DOCUMENT RESUME

ED 467 807 TM 034 349

AUTHOR Reese, Lynda M.; Schnipke, Deborah L.

TITLE An Evaluation of a Two-Stage Testlet Design for Computerized Testing. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

INSTITUTION Law School Admission Council, Princeton, NJ.
REPORT NO LSAC-R-96-04
PUB DATE 1999-03-00
NOTE 16p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; *Computer Assisted Testing; Difficulty Level; Estimation (Mathematics); Research Design; Simulation; Test Construction
IDENTIFIERS *Testlets

ABSTRACT

A two-stage design provides a way of roughly adapting item difficulty to test-taker ability. All test takers take a parallel stage-one test, and based on their scores, they are routed to tests of different difficulty levels in the second stage. This design provides some of the benefits of standard computer adaptive testing (CAT), such as increased precision of ability estimates over a paper-and-pencil design. In addition, the item selection and scoring algorithms in a two-stage design may be easier for test takers and test-result users to understand—an important feature for gaining public acceptance of new test designs. This simulation study incorporated testlets (or collections of items) into the two-stage design and compared the precision of the ability estimates derived from this design with those derived from a standard CAT design and from a paper-and-pencil test design. The results indicate that if the testlets are carefully assembled, a two-stage testlet design can produce more precise ability estimates in the middle ability range than those obtained from a paper-and-pencil design with twice as many items. Results of this study provide a baseline against which future research that incorporates content constraints can be compared. (Contains 6 figures, 5 tables, and 12 references.) (Author/SLD)

TM

An Evaluation of a Two-Stage Testlet Design For Computerized Testing

Lynda M. Reese and Deborah L. Schnipke
Law School Admission Council

Law School Admission Council
Computerized Testing Report 96-04
March 1999

LAW Services

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

In a standard computer-adaptive testing (CAT) design, test takers are first administered a test question of approximately medium difficulty. Based on their response, an attempt is made to choose subsequent items for administration to the test takers that are appropriate to their ability level. Testing proceeds until some termination criterion, such as a fixed test length or a sufficiently precise ability estimate, is achieved. In this pure form, CAT holds many theoretical advantages. Because the test taker's time is not wasted on test items that are too difficult or too easy, test length may be reduced, usually by about one half, without loss of precision.

As large-scale, high-stakes testing programs, such as the Law School Admission Test (LSAT), consider converting to a computer-adaptive mode of test administration, a standard computer-adaptive test, as described above, is rarely practical. For example, most large-scale testing programs contemplating CAT must face the challenge of maintaining content balancing requirements which usually compromise the efficiency and precision that make CAT attractive. Some researchers have advocated the use of testlets (or collections of items) as an alternative to individually selected and delivered items. These testlets may be preassembled to achieve certain content coverage requirements; employing these requirements may help to control context effects.

Prior to the advances in computer technology that made CAT feasible, the concept of two-stage testing emerged as a rudimentary means of tailoring the difficulty level of the test to the ability level of the test taker. In the first stage of this procedure, all test takers take a "routing test" of medium difficulty. Based on their scores on the routing test, test takers are branched to a second stage "measurement test" that is roughly adapted to their ability level. The test taker's ability is then estimated based on the items administered at both testing stages.

This simulation study evaluated the efficiency of a two-stage testlet design as compared with that achieved by a standard computer-adaptive test administration and a paper-and-pencil test. The precision of the ability estimates derived from a 25-item two-stage testlet design was compared to those estimates derived from a standard CAT of 25-items and both a 25-item and a 50-item paper-and-pencil test. The results indicate that if the testlets are carefully assembled, the 25-item two-stage testlet design results in greater precision than a 50-item paper-and-pencil test. This study provides a baseline against which future research that incorporates content constraints can be compared.

Abstract

A two-stage test design provides a way of roughly adapting item difficulty to test-taker ability. All test takers take a parallel stage-one test, based on their score they are routed to tests of different difficulty levels in the second stage. This design provides some of the benefits of standard computer-adaptive testing (CAT), such as increased precision of ability estimates over a paper-and-pencil test design. Additionally, the item selection and scoring algorithms in a two-stage design may be easier for test takers and test-result users to understand—an important feature for gaining public acceptance of new test designs. This study incorporates testlets (or collections of items) into the two-stage design and compares the precision of the ability estimates derived from this design with those derived from a standard CAT design and from a paper-and-pencil test design. The results indicate that if the testlets are carefully assembled, a two-stage testlet design can produce more precise ability estimates in the middle ability range than those obtained from a paper-and-pencil design with twice as many items. This study provides a baseline against which future research that incorporates content constraints can be compared.

