ED 431 797 TM 029 883

AUTHOR	Davey, Tim; Pommerich, Mary; Thompson, Tony D.
TITLE	Pretesting alongside an Operational CAT.
PUB DATE	1999-04-00
NOTE	22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, Canada, April 20-22, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Item Banks; Measurement Techniques; Prediction; *Pretests Posttests; Simulation; Test Construction; *Test Items; Test Length; Testing Problems
IDENTIFIERS	*Embedded Items

ABSTRACT

In computerized adaptive testing (CAT), new or experimental items are frequently administered alongside operational tests to gather the pretest data needed to replenish and replace item pools. The two basic strategies used to combine pretest and operational items are embedding and appending. Variable-length CATs are preferred because of the control they give test developers over what is measured and the precision of that measurement. When variable-length testing and embedded items are combined, problems arise because of the difficulty of distributing pretest items evenly across tests of unknown length. Simulation studies focused on the procedures for making the necessary adjustments and prediction of test length and the impact of these procedures on operational testing. Accurate prediction was found to be possible provided prediction begins only after at least five items have been administered. Predictions also improve with increasing test length. Very simple regression models proved both effective and highly stable in cross validation. (Contains 1 table, 5 figures, and 14 references.) (SLD)

Pretesting Alongside an Operational CAT

Tim Davey
Mary Pommerich
Tony D. Thompson

This paper is prepared for the:
Annual Meeting of the National Council on Measurement in Education
Montreal Canada, April 1999

Pretesting Alongside an Operational CAT

Tim Davey
Mary Pommerich
Tony D. Thompson
ACT

1. Overview

To be self-sustaining is a goal held by creatures ranging from bacteria to bureaucracies. Computerized adaptive testing programs prove no exception, administering new or experimental items alongside operational tests to gather the pretest data needed to replenish and replace their own item pools. The two basic strategies commonly used to combine pretest and operational items are *embedding* and *appending*. Embedded pretest items appear in the midst of an operational test, either in a contiguous block or scattered throughout. Appended pretest items are simply tacked to the end of the operational test.

The differences between embedding and appending are important from several perspectives. The first is the view that, for best effect, examinees should not know whether a particular item is operational or being pretested. Because the purpose of pretesting is to predict how items will perform once operational, pretest data are most useful when gathered under conditions that most approximate operational testing. An embedded approach that mingles pretest with operational items is therefore preferred to appending, where pretest items are segregated at the back of the test.

A second important difference becomes clear when examinees are put under time pressure. Examinees would certainly prefer to run out of time during an unscored pretest section after having completed all operational items. A test with appended pretest items is therefore functionally less speeded than one with both pretest and operational items scattered throughout. Whether or not this is a desirable outcome depends on the nature of what is being measured and the uses to which test scores are put.

In some cases, speededness is an unwelcome measurement intrusion that obscures the relationship between test scores and criterion measures. For example, the test taken to obtain a driver's license is perhaps best left unspeeded, even if most individuals prefer that their fellow motorists be able to think and react quickly. However, in other cases, the speed with which examinees respond is an important predictor of success on criterion measures. For example, air traffic controllers work in a fast-paced environment in which they are routinely asked to make split-second decisions that put at risk the lives of hundreds of travelers. It is therefore only proper that a selection test for these positions would pressure candidates to think and respond quickly, and that the ability to do so would be predictive of job success.

Where embedded items are located also affects the extent to which a test is speeded. Putting all of the pretest items forward in a test increases speededness by guaranteeing that examinees who don't finish leave only operational items unanswered. Clustering pretest items near the back of the test approximates appending and reduces speededness. Distributing pretest items evenly throughout the test is a nice balance that penalizes examinees in direct proportion to the number of items left unanswered.

The remainder of this paper begins, curiously enough, with a brief discussion of fixed- and variable-length adaptive tests. The intent of this section is to promote the use of variable-length tests, with one notable caveat. An elaboration of the pretest strategies presented above follows next, arguing in favor of the embedded approach, particularly for speeded tests. It is when variable-length testing and embedded pretesting are conjoined that problems arise, because evenly distributing pretest items across operational tests of unknown length can be a tricky proposition. To do so properly requires that the expected length of each examinee's test be continually predicted as the test progresses. Examinees expected to receive fewer operational items would be administered more pretest items throughout their test; examinees expected to take longer operational tests would receive fewer pretest items throughout. The paper concludes by focusing on the procedures for making the necessary predictions, and by discussing their impact on operational testing.

2. Fixed- and variable-length adaptive tests

The relative advantages and disadvantages of fixed- and variable-length adaptive tests have been debated elsewhere. Arguments favoring fixed-length tests cite the method's simplicity and its avoidance of a certain sort of measurement bias (Stocking, 1987). Proponents of variable-length tests contend that they are more efficient and allow test measurement properties to be precisely specified (Davey & Thomas, 1996; Thompson, Davey & Nering, 1998). Both views are briefly summarized below.

As its label suggests, fixed-length adaptive tests administer the same number of items to each examinee. The number of items administered is determined by weighing such factors as content coverage, measurement precision, and the time available for testing. Measurement precision is usually specified in the aggregate, or averaged across examinees at different proficiency levels (Thissen, 1990). However, the measurement models that underlie adaptive tests recognize that precision varies across examinees. Examinees whose proficiency levels are quickly and accurately identified can be repeatedly targeted with items of an appropriate difficulty and consequently measured very efficiently and reliably. Examinees whose performance levels are located in a range where an item pool is particularly strong are also likely to be well measured. Conversely, examinees who are difficult to target or whose proficiency levels fall where the item pool is weak will be measured more poorly.

The function traced by measurement precision over proficiency level can be manipulated in limited ways by test developers. Item pools can be bolstered where they are weak and weakened where they are unnecessarily strong. Test length can be shortened or lengthened. Item selection and exposure control procedures can be finessed. However, the level of control is far short of complete, leaving conditional measurement precision more a function of chance than of design.

Variable-length tests allow measurement precision to be addressed directly by using it as the criterion of when a test ends. Rather than administering a specified number of items to each examinee, variable-length tests instead administer items until a specified level of precision is met. Examinees who are measured efficiently, because they are well targeted for example, will reach the criterion quickly and take shorter tests. Other examinees in other circumstances will take longer tests. However, regardless of test length both sorts of examinees will be measured with the precision specified. Test precision is dictated rather than left to fate.

Being able to exactly specify test precision is a crucial advantage in test development. However, it does not come without cost, as variable-length tests have two faults, one relatively trivial and the second potentially important. The first problem is that the rule by which tests are stopped is necessarily a function of *estimated* examinee proficiency. Properly, it should be a function of *true* proficiency which, alas, is unavailable to us. The result is that bias in proficiency estimates influences, and in turn is influenced by, test length. Specifically, low-proficiency examinees who are administered shorter tests are generally underestimated, as are high-proficiency examinees who receive longer tests. Conversely, longer tests administered to low-proficiency and shorter tests administered to high-proficiency examinees will generally produce higher-than-expected proficiency estimates. However, the effect is subtle and disappears almost entirely as test length and reliability increase.

