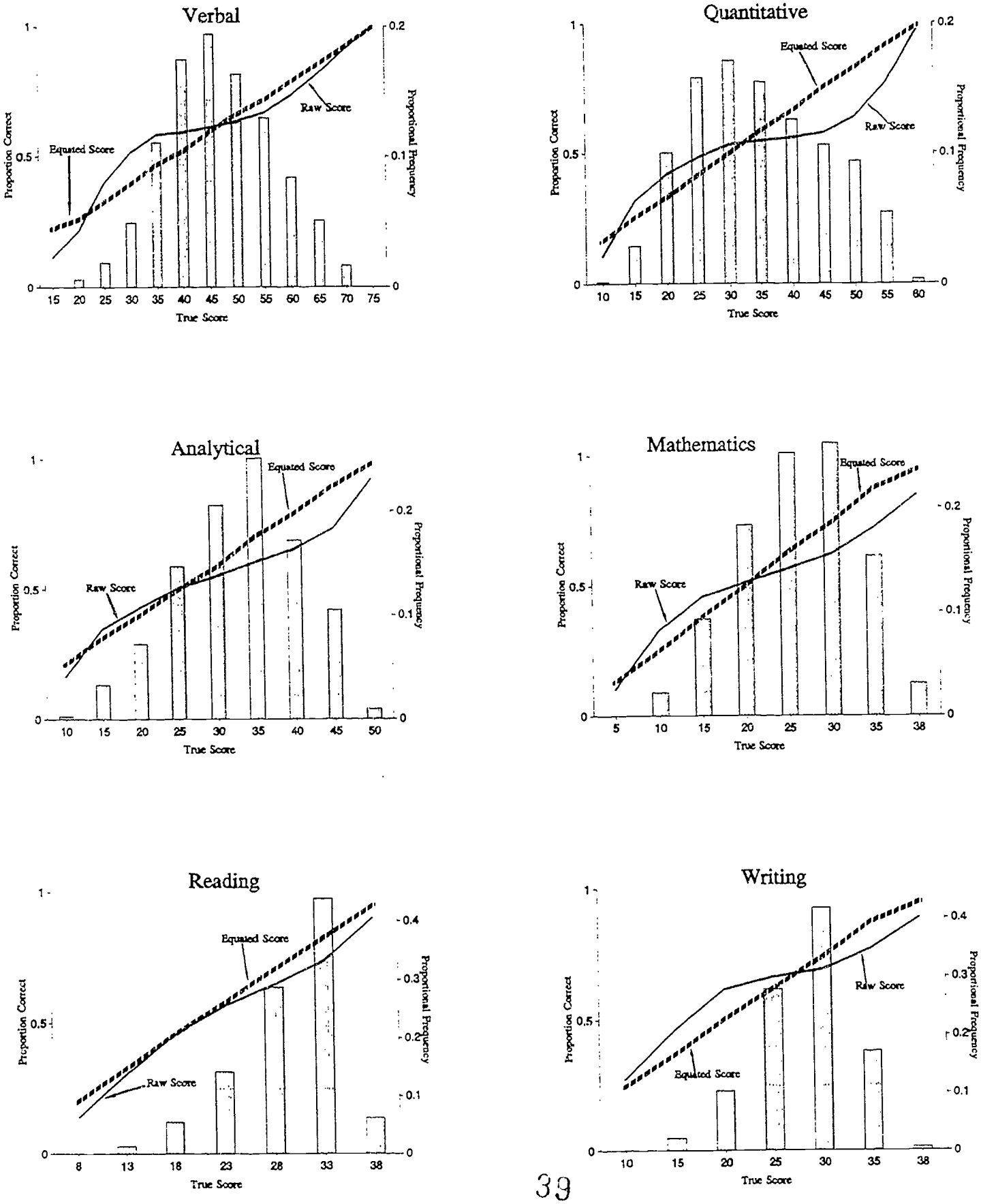


Figure 2: Raw and equated proportions correct for number right scoring of adaptive tests.