Introduction

In computer-adaptive testing (CAT), items are selected for administration to test takers based on their responses to previous items on the test. In this way the difficulty level of the test is tailored to the ability level of the test taker. In a standard CAT, a test taker is first administered a test question of about medium difficulty. Based on the response, an attempt is made to choose subsequent items for administration to the test taker that are appropriate to the test taker's ability level. Testing proceeds until some termination criterion, such as a fixed test length or a

sufficiently precise ability estimate, is achieved. In this pure form, CAT holds many theoretical advantages. Because test-taker time is not wasted on test items too difficult or too easy, test length may be reduced, usually by about one half, while still maintaining the same level of precision (Weiss, 1982).

As large-scale, high-stakes testing programs consider converting to a computer-adaptive mode of test administration, a standard CAT, as described previously, is rarely practical. Most large-scale testing programs contemplating CAT must face the challenge of maintaining content balancing requirements that usually compromise the efficiency and precision that make CAT attractive. The selection of test items for administration to test takers within the CAT must be constrained in some way (Kingsbury & Zara, 1989). Researchers have proposed various ways of addressing the issue of content balancing in CAT. For example, in the weighted deviations model (Swanson & Stocking, 1993), psychometric and nonpsychometric constraints on item selection are defined, and an attempt is made to minimize deviations from the constraints. As an alternative, some researchers have advocated the use of bundles of items called testlets, rather than individually selected and delivered items (Thissen, Steinberg, & Mooney, 1989; Wainer, Kaplan, & Lewis, 1992; Wainer & Kiely, 1987; Wainer & Lewis, 1990; Wainer, Lewis, Kaplan, & Braswell, 1991). These testlets may be preassembled to achieve certain content coverage requirements and may help to control context effects.

Prior to the advances in computer technology that made CAT feasible, the concept of two-stage testing emerged as a rudimentary means of tailoring the difficulty level of the test to the ability level of the test taker. In the first stage of this procedure, all test takers take a "routing test" of medium difficulty. Based on their scores on the routing test, test takers are branched to a "measurement test" roughly adapted to their ability level. The test taker's ability is then estimated based on the items administered at both testing stages (Lord, 1980; Weiss, 1985). By tailoring item difficulty to test-taker ability in the second stage (the measurement test), large gains in precision or efficiency should be obtained, as compared with a standard paper-and-pencil test where all test takers receive the same items, regardless of ability.

In the application of CAT within a large-scale testing program, the two-stage test design holds some practical advantages. One issue to be addressed within the standard CAT is item exposure. A standard CAT design could overexpose the best items in the item pool, leading to security problems. If the stages are well-designed, a two-stage test could assure that the entire pool is being utilized, thus keeping exposure of each item to a minimum. Also, depending on how test takers are routed from one stage to the next, the two-stage design could lead to a simpler scoring model that is easier for test takers to understand.

The present simulation study applies the concept of a two-stage test design while incorporating testlets to facilitate future content constraint issues. The precision of the two-stage testlet design is compared with that achieved by both a typical paper-and-pencil design and a standard CAT design without content constraints. The goal for the two-stage testlet design would be to achieve precision as close as possible to a standard CAT design, with the paper-and-pencil design providing a minimum criterion for precision. The two-stage design should perform at least as well as the paper-and-pencil test of the same length which does not adapt item difficulty to test-taker ability at all.

Test Designs

Two-Stage Testlet Design

In the two-stage testlet design, bundles of items (i.e., testlets), rather than individual items, were selected for administration. Testlets were assigned to Stage One (the routing test) or to Stage Two (the measurement test). Testlets within Stage Two were further classified as "low," "medium," or "high," based on item difficulty (described below). Testlets were roughly parallel to other testlets of the same classification. Number-right score on Stage One served to route test takers to Stage Two where item difficulty more closely matched test-taker ability (e.g., high-difficulty testlets for high-ability test takers).

In Stage One, two testlets were randomly selected for each simulated test taker. The simulated test taker's number-right score was calculated and was used to route the simulated test taker to a low, medium, or high

Stage Two level (which scores were routed to each level is described below). In Stage Two, three testlets were randomly selected at the appropriate level (low, medium, or high, based on the Stage One number-right score) for each simulated test taker.

After all items were administered, the final θ estimate was calculated using Bayesian modal scoring (Hambleton, Swaminathan, & Rogers, 1991), based on all 25 responses, after a normal prior with a mean of 0 and a standard deviation of 1 was set for the θ distribution. (Bayesian modal scoring requires an initial θ estimate. The initial estimate was obtained with Owen's Bayesian sequential scoring (Owen, 1969) which updated the θ estimate after each item was administered.)

Routing in stage two based on stage one number-right score. The purpose of Stage Two is to tailor item difficulty more closely to test-taker ability. Matching item difficulty to test-taker ability maximally decreases measurement error for a fixed number of items. Thus, the level (low, medium, or high) of Stage Two that is expected to decrease measurement error the most for a given test taker is the one that should be administered to that test taker. We determined for each number-right score what the error would be if each level of Stage Two were administered separately.