The larger problem with variable-length tests becomes apparent when they are administered under time limits. Equity concerns abound unless these limits are generous enough to allow all examinees to comfortably finish even the longest test they might be administered. Short of this, examinees who receive longer tests are put at a disadvantage. This situation must be addressed before variable-length tests, with their attendant benefits, can be a viable option for most high-stakes testing programs. One solution is to vary test time with test length. However, even this is open to the criticism that examinees taking longer tests are subject to increased fatigue. An alternative solution will be described below.

3. Pretesting

Pretesting has two basic goals. The first is to gather data on new or experimental items to determine whether they are suitable for future operational use and, if so, how they should be used. The second is to avoid interfering with operational testing. Unfortunately, these goals can conflict with each other. Pretest data, which are used to predict how experimental items will perform operationally, are most useful if gathered under conditions that are most nearly operational. Pretesting alongside operational tests is therefore preferable to pretesting with special test forms or with special examinee populations. The idea is to hide pretest items inside the operational test and leave examinees none the wiser as to their location. Examinees unsure whether or not their answer to an item will contribute toward a score they care about are apt to respond to the best of their ability and provide the most data.

The danger is that mixing pretest with operational items can somehow affect performance on the latter. This is a particular concern with speeded tests. An ambiguous, poorly written, or just inordinately difficult pretest item can cause an examinee to waste time better put to more productive use on items that count.

Three strategies have evolved for pretesting, each offering a different balance between obtaining quality data and avoiding impact on the operational test (Millman & Greene, 1989):

1. Separately timed sections. This approach completely prevents pretest items intruding on operational testing by isolating them in separately timed "special" sections. Although these may look and function like the real thing, savvy examinees can readily identify them as pretests. However, there is a long history of effective use with high-stakes testing programs. When stakes are high, any uncertainty as to the value of an item apparently ensures a motivated response.

2. Appending to operational tests. Here, pretest items are attached to the end of an operational test, with the whole administered under a single time limit. The hope is that the pretest items at the end of the test will not affect how examinees respond to the operational items at the front. Since no indication is given of where the operational test ends and the pretest begins, there is also the hope that examinees will treat all as operational. However, given a little instruction and practice most examinees could be taught to identify the boundary that separates the important scored items from the time-wasting pretest. The cues might be a dramatic change in item difficulty, a change in item format, a change in the number of options per item, or a change in the number of items attached to stimulus passages. In any case, examinees would know that it's better to leave items at the end of the test unanswered rather than those at the front or in the middle. The test is functionally less speeded for examinees who know this or are able to recognize the pretest component and focus their time and attention elsewhere.

3. Embedding in operational tests. A third strategy is to better disguise pretest items by surrounding them by the operational test. Although the clues listed above would still be valid, the contrast between pretest and operational items is less stark. Spotting pretest items distributed individually or in small blocks throughout an operational test is a daunting task. Few examinees would be confident enough in their judgement to risk slighting what they believe to be unscored items. An embedding strategy is therefore most likely to yield the best possible pretest data.

Embedding is also likely to be more equitable with speeded tests. Assume that the penalty for failing to finish a test is proportional to the number of operational items left unanswered. Other sorts of penalties will be discussed later. When pretest items are appended, the penalty function declines to zero at the end of the operational test. For example, consider a 50-item test where the first 40 items are operational and the last 10 pretest. Examinees who finish fewer than forty items are penalized in proportion to the number of items unanswered. An examinee who answers 26 items is less penalized than one answering 25. But since the penalty drops to zero at the end of the operational test, examinees who finish the entire test derive no advantage over examinees who answered only the first 40 items. This is satisfactory if test developers have explicitly decided that response speed should impact the scores of only the slowest examinees. However, in many cases speed of responding is worth measuring for all examinees. This is best done by distributing pretest items evenly throughout the test. In the example, pretest items might take positions 5, 10, 15,...,50 rather than the last ten slots.

The problem with embedding is that it maximizes the risk of pretesting impacting operational test scores. Since most examinees work through a test from front to back, time unnecessarily spent on pretest item early in a test can detract from performance on later operational items. This is particularly problematic if pretest items differ across examinees. Examinees receiving difficult, ambiguous or otherwise time-consuming pretest items are at a disadvantage to examinees who take an easier, shorter more straightforward pretest. The problem is not simply that different examinees took different items. This occurs whenever alternate test forms are administered and is a necessity of adaptive testing. However, in these cases the operating characteristics of items are usually known and taken into account when forms are assembled or adaptive tests selected. Pretest items, in contrast, are by definition unknown quantities. Minimizing disparate impact might therefore depend on little more than randomizing sets of pretest items across examinees and hoping for the best. A more satisfying alternative is to carefully screen pretest items either judgmentally or with small samples before administering them more widely.

4. Variable-length testing and embedded pretesting

The previous sections have extolled the virtues of variable-length adaptive testing and embedded item pretesting. This section tries to fit these ideas together. Recall that the drawback of variable-length tests was their unfairness under speeded conditions. Examinees taking shorter tests have an advantage over those taking longer tests when time limits are equal. We propose to correct this by attaching to each variable-length operational test a variable number of pretest items so that all total tests are brought to an equal length. Suppose that total test length is fixed at forty items. Then examinees taking 32 and 35-item operational tests would be administered eight and five pretest items, respectively, bringing each test to 40 total items. Provided pretest items are not identified as such, tests will appear to examinees to be of equal length.

Equalizing perceived test length goes a long way toward the goal of minimizing differential speededness across examinees. However, two other factors are also important. The first is how examinees are penalized for not completing a test. With conventional tests, omitted or unreached items are usually scored as incorrect. When tests are scored by number correct, it is therefore in the examinee's best interest to answer all items, even if with random guesses. Any correct guesses will increase a score and wrong answers don't carry additional negative weight. The situation is different when a test is formula scored, where right answers count positively and wrong answers fractionally negative[1]. Unless an examinee is certain of their ability to guess correctly more often than chance would suggest, there is no benefit to completing a test with random responses.

Penalty schemes and optimal guessing strategies are more complicated with adaptive tests. One approach tried was to require examinees to complete a certain minimum number of items in order to receive a score, but attach no penalty to failing to complete the entire test (Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995). Unfortunately, examinees quickly learned that there was no reward for completing the whole test and so instead spent all of their time on the minimum number of items. This

[1] The usual formula is to count right answers as +1 and wrong answers as $-1/m$, where m is the number of response options.

left the test less speeded than designed and introduced inequities between examinees who were aware of this strategy and those who were not.