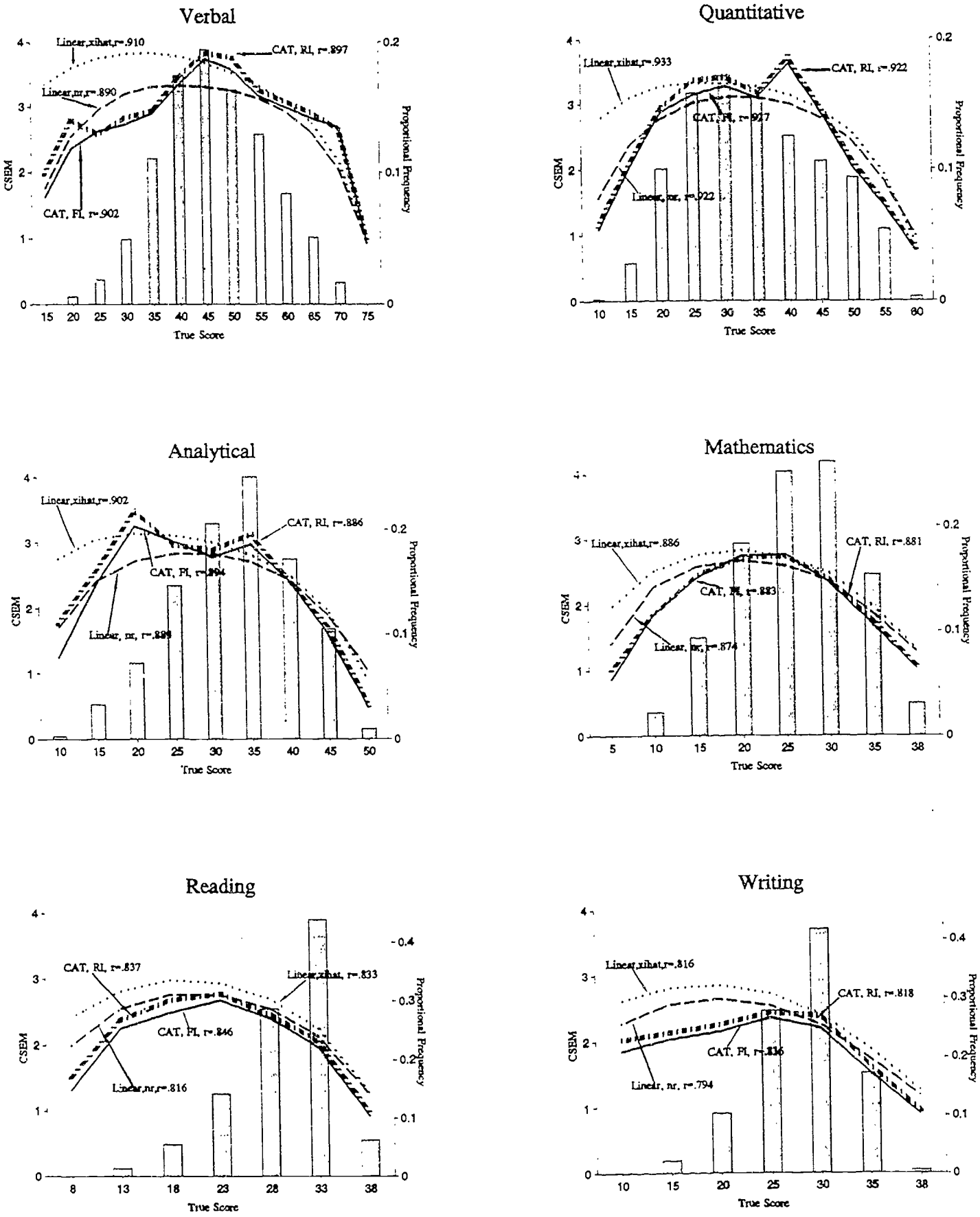


Figure 3: CSEMs for adaptive and linear tests scored by two different methods.