Specifically, the mean squared error (*MSE*) of ability (θ) was used to determine which Stage One number-right scores would be routed to each level (low, medium, or high) of Stage Two. Test takers with a given Stage One number-right score should be routed to the Stage Two level that leads to the lowest *MSE* for that number-right score. To determine the cutoff scores for routing simulated test takers to Stage Two levels of low, medium, and high, all simulated test takers were first administered two Stage One, five-item testlets, and their number-right score was then calculated. Regardless of Stage One number-right score, each simulated test taker was administered three randomly selected low Stage Two testlets; and θ estimates were obtained. All simulated test takers were next administered three randomly selected medium Stage Two testlets, and new θ estimates were obtained using the responses to items on Stage One and the medium Stage Two testlets that were administered to the test taker. Finally, all simulated test takers were administered three randomly selected high Stage Two testlets, and a third θ estimate was obtained for each test taker using Stage One and the high Stage Two testlets.

MSE_s was calculated separately for the three θ estimates (one from each level of Stage Two) at each Stage One number-right score, *s*. MSE_s is given by

$$MSE_s = \frac{\sum_{j=1}^{N} (\theta_j - \hat{\theta}_j)^2}{N}$$

where

θ_j represents the true value of ability parameter for test taker *j*,
$\hat{\theta}_j$ represents the estimated ability value for test taker *j*, and
N is the number of simulated test takers who obtained a number-right score of *s*.

Figure 1 shows MSE_s values for one of the item pools (item pools are described under *Simulations*). Test takers who obtained a low Stage One number-right score (less than 6 correct) presumably have a low true θ, and a low Stage Two testlet leads to the lowest measurement error (*MSE*). Similarly, test takers who obtained a high Stage One number-right score (9 or more correct) presumably have a relatively high true θ, and a high Stage Two testlet leads to the lowest measurement error (*MSE*). The locations at which the low and medium and the medium and high lines cross determined the Stage One number-right cutoffs among the low, medium, and high levels. Similar analyses were carried out for each item pool studied, and the cutoff values for routing test takers at Stage Two turned out to be identical regardless of the item pool.

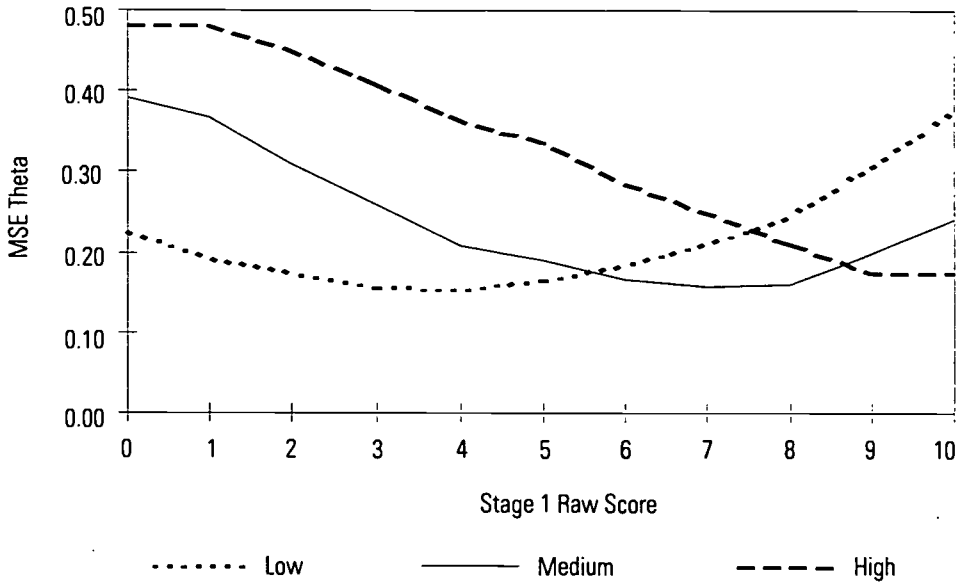

FIGURE 1. *Choosing number-right score cutoffs*

Maximum-Information (Standard CAT) Design

A standard maximum-information CAT (standard CAT) design was simulated to indicate how precise ability estimates could be if item difficulty were adapted to test-taker ability after every item. The design was based on item information, as specified by item response theory (IRT). Item information, $I_i(\theta)$, was calculated at ninety-seven θ values (from -3 to 3 in increments of 0.0625) for each item using the formula

$$I_i(\theta) = \frac{2.89a_i^2\ (1 - c_i)}{[c_i + e^{1.7a_i(\theta - b_i)}]\ [1 + e^{-1.7a_i(\theta - b_i)}]^2}$$

where

i indicates the item,
a is the IRT discrimination parameter for item *i*,
b is the IRT difficulty parameter for item *i*, and
c is the IRT lower asymptote parameter for item *i* (Hambleton, Swaminathan, & Rogers, 1991).