A second approach is to mimic conventional tests by scoring unanswered items as incorrect. However, this has two fundamental problems. This first is determining exactly which items are to be scored as incorrect. Adaptive tests select subsequent items depending on current responses. Requiring that a test conclude with a string of incorrect responses will cause successively easier items to be selected as the examinee's ability estimate declines. How much the ability estimate declines is based not just on the number of items unanswered, but also on the ability estimate when time ran out and on the items selected subsequently. The strong random component to item selection due to exposure control also causes different examinees to be affected in different ways despite their having left the same number of items unanswered. One approach to this problem is to determine an average or expected penalty for incompletion and to apply it uniformly to all examinees in the same circumstances (Schaeffer, Steffen, Golub-Smith, Mills & Durso, 1995).

The second problem with scoring unanswered questions as wrong concerns the process that examinees must follow to respond to a computerized questions. With paper-and-pencil tests it takes no more than a few seconds to quickly fill in any remaining ovals on the answer sheet just before time is called. But clicking your way through a computerized test to accomplish the same thing will likely take much longer. Telling examinees that it is in their best interest to answer all items is no help if they are logistically unable to do so. It may therefore be unfair to count nonresponses as incorrect.

A better approach may be to do for examinees what they would have done for themselves had they been able to, namely fill in unreached items with random guesses. Because the problem of figuring out exactly which items are guessed at still applies, the idea of determining and applying an average or standardized penalty still makes sense.

The other major influence on test speededness is where pretest items are located within the operational test. Distributing them evenly has already been argued as most equitable. We will try to strengthen and extend this argument in the context of a variable-length CAT.

The difference between appending and embedding pretest items in variable-length tests can be made concrete by considering a test that administers 40 total items. Now compare the tests of two examinees, one (A) who receives 36 operational and 4 appended pretest items, and a second (B) who receives 32 operational and 8 appended pretest items. Their tests are represented schematically below:

```
             ------------------------------------------------------xxxx
Examinee A  ____________|___________|___________|___________|  36 operational (-) / 4 pretest (x) items.
                       10          20          30          40

             --------------------------------------------------xxxxxxxx
Examinee B  ____________|___________|___________|___________|  32 operational (-) / 8 pretest (x) items.
                       10          20          30          40
```

Suppose both examinees were able to finish only 30 items. Then Examinee A would have the six remaining operational items completed with random responses, 1/6 of the operational test. But Examinee B would have only two random responses, a mere 1/16 of the operational test. Examinee A would thus pay a much stiffer price for non-completion although both examinees worked at the same rate.

Now suppose that pretest items were embedded and evenly distributed throughout the test. Then examinees are matched at every point of their tests in the proportions of operational items taken, even when the number of operational items taken differs. So after 10 items have been administered, every examinee should have taken 25% of their operational items. After 20 items have been administered, every examinee should have taken 50% of their operational items, after 30 items 75%, and so on. This situation can be represented as follows:

```
            -------------x--------------x-------------x--------------x
Examinee A  _____________|______________|_____________|______________|  36 operational (-) / 4 pretest (x) items.
                        10             20            30             40

            ------x------x------x------x------x-------x------x------x
Examinee B  _____________|______________|_____________|______________|  32 operational (-) / 8 pretest (x) items.
                        10             20            30             40
```

If both examinees completed 30 items, A would have 9 operational items answered randomly, ¼ of the total. Examinee B would have only 8 guessed responses, but this is the same ¼ of the total operational test. Equity reigns.

How to distribute pretest items evenly within the operational test is the problem. The length of the operational test is not known until it has ended, leaving the necessary number of pretest items also unknown. The trick is to evenly mix these two sets of items of unknown size. One solution is to continually predict the final length of the operational test as it proceeds and use these predictions to determine how many pretest items will be needed and where they should be located. Examinees predicted to take shorter operational tests would take more, and more closely spaced, pretest items. Examinees expected to take longer operational tests would be administered fewer, more widely spaced pretest items. Every pretest item administered would be followed by an operational item, and the prediction computed again. Equitable variable-length testing is therefore possible to the extent that accurate prediction of final test length is possible. The next section describes our attempts at making these predictions.

5. Predicting variable test lengths

The particular sort of variable-length adaptive tests considered here use test (Fisher) information as the stopping criterion (Thissen & Mislevy, 1990). Fisher information is calculated after each item has been answered as a function of the items so far administered and of the current ability estimate. Information values computed following each response are continually compared with a specified target value. The test ends when accumulated information exceeds the target[2].

Various data are available following each item administration to be used to predict final test length. Values that seem like they should be useful as predictors include the number of items already administered, the current ability estimate, the amount of information so far accumulated at the current ability estimate and the amount of additional information needed to meet the stopping target. Other values are less obvious, but may be important as well. For example, the stability of the current ability estimate indicates how likely it is that information is being accumulated at the right place. It is not unusual early in a test to see information at the current ability estimate actually decline as that estimate changes. The rate at which information is accumulating is another potential predictor. As tests lengthen and ability estimates stabilize, item targeting and test efficiency improves and information accumulates more quickly.

A number of models were investigated to determine whether accurate prediction of test length was possible. This was examined in a realistic simulation context in which examinees and items were modeled multidimensionally (Davey, Nering & Thompson, 1997). Although simpler models adequately capture the gross properties of observed response data, they lack the complex, subtle, 'noisy' features characteristic of actual response behavior. Accordingly, more realistic simulation procedures based on high-dimensional item response models were developed and implemented. All evaluations of testing procedures are conducted exclusively under these simulation conditions. It is not enough that a testing procedure work effectively when examinees respond as we *hope* that they will. A test procedure must also be robust enough to work effectively when examinees respond as we *know* that they will.

The simulation process begins by fitting high-dimensional compensatory logistic item response models to real examinees and real test items. This can be done with any of several software packages (Fraser, 1986; Wilson, Wood & Gibbons, 1991). No attempt is made to interpret the resulting solution. The fitted model is simply treated as a template from which new data can be generated. There is also no concern regarding overfitting or "capitalizing on chance". In fact, since the intent is to generate new data that embodies all of the characteristics of the real data, a certain degree of overfitting is desirable.

The estimated multidimensional item response functions are used to generate data by procedures directly analogous to those used with unidimensional simulations. The important difference is that multiple ability parameters are generated for each simulated examinee, all of which influence each item response. The item and ability parameters combine to produce probabilities of correct responses, just as they do with simpler unidimensional models. Data are generated by comparing these probabilities to random draws from a uniform distribution.

[2] Other conditions can be added to the stopping rule. For example, tests that impose content constraints on item selection can require that these constraints be adequately satisfied before the test is allowed to end. The prediction methods outlined are unaffected by these additional conditions.