The information values were used to sort items at each θ-level. The sequence of items from highest to lowest information at each θ-level was saved in an information look-up table. This table was used during the simulations to select the items with the highest information at a given θ-level.

To prevent items from becoming "overexposed" (administered to too large a proportion of simulated test takers), a 10-9-8- ... exposure control method (Kingsbury & Zara, 1989) was incorporated into the simulations. The first item to be administered to a simulated test taker was randomly selected from the 10 items with the highest information values at $\theta = 0$ (the starting value for all simulated test takers). The second item was randomly selected from the nine best items at the new estimate of θ. The third item was randomly selected from the eight best items, and so on until, beginning with the tenth item, the item with the highest information was selected (unless, of course, the item had already been administered to that simulated test taker, in which case the next best item was selected).

As in the two-stage testlet design, after each item was selected, the simulated test taker's response (right/wrong) was determined, and the simulated test taker's estimated θ was updated using Owen's Bayesian sequential scoring (Owen, 1969). After all items were administered, a Bayesian modal score

(e.g., Hambleton, Swaminathan, & Rogers, 1991) was calculated and was used as the final θ estimate. A fixed-length 25-item CAT was simulated.

Paper-and-Pencil Design

In addition to the standard CAT design, the two-stage testlet design was compared with simulated paper-and-pencil tests of both 25 and 50 items. The items in the paper-and-pencil designs were taken from two intact Law School Admission Test (LSAT) sections. As a rule, a standard maximum-information CAT promises a test length reduction of approximately 50% over a paper-and-pencil test (Weiss, 1982). Thus we would expect the 25-item standard CAT to produce θ estimates that are as precise as those from a 50-item paper-and-pencil test. Of interest is how the 25-item two-stage testlet design compares with the 50-item paper-and-pencil test and the 25-item standard CAT. The two-stage testlet design of 25 items should perform at least as well as, if not better than, the 25-item paper-and-pencil test.

Simulations

Two groups of simulated test takers, one pool of real items, and two pools of simulated items were created, as described below. The two-stage testlet design was simulated using the real item pool and both simulated item pools, whereas the standard CAT was simulated using only the real item pool. The real item pool was taken from the form assembly pool for the LSAT logical reasoning (LR) item type. The 25- and 50-item paper-and-pencil tests were simulated using item parameters from intact LSAT LR test sections. The intact LR test sections were assembled from the same pool of items as the items drawn for the real pool, and thus met the same statistical criteria. The items on the intact sections were not included in the real item pool simulations.

Simulated Test Takers

Two groups of simulated test takers were created. One group was used to establish the cutoffs for the two-stage testlet design, and the other group was used for the simulations of all three test designs. For the group that was used to establish the cutoffs for the two-stage testlet design, 50,000 simulated test takers were created by randomly sampling ability (θ) parameters from a normal distribution with a mean of 0 and a standard deviation of 1. (A large sample size was chosen to establish the cutoffs to minimize chance variations that would affect the cutoffs.)

For the group used to simulate all three test designs, 7,000 simulated test takers were created by randomly sampling ability (θ) parameters from a normal distribution with a mean of 0 and a standard deviation of 1. (A more moderate sample size was chosen for the simulations to be closer in size to what we might obtain operationally.)

Real Item Pool

A pool of real items was specified by the item statistics from 538 LSAT LR items. This pool consisted of all currently developed LR items that met the statistical criteria for use in test assembly. The LR items were selected from among the item types available on the LSAT because the LR items are administered discretely, as opposed to analytical reasoning (AR) and reading comprehension (RC) items, which are administered in sets. Using discrete items allowed us to group items arbitrarily into testlets without worrying about the natural division of items into testlets that would occur with items that refer to a common stimulus. Table 1 describes the distribution of items among the stages and levels for all item pools utilized in this study. Descriptive statistics for the *a*, *b*, and *c* parameters for this item pool are shown in Table 2.

TABLE 1

Number of items for each stage and level for the real and simulated item pools

Stage & Level	Real Pool	Simulated Pool I	Simulated Pool II
Stage 1	298	663	500
Stage 2: Low	48	210	750
Stage 2: Medium	173	221	750
Stage 2: High	106	206	750

TABLE 2

Summary statistics of the IRT item parameters for the real item pool

Variable	N	Mean	*SD*	Minimum	Maximum
a	538	0.74	0.22	0.2596	1.5403
b	538	0.30	1.13	-2.9235	2.8910
c	538	0.17	0.11	0.0069	0.5000

Items were grouped into five-item testlets using the *a* and *b* values to categorize items into stages/levels. Where appropriate, we allowed items to contribute to more than one stage or level, but assured that the same item was never administered to a simulated test taker more than once. Items with *a*-parameter values less than .8 were classified as Stage One, regardless of the value of the *b*-parameter (298 items). For items with *a*-parameter values greater than .7, those with *b*-parameter values less than -.25 were classified as Low Stage Two (48 items), those with *b*-parameter values greater than .75 were classified as High Stage Two (106 items), and those with *b*-parameter values falling between these two cutoffs were classified as Medium Stage Two (173 items). We felt it was more suitable to use lower discriminating items (lower *a*'s) in Stage One when less is known about test-taker ability and the items are not necessarily of the appropriate difficulty level for the test taker. Testlets were assembled by randomly grouping five items from those classified into each stage/level. Figure 2 shows the *b*-parameter values for items in each testlet at each stage/level (indicated by different markers).

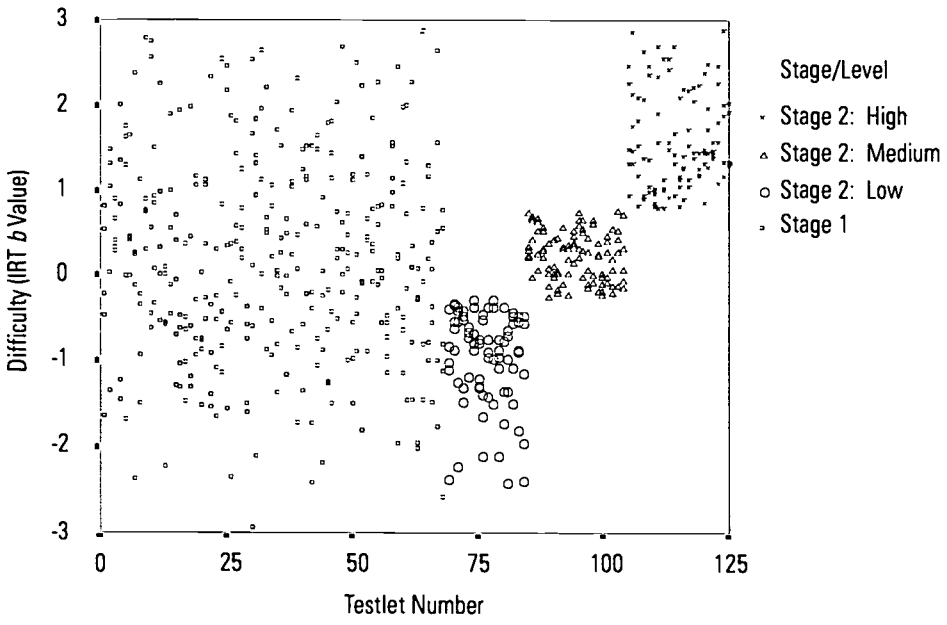

FIGURE 2. *Testlets in real item pool*

Simulated Item Pool I

A simulated item pool of 1,300 items was also created. The *a*-, *b*-, and *c*-parameter values were randomly sampled from the distributions indicated in Table 3.

TABLE 3

Description of the distributions of the IRT item parameters for simulated item pool I

Variable	Stage 1	Stage 2: Low	Stage 2: Medium	Stage 2: High
a	normal (.75, .25)	normal (.9, .25)	normal (.9, .25)	normal (.9, .25)
b	normal (0, .8)	normal (-1, .5)	normal (0, .5)	normal (1, .5)
c	uniform (.15, .25)	uniform (.15, .25)	uniform (.15, .25)	uniform (.15, .25)

Note. The values in parentheses represent the lower and upper ranges for the uniform distribution and the mean and *SD* for the normal distribution.

The *c* (lower asymptote) parameters were created to be roughly comparable to five-option multiple-choice items. The *b* (difficulty) parameters varied according to the stage/level to which the item was assigned. In all stages/levels, the *b*-parameter values were sampled from a normal distribution. The *b*-parameter values in Stage Two were centered at -1, 0, or 1 (for low, medium, and high) and had a smaller *SD* so that the item difficulty would be less spread out for a given level than the difficulty values at Stage One. Figure 3 shows the *b*-parameter values for items in each testlet at each stage/level (indicated by different markers).

The *a*- parameter values were sampled from a normal distribution and varied according to the stage the item was assigned. The *a*'s for Stage Two are higher on average than those for Stage One. As in the real pool, we felt that higher values of *a* were more appropriate in Stage Two when item difficulty is being tailored to simulated test-taker ability.