Adaptive test data were simulated under the particular conditions set forth in Table 1. Tests included a minimum of one and a maximum of 60 items, with test length averaging about 25 items. Most examinees took between 20 and 40 items. The test ended when accumulated information exceeded a specified target, one that varied across ability levels. Items selected for administration were those that had maximum information at the current ability estimate. Item selection was further controlled by imposing content constraints and an exposure control procedure known as hybrid (Nering, Davey & Thompson, 1998). Current or provisional ability estimates were by Bayes mean (EAP). Once each test was completed, a maximum likelihood final ability estimate was computed.

Table 1. Test Conditions used to Create Prediction and Cross-Validation Samples.

Minimum test length	1
Maximum test length	60
Provisional ability estimate	EAP
Final ability estimate	MLE
Item selection	Maximum information
Stopping criterion	Information exceeds target[3]
Exposure control	Hybrid

The simulated CAT environment was used to generate two independent samples, each containing a complete examinee record of item responses, provisional ability estimates, standard errors of ability estimates, and target and examinee information, over all items taken for 4800 total examinees. Counting each presented item (and its accompanying statistics) as an observation, these samples contained about 120,000 observations each. The first sample (the prediction sample) was used to develop and examine various prediction models. The second sample was used to cross-validate the model that looked most promising.

The prediction sample was reduced by randomly selecting one item (observation) from each examinee. However, the last item taken was not eligible for selection because at that time the stopping criterion is met, and there is no need to predict final test length[4]. The item selected for each of the 4800 examinees ranged between 1 and 54, and was about 12.6 on average. The item selected will be referred to as "item position" because it represents the position in the sequence of items taken.

A variety of regression models were fitted to predict the final test length from variables characterizing items and examinees at the item position. Equivalent models predicting the number of

[3] The target information is presented in detail in Fan, Thompson & Davey (1999).

[4] In the case of a handful of examinees, the stopping criterion was not met at the maximum test length, so that final test length was the maximum test length.

items remaining to be taken at the item position (final length – item position) were also explored. Models were chosen by a combination of stepwise regression procedures, clinical judgment and women's intuition. Predictions of the number of items remaining were consistently poorer than predictions of final test length. All results presented here are therefore based on predictions of final test length.

Predictors characterizing the items and examinees at the item position included (in addition to item position itself) the current ability estimate, the proportion of information target met at the current ability estimate and the standard error of the current ability estimate. Transformations of these variables were also considered, for example the arc sine of the proportion of target met and the absolute value of the current ability estimate. More complicated predictors summarized test progress so far by measuring the rate at which observed information was approaching the target criterion or the rate of change of the current ability estimate.

Preliminary model fitting generally yielded predictions that were unacceptably poor. However, things improved markedly when prediction was attempted only after at least five items had been administered. The results presented here are based on this restricted sample (N=3695). By far the best single predictor of final test length was the rate at which information was accumulating, as measured by the change averaged across adjacent items (R^2=.48). Combinations of different variables, however, substantially outperformed the single predictor. Simpler variables were also consistently better than more complex alternatives. Further additions of some interaction terms increased R^2 substantially.

The model judged as best included five predictors: item position, the difference between the observed information and the target, the current ability estimate, an interaction term between item position and difference in target and observed information, and an interaction term between provisional ability estimate and difference in target and observed information. The model fit was significant overall with $p < .0001$, and individual t-statistics for each variable were also significant with $p < .0001$. In all, the model accounted for 77% of the variance in final test length. Figure 1 shows the predicted final lengths versus the observed final lengths based on the final model. Numeric results are shown in Table 2.

As might be expected, predictions were poorest when made from the earliest item positions. As the item position increased, the quality of the prediction improved. However, smoothing predictions by computing the predicted final length as the average of the predicted score for the current item and the two previous items did not increase R^2.

The selected model and its estimated parameters were cross-validated by applying them to the second data sample. For convenience, this sample was also reduced by randomly selecting one observation from each examinee. Results demonstrated remarkable stability, the model attaining an R^2 in the second sample essentially equal to that from the first.

To gain insight into how the prediction methods worked, and when they did not, we looked in detail at individual examinees. Results for four selected test administrations are given by the plots in Figure 2-5. The plotted symbols show the series of predictions of final test length made after each item response.

Recall that no prediction is possible until at least five items have been answered. The solid line on each plot indicates the final test length actually observed. Figures 2 and 3 are typical of the large bulk of tests, which are generally well-predicted. Final test length is predicted accurately and consistently from the fifth item through the next-to-last.

Figures 4 and 5 are the exception, representing cases where the prediction methods failed to varying extent. In Figure 4, final test length is badly underpredicted until the test was about half completed, after which predictions began tracking reality. Close examination of the items administered and responses offered revealed the explanation. This simulated examinee had a "true"[5] ability of around zero. However, after five items the current ability estimate of less than –1.0, the responses following the classic pattern of a "slow starter." Because the information target at abilities below –1.0 was much lower than the target around zero, the prediction system quite reasonably assumed the test would be a short one. In fact it would have been had the ability estimate remained at –1.0. Instead, the examinee recovered from the poor start and the ability estimate crept up toward zero. Because the information target increased steadily as the ability estimate approached zero, the predicted test length increased correspondingly each step of the way. Recovery was complete by the 15- or 20-item mark, with predicted lengths becoming accurate from that point onward. This sort of prediction error will inevitably affect a small percentage of examinees, and little can be done to correct it.

The situation in Figure 5 is different, with test length consistently and badly underpredicted throughout. This examinee also had a true ability near zero, but responded in a much steadier fashion than the previous simulee, the ability estimate never straying far from zero. What went wrong was the fault of exposure control. Since the information target at zero is relatively high, examinees there need to receive more than a few discriminating items in order to reach the stopping criterion. Unfortunately, these are the very items that are most stingily protected by the exposure control procedures. We routinely compute a statistic that indicates the extent that an examinee's test was degraded due to exposure control. Essentially, this statistic counts the average number of items denied from use by exposure control for each item that is allowed to be used. This average was much higher for the examinee in Figure 5 than it is generally. The predictions underestimated test length because they overestimated the discriminations of the items that would be administered. More will be said about this below.

6. Using test length predictions

The primary reason for predicting test length is to allow pretest items to be evenly distributed throughout each examinee's operational test. Assume that each examinee is to receive 40 total items.

[5] Because simulated examinees were modeled in multiple dimensions, a single "true" ability does not exist. Each examinee instead has a vector of 50 true abilities. However, for convenience it's nice to be able to compare the unidimensional ability estimates produced during the CAT administration to some unidimensional reference value. Happily, there are numerous ways of projecting or summarizing the true 50-dimensional ability vector as a single value. Here, we use a type of "true score" obtained by summing for an examinee the true (multidimensionally determined) response probabilities across all pool items. This is the examinee's expected number correct, if the

Suppose that after the fifth item we predict that a given test will end after 35 items. Then we have five pretest items to embed, at positions 8, 16, 24, 32 and 40. We would thus hold off administering a pretest item for the time being. If. on the other hand, final test length was predicted at 30 items then we would have ten pretest items to distribute. These would ideally appear in positions 4, 8, 12..., 40. Since we have already administered five operational items, we are behind schedule and would immediately administer a pretest item.