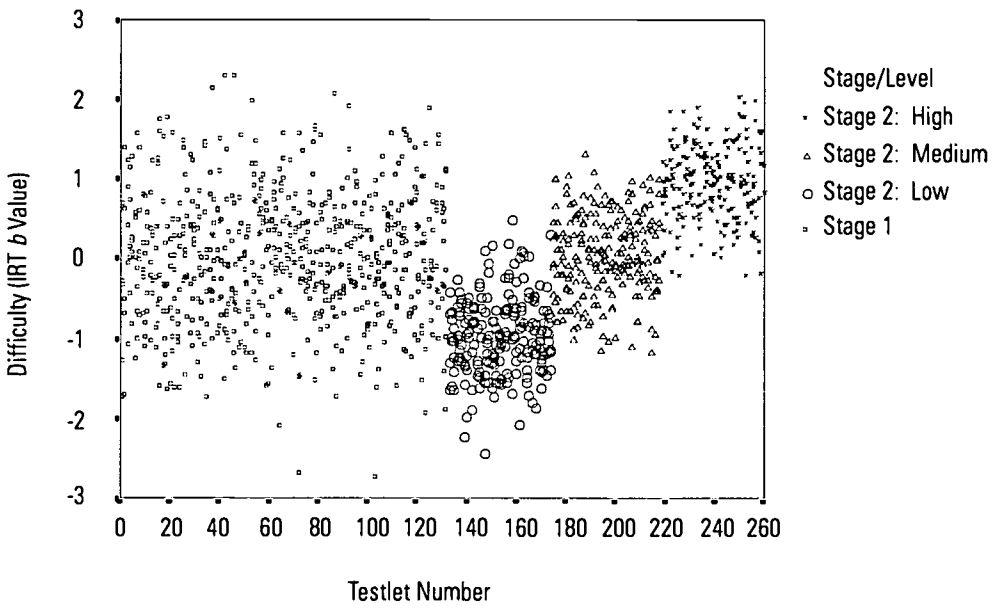

FIGURE 3. *Testlets in simulated item pool I*

Simulated Item Pool II

A second simulated item pool of 2,750 items was also created. As will be described more fully in the Results section, the need for a second simulated pool, in which the Stage One testlets were somewhat easier and the testlets were assembled in a more constrained way, became apparent when the results obtained with the real item pool and simulated pool I were examined. Table 4 describes the distributions used to assemble this item pool.

TABLE 4

Description of the distributions of the IRT item parameters for simulated item pool II

Variable	Stage 1	Stage 2: Low	Stage 2: Medium	Stage 2: High
a	normal (.80, .22)	normal (.90, .22)	normal (.90, .22)	normal (.90, .22)
b	normal (-.50, .80)	normal (-1, .80)	normal (0, .80)	normal (1.0, .80)
c	uniform (.15, .25)	uniform (.15, .25)	uniform (.15, .25)	uniform (.15, .25)

Note. The values in parentheses represent the lower and upper ranges for the uniform distribution and the mean and *SD* for the normal distribution.

Primary emphasis was placed on the assembly of the testlets for this item pool. For each stage/level, item parameters were generated for one testlet at a time, beginning with the *b*-parameter. The *b*-parameter values were generated for a five-item testlet by selecting randomly from the distributions indicated in Table 4 and assuring that the difference between the lowest and highest *b*-parameter value for that testlet ranged from 1.5 to 2.0, and that the mean of the *b*-parameter for that testlet was within .3 of the mean values specified in Table 4. Any testlet that did not meet these requirements was rejected. This ensured that the testlets would be centered near the specified mean and would have a range of *b* values that was consistent across testlets, thus creating testlets more parallel to one another than the testlets in the real pool and those in simulated pool I. Once the *b*-parameter values were generated satisfactorily, *a*- and *c*-parameter values were generated as specified in Table 4. The *b*-parameter values for items in each testlet at each stage/level (indicated by different markers) are displayed in Figure 4.