Simply put, the strategy is to predict final test length following each item response and, contingent on that prediction, decide whether or not a pretest item should be administered next. However, several considerations complicate this decision. The first, and most important is to prevent pretesting from compromising the operational test. A good example of how compromise might occur is offered by the examinees in Figures 4 and 5. Both of their test lengths are underpredicted, at least in their early stages. This could lead to a large number of pretest items being administered early on. By the time it was realized that the test was actually to be much longer than anticipated, too many pretest items may have been administered to allow the operational test to meet the information target. The examinee would therefore be measured less precisely than specifications called for.

Overprediction of test length is also possible. The problem in this case is that pretest items would be spread too thinly across the test. The remaining pretest items needed to complete the test might then be concentrated toward the end, making the test functionally less speeded than is desirable. The decision rule that dictates when pretest items are administered must weigh the danger of reduced speededness against the competing problems of degraded measurement.

Test length predictions can also be more tightly integrated into item selection to correct a number of related problems that beset adaptive tests generally and variable-length tests in particular. Consider dividing an item pool into two or more strata based on item discrimination. Doing so will roughly divide items by the frequency with which they are selected for administration, with more discriminating items generally being more frequently selected. Chang and Ying (1999) suggest an item selection procedure that chooses items from successively more discriminating strata as a test proceeds. Thus, early selections are made from among the least discriminating items, items in the middle of the test are chosen from middle strata, and final items drawn are most discriminating.

Test length prediction can improve Chang and Ying's procedure by transitioning item selection from one stratum to the next based on need rather than some programmed schedule. An examinee expected to take a long test would be given proportionally more items from higher-discriminating strata. Conversely, examinees predicted to take shorter tests would receive items mostly from less discriminating strata. This has several salutary effects on the tests delivered. First, test length will be more nearly equalized across examinees. Although tests will almost certainly remain varied in length, the range of difference will be

entire pool were presented. The true score is then converted back to the unidimensional ability metric using the (inverted) unidimensional test characteristic curve for the full pool.

smaller. This makes the problem of evenly distributing pretest items more tractable by lessening the chance that poor predictions will leave pretest items clustered at the back of the test, or that operational tests will be deficient because of excessive pretesting. However, occasional poor predictions like those shown in Figure 4 remain inevitable and little can be done about them.

A second, related positive effect is that very short operational tests are made unlikely. By dividing required test information over more rather than fewer operational items, problems with extreme ability estimates are avoided. With very short tests, the odds are good that some number of examinees will respond either entirely correctly or entirely incorrectly. Lengthening these tests by administering less discriminating items makes perfect patterns far less probable.

Test length predictions can also be used to amend exposure control procedures and prevent the situation exemplified by Figure 5. Recall that this examinee was forced to take an exceptionally long test because exposure control procedures barred access to most of the pool's discriminating items. That some small number of examinees will be treated this way is inevitable with any sort of probabilistic exposure control. Where test length prediction can help is by again distributing discriminating items by need rather than by chance. Exposure control would be relaxed for examinees expected to take longer tests and stiffened for examinees destined for shorter tests. This would lead to more equal test lengths and perhaps even improved exposure control.

7. Discussion

We began by arguing the advantages of both variable-length adaptive tests and embedded item pretesting. Variable-length tests were preferred largely because of the control they give test developers over what is measured and the precision of that measurement. Efficiency is a second advantage. Embedded pretesting was advocated both because it is more equitable under speeded test conditions and because it is likely to yield better pretest data. The problem was to evenly distribute embedded pretest items within a variable length operational test. Test length prediction was proposed as one way of attacking this problem.

Accurate prediction of test length was shown possible, at least in the context of a realistic simulation. Multiple correlations of nearly .80 were demonstrated provided prediction begins only after at least five items have been administered. Predictions also improve with increasing test length, a welcome state of affairs since decisions made early in a test are less crucial than those made later when there is less time to recover from error. Very simple regression models proved both effective and highly stable in cross validation.

Obviously, much remains to be done. Decision rules for spacing embedded pretest items given test length predictions need to be developed and evaluated. Similar rules for equalizing test lengths across examinees need to also be examined. The current work is perhaps best seen as a feasibility study for these next steps.

8. References

Chang, H-H. & Ying, Z (1999). Stratified multistage computerized testing. *Applied Psychological Measurement.* To appear.

Davey, T.C. & Thomas, L.A. (1996). *Developing adaptive tests to parallel conventional programs.* Presented at the annual meeting of the American Educational Research Association, New York.

Davey, T.C., Nering, M.L. & Thompson, T.D. (1997). Realistic simulation of item response data. Presented at the annual meeting of the American Educational Research Association, Chicago.

Davey, T.C. & Nering, M.L. (1998). Controlling item exposure and maintaining test security. Presented at Computer-Based Testing: Building the Foundation for Future Assessments. Philadelphia, PA.

Fan, M., Thompson, T.D. & Davey, T.C. (1999). *Constructing adaptive tests to parallel convetional programs.* Presented at the annual meeting of the National Council on Measurement in Education. Montreal.

Fraser, C. (1986). *NOHARM: An IBM PC Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory* [Computer Program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Millman, J., & Greene, J. (1989). The specification and development of tests in achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., Chapter 8, pp. 335-366). New York: Macmillan.

Nering, M.L., Davey, T.C. & Thompson, T. (1998). *A hybrid method for controlling item exposure in computerized adaptive testing.* Presented at the annual meeting of the Psychometric Society, Champaign, IL.

Schaeffer, G.A., Steffen, M., Golub-Smith, M.L., Mills, C.N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE General Test* (Research Report No. 95-20). Princeton, NJ: Educational Testing Service.

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

Thissen, D. (1990). Ability and measurement precision. In Wainer, H. (Ed.), *Computer adaptive testing: A primer* (Chapter 7, pp. 161-186). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. & Mislevy, R.J. (1990). Testing algorithms. In Wainer, H. (Ed.), *Computer adaptive testing: A primer* (Chapter 5, pp. 103-136). Hillsdale, NJ: Lawrence Erlbaum.

Thompson, T., Davey, T.C. & Nering, M.L. (1998). Constructing Adaptive Tests to Parallel Conventional Programs. Presented at the annual meeting of the American Educational Research Association, San Diego.

Wilson, D.T., Wood, R. & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* [Computer Program] Scientific Software, Inc. Chicago.

Table 2. Regression Results for Final Model

Model Summary:

F	p	R^2	RMSE
2495.32	.0001	.78	4.63

Parameter Summary:

Variable	Estimate	se	t	p
Intercept	2.06	0.27	7.58	.0001
Item Position	0.96	0.01	70.07	.0001
Difference in Information	1.78	0.05	37.67	.0001
Ability Estimate	0.76	0.09	8.06	.0001
Item Position*Information Difference	0.06	0.00	15.89	.0001
Ability Estimate*Information Difference	-0.52	0.03	-20.51	.0001

Figure 1. Predicted and Observed Final Test Lengths for Prediction Sample.