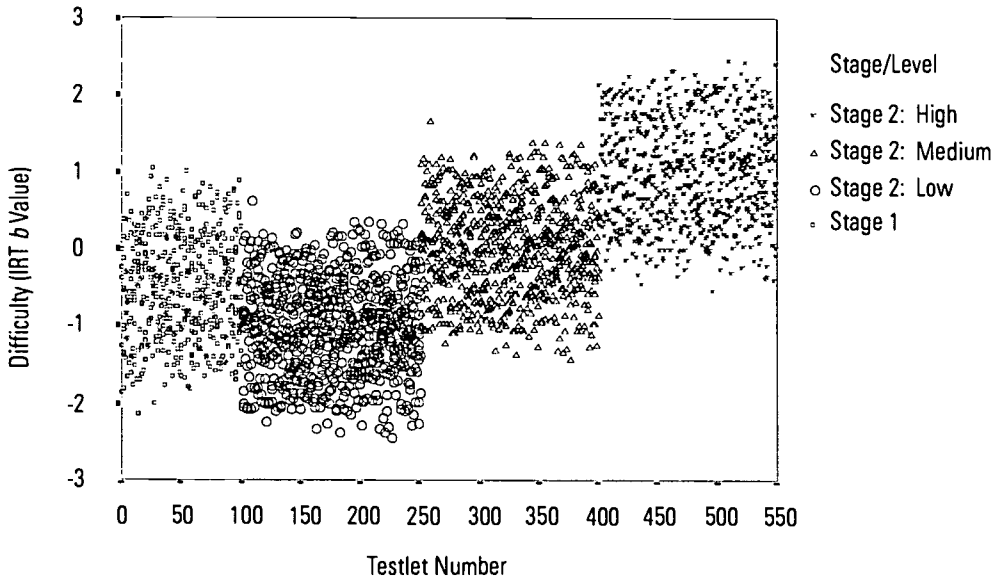

FIGURE 4. *Testlets in simulated item pool II*

As described for simulated pool I, the *c* (lower asymptote) parameters were created to be roughly comparable to 5-option multiple-choice items. The *b* (difficulty) parameters varied according to the stage/level to which the item was assigned. The *b*-parameter values in Stage One were centered at -.50, which is slightly easier than the other item pools. As in simulated pool I, the *b*-parameter values in Stage Two were centered at 1, 0, or 1 (for low, medium, and high), and the *a*'s for Stage Two were higher on average than those for Stage One. As indicated in Table 4, the standard deviation of the *b*-parameter for all stages/levels was defined to be .80. However, when further restrictions were imposed on parameter generation, the actual standard deviation for the *b*-parameter ranged from .62 to .66 for the various stages/levels.

Analyses

To indicate the amount of error in the ability estimates, the root mean squared error (*RMSE*) was calculated and plotted for each test design along the θ scale. To indicate whether ability is over- or underestimated, the bias statistic was computed and plotted along the θ scale.

To calculate both *RMSE* and bias, the θ's were grouped into intervals (<-2, -2 to -1, -1 to 0, 0 to 1, 1 to 2, and >2). *RMSE* and bias were calculated for all simulated test takers in each of these intervals, and the values were plotted to show how *RMSE* and bias vary across θ.

The *RMSE* is given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\theta - \hat{\theta})^2}{N}},$$

where

θ represents the true value of ability parameter,
$\hat{\theta}$ represents the estimated ability value,
i represents the item, and
N represents the number of simulated test takers in a particular θ interval.

The bias statistic was calculated similarly for each test design and θ interval and is given by

$$Bias = \frac{\sum_{i=1}^{N} (\theta - \hat{\theta})}{N}.$$

Positive bias values indicate that ability was underestimated; negative values indicate ability was overestimated.

Results

RMSE is shown in Figure 5, revealing that at the extremes of the ability scale, the 25-item standard CAT produced θ estimates with less error than any of the other test designs including the 50-item paper-and-pencil design. The 50-item paper-and-pencil design had less error than the 25-item paper-and-pencil design, as expected. Surprisingly, the 25-item two-stage testlet design with the real pool and simulated pool I performed no better than the 25-item paper-and-pencil design. However, when simulated pool II was used, the 25-item two-stage testlet design performed quite well, producing estimates with less error in the medium ability range than the 50-item paper-and-pencil design and the standard CAT.

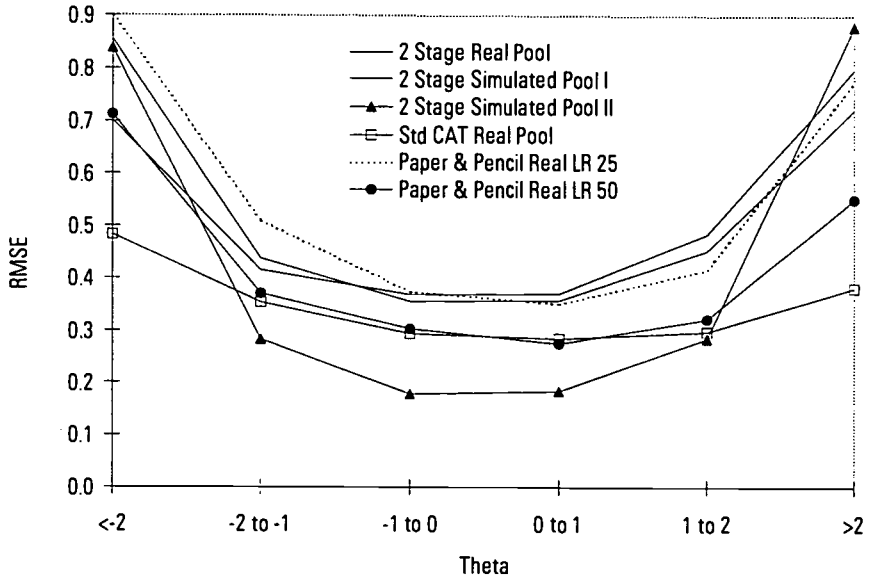

FIGURE 5. *Comparing test designs: RMSE*

The results for the bias statistic are presented in Figure 6. Here, the standard CAT outperformed all other test designs simulated, with very little bias throughout the entire ability range. The two-stage testlet design for all item pools performed similarly to the 25-item paper-and-pencil test.