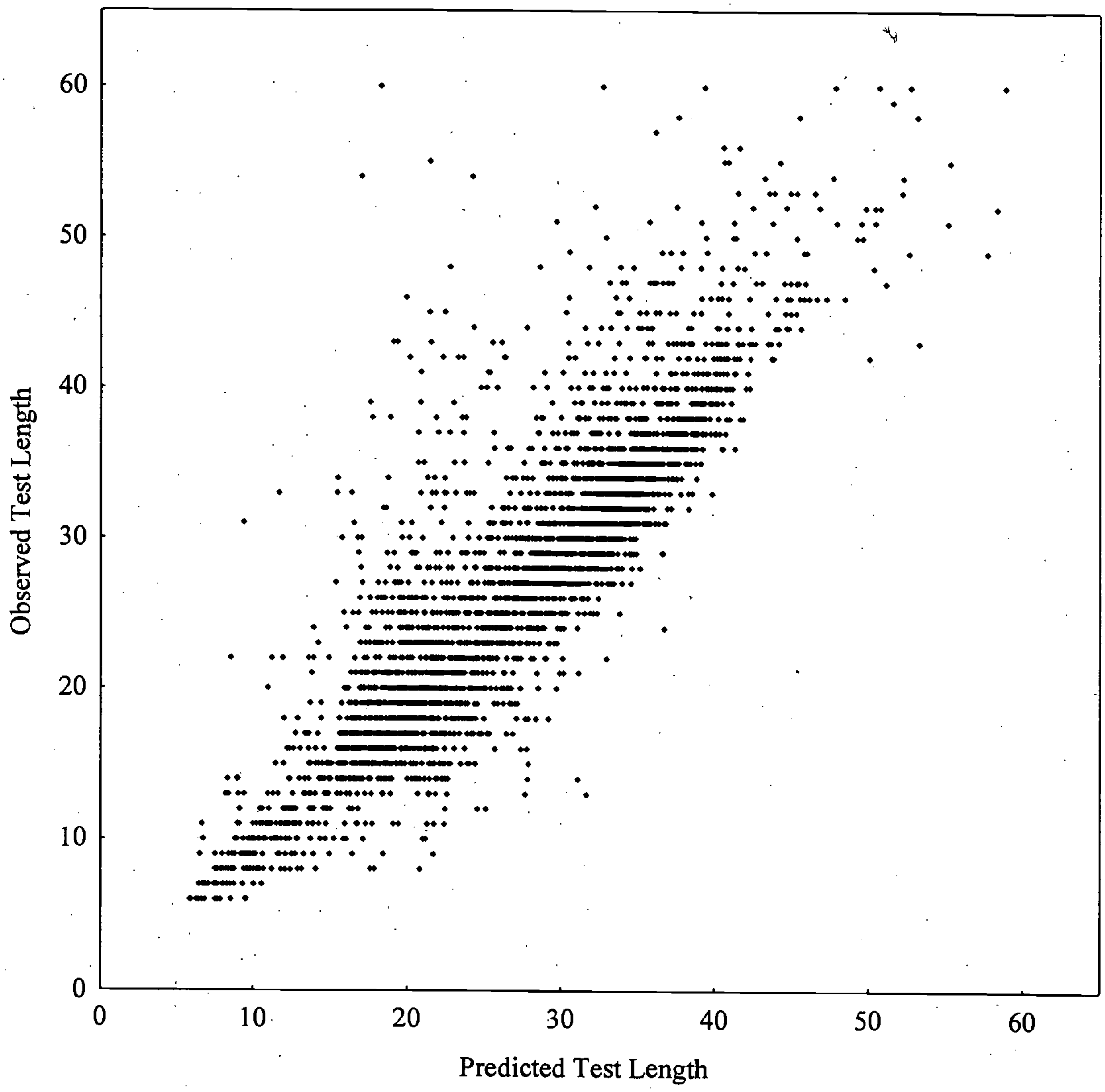


Figure 2. Predicted and Observed Final Test Lengths for an Examinee (Final Test Length = 17 Items).