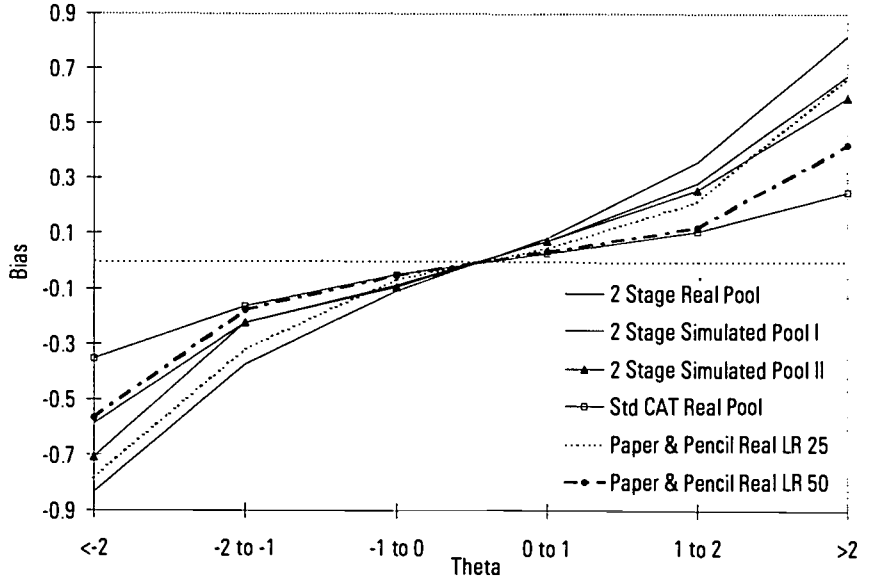

FIGURE 6. *Comparing test designs: Bias*

TABLE 5

Number of simulated test takers routed to each stage two level

Item Pool	Low Difficulty	Medium Difficulty	High Difficulty
Real Pool	5,024	1,794	182
Simulated Pool I	4,146	2,351	485
Simulated Pool II	2,303	2,358	2,339

Table 5 shows the number of simulated test takers routed to each level of Stage Two for the three item pools. It is clear that test takers were not evenly distributed to the levels with the real pool or simulated pool I; very few test takers were routed to the high difficulty group. For this reason, Stage One was shifted to a mean difficulty of -0.5 (rather than 0.0) when the revised simulated pool (pool II) was created. Examinees were distributed more evenly among the three Stage Two levels when simulated pool II was used.

Conclusions and Future Directions

With the two-stage testlet design, how the testlets were constructed and the average difficulty of Stage One greatly affected the precision of the θ estimates. If testlets are carefully constructed, one of the major advantages of CAT—improved precision with a reduction in test length—can be achieved with a two-stage testlet design; at the center of the ability scale, the precision of the standard CAT was even exceeded. Additionally, many of the other advantages of CAT, such as on-demand testing and the possibility of innovative item types, would also be possible with the two-stage testlet design. The use of testlets would also allow for simpler content coverage, a factor that will generally compromise the precision of a standard CAT design. By using a two-stage design, it may be possible to implement a scoring scheme and an item selection method that are easier for test takers, the test-result users, to understand. This may facilitate public acceptance of computerized testing.

In the two-stage testlet design studied here, simulated test takers were not permitted to change levels within the Stage Two test. An extension of this study is planned in which simulated test takers who appear to be misclassified will be allowed to change levels within the Stage Two test. This modification should result in improved precision, especially at the extreme ability values where the two-stage testlet design produced ability estimates that contained more error.

In closing, it was stated at the outset that this study would provide a baseline against which future modifications could be compared. We feel certain that with minor modifications to the model, greater precision will be achieved. Given the many advantages to this design, such future research is certainly warranted.

References

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications, Inc.

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

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

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report RB-69-92). Princeton, NJ: Educational Testing Service.

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

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247-260.

Wainer, H., Kaplan, B., & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement, 29,* 243-251.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., Lewis, C., Kaplan, B., & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. *Journal of Educational Measurement, 28,* 311-323.

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

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Counseling and Clinical Psychology, 53,* 774-789.

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

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").