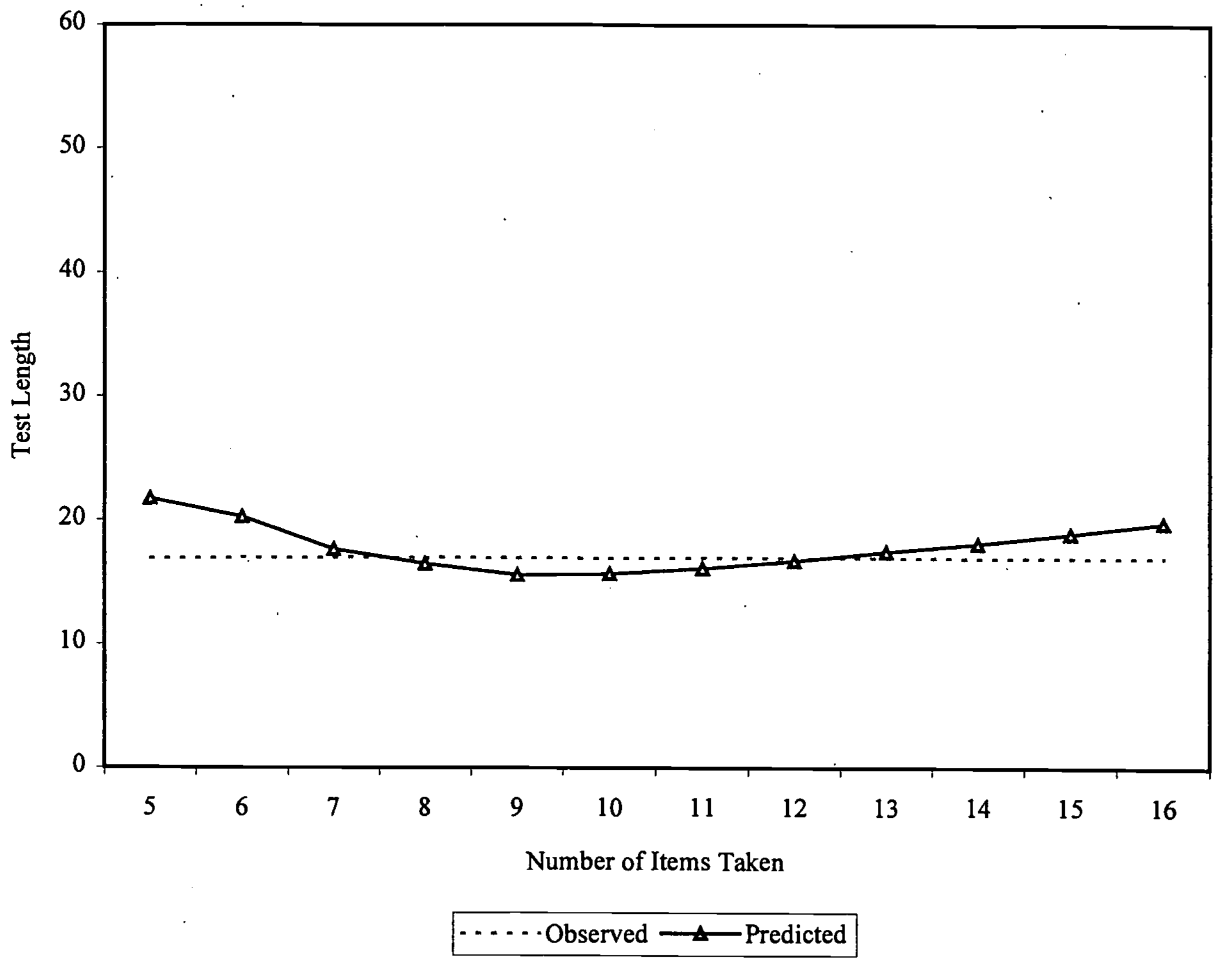


Figure 3. Predicted and Observed Final Test Lengths for an Examinee (Final Test Length = 27 Items).

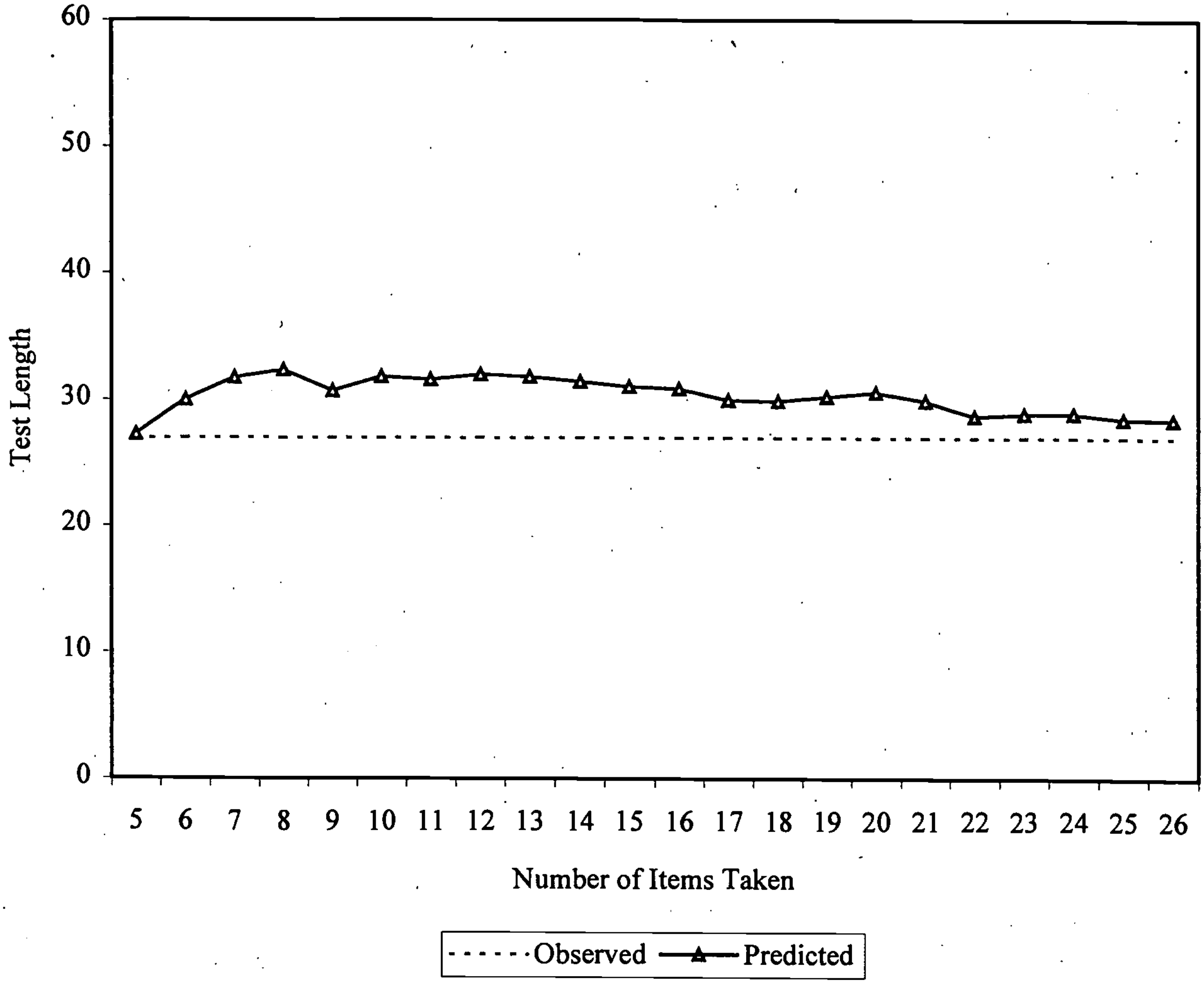


Figure 4. Predicted and Observed Final Test Lengths for an Examinee (Final Test Length = 34 Items).

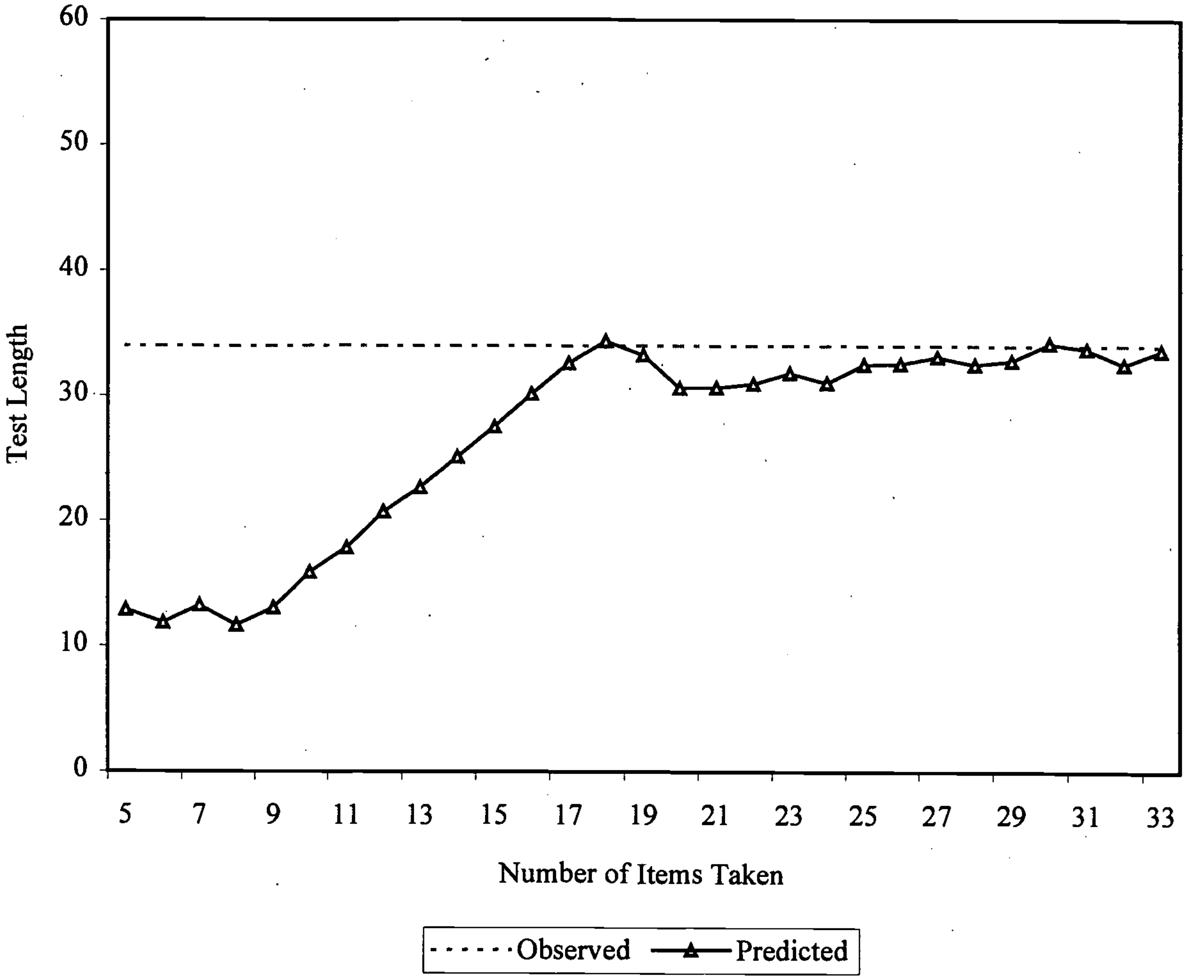
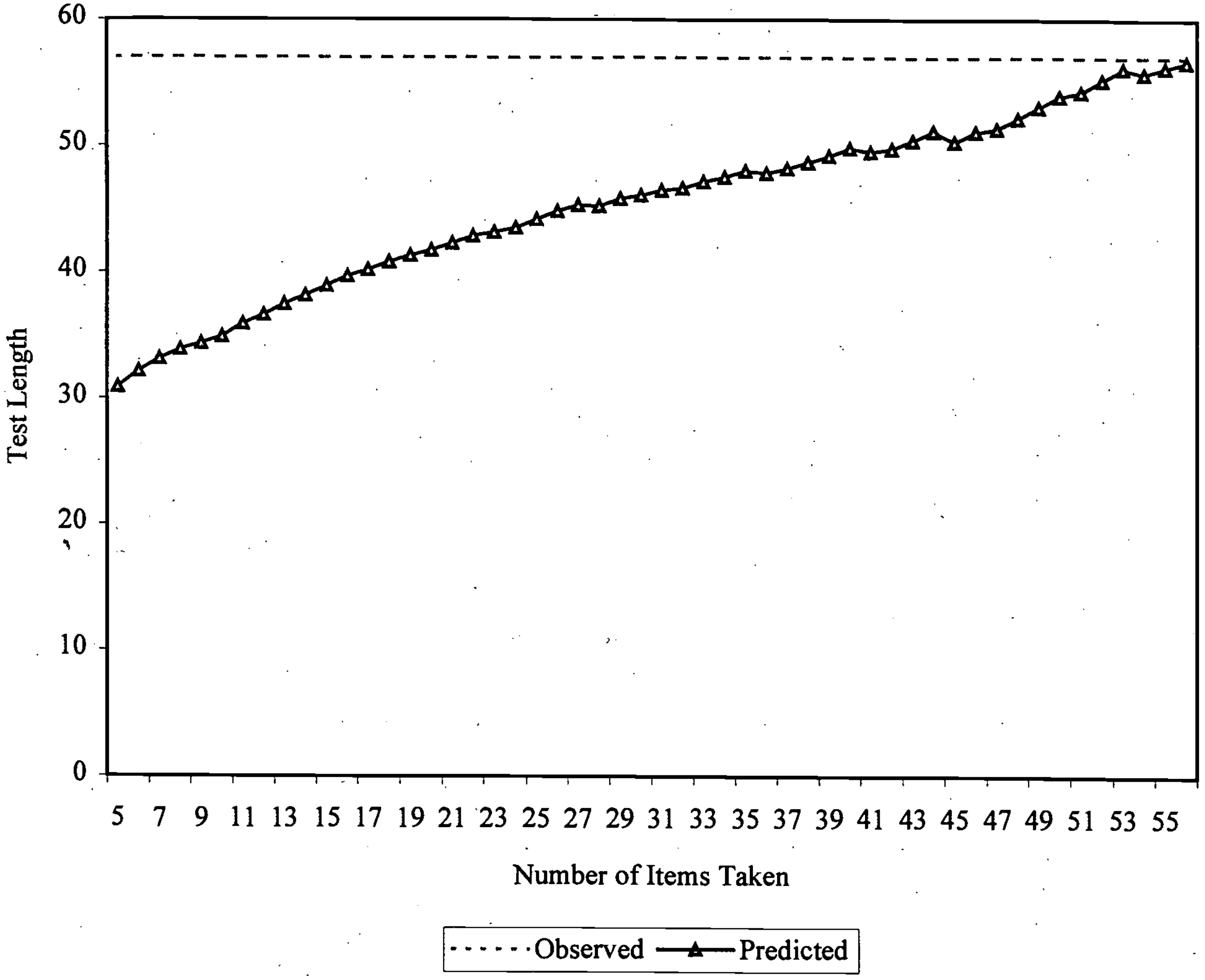


Figure 5. Predicted and Observed Final Test Lengths for an Examinee (Final Test Length = 57 Items).



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

TM029883

REPRODUCTION RELEASE

(Specific Document)

NCME

I. DOCUMENT IDENTIFICATION:

Title: Pretesting Alongside an Operational CAT

Author(s): Tim Davey, Mary Pommerich, Tony D. Thompson

Corporate Source:

Publication Date: NCME 1999 Presentation

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 at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

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)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

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.

Sign here, → please

Signature: Tim Davey

Printed Name/Position/Title: Tim Davey

Organization/Address:
ACT, Inc./2201 N. Dodge St., P.O.
Box 168, Iowa City, IA 52243

Telephone: 319/337-1359

FAX: 319/339-3021

E-Mail Address: davey@act.org

Date: 5/24/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.