DOCUMENT RESUME

ED 467 816 TM 034 360

AUTHOR Reese, Lynda M.; Schnipke, Deborah L.; Luebke, Stephen W.
TITLE Incorporating Content Constraints into a Multi-Stage Adaptive Testlet Design. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION Law School Admission Council, Princeton, NJ.
REPORT NO LSAC-R-97-02
PUB DATE 1999-08-00
NOTE 18p.
PUB TYPE Reports - Research (143)
EDRS PRICE EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Computer Assisted Testing; Simulation; *Test Construction; Testing Programs
IDENTIFIERS *Constraints; *Testlets

ABSTRACT

Most large-scale testing programs facing computerized adaptive testing (CAT) must face the challenge of maintaining extensive content requirements, but content constraints in computerized adaptive testing (CAT) can compromise the precision and efficiency that could be achieved by a pure maximum information adaptive testing algorithm. This simulation study first evaluated whether realistic content constraints could be met by carefully assembling testlets and appropriately selecting testlets for each test taker that, when combined, would meet the content requirements of the test and would be adapted to the test taker's ability level. The second focus of the study was to compare the precision of the content-balanced testlet design with that achieved by the current paper-and-pencil version of the test through data simulation. The results reveal that constraints to control for item exposure, testlet overlap, and efficient pool utilization need to be incorporated into the testlet assembly algorithm. More refinement of the statistical constraints for testlet assembly is also necessary. However, even for this preliminary attempt at assembling content-balanced testlets, the two-stage computerized test simulated with these testlets performed quite well. (Contains 5 figures, 5 tables, and 12 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

TM

Incorporating Content Constraints into a Multi-Stage Adaptive Testlet Design

Lynda M. Reese, Deborah L. Schnipke, and Stephen W. Luebke
Law School Admission Council

Law School Admission Council
Computerized Testing Report 97-02
August 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.

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 ;y of the Law School Admission Council.

Table of Contents

Executive Summary

In a standard computerized adaptive test (CAT) design, test takers are first administered a test question of approximately middle difficulty. Based on their response, an attempt is made to choose subsequent items for administration that are more appropriate for 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 LSAT consider converting to a computerized adaptive mode of test administration, a standard CAT, as described above, is rarely practical. 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. Other concerns about a CAT include how to deal with set-bound items (items that refer to a common stimulus) and whether to allow item review (i.e., allow test takers to change previous responses). Efficient utilization of the item pool is also a concern when developing a CAT design. 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 pre-assembled to achieve certain content coverage requirements. Testlets may also facilitate administering set-bound items, allowing item review, and efficient item pool utilization.

This study first evaluated whether realistic content constraints could be met by carefully assembling testlets and appropriately selecting testlets for each test taker that, when combined, would meet the content requirements of the test and would be adapted to the test taker's ability level. Second, the precision of the content balanced testlet design was compared with that achieved by the current paper-and-pencil version of the test through data simulation. The results revealed that constraints to control for item exposure, testlet overlap, and efficient pool utilization need to be incorporated into the testlet assembly algorithm. More refinement of the statistical constraints for testlet assembly are also necessary. However, even for this preliminary attempt at assembling content-balanced testlets, the two-stage computerized test simulated with these testlets performed quite well.

Abstract

Most large-scale testing programs contemplating computerized adaptive testing (CAT) must face the challenge of maintaining extensive content requirements. However, content constraints can compromise the precision and efficiency that could be achieved by a pure maximum information adaptive testing algorithm. Other concerns about a CAT include how to deal with set-bound items (items that refer to a common stimulus) and whether to allow item review (i.e., allow test takers to change previous responses). Efficient utilization of the item pool is also a concern when developing a CAT design. Building a test from testlets (bundles of items) may facilitate managing content constraints, administering set-bound items, allowing item review, and efficient item pool utilization. This study first evaluated whether realistic content constraints could be met by carefully assembling testlets and appropriately selecting testlets for each test taker that, when combined, would meet the content requirements of the test and would be adapted to the test taker's ability level. Second, the precision of the content-balanced testlet design was compared with that achieved by the current paper-and-pencil version of the test through data simulation. The results revealed that constraints to control for item exposure, testlet overlap, and efficient pool utilization need to be incorporated into the testlet assembly algorithm. More refinement of the statistical constraints for testlet assembly is also necessary. However, even for this preliminary attempt at assembling content-balanced testlets, the two-stage computerized test simulated with these testlets performed quite well.

Introduction

Most large-scale testing programs contemplating computerized adaptive testing (CAT) must face the challenge of maintaining extensive content requirements. Content constraints are a reality for any testing program, and incorporating them into an adaptive test design is mandatory for an operational test (Kingsbury & Zara, 1989). However, content constraints can compromise the precision and efficiency that could be achieved by a pure maximum information adaptive testing algorithm, although the CAT will still perform better than a non-adaptive (e.g., linear paper-and-pencil) test. Other concerns about a CAT include how to deal with set-bound items (items that refer to a common stimulus) and whether to allow item review (i.e., allow test takers to change previous responses). Efficient utilization of the item pool is also a concern when developing a CAT design.

Building a test from testlets (bundles of items) may facilitate managing content constraints, administering set-bound items, allowing item review, and efficient item pool utilization. Testlets are particularly appealing when items naturally occur in sets (e.g., reading comprehension items that refer to a common reading passage), as is the case for half of the items on the current paper-and-pencil version of the test being considered here. Testlets may be pre-assembled to achieve certain content coverage requirements within the testlet. By administering appropriate combinations of carefully assembled testlets, the content constraints for the test can be satisfied. Rather than having a complicated item selection algorithm that incorporates many constraints (e.g., Stocking & Swanson, 1992), a "testlet selection algorithm" could be much simpler because many of the constraints would be incorporated into the testlet construction. With regard to set-bound items, a testlet can be constructed for each set, thereby solving the issue of how to administer the set-bound items. In terms of item review, because the items within a testlet are preselected, allowing item review within a testlet will not cause any difficulties (see Stocking, 1996). Finally, with regard to efficient pool utilization, items that individually might not contribute to an adaptive test, perhaps because of a low a-parameter value, might be part of an efficient testlet.

Because of the benefits of using testlets, an adaptive testlet design is being contemplated for the Law School Admission Test (LSAT). The purpose of the present study is to evaluate whether realistic content constraints can be incorporated into a two-stage adaptive testlet-based test design while still achieving the reduced test length and improved efficiency promised by adaptive testing. Previous research (Reese & Schnipke, 1999; Schnipke & Reese, 1997) indicates that a two-stage testlet design may be appropriate for a computerized LSAT.

This study first evaluated whether realistic content constraints could be met by carefully assembling testlets and appropriately selecting testlets for each examinee (test taker) that, when combined, would meet the content requirements of the test and would be adapted to the examinee's ability level. Second, the precision of the content-balanced testlet design was compared with that achieved by the current paper-and-pencil version of the test through data simulation.

Test Designs

Content-Balanced Two-Stage Testlet Design

In the two-stage testlet design, bundles of items (i.e., testlets) were selected for administration, rather than individual items. Testlets were assigned to Stage 1 (the routing test) or Stage 2 (the measurement test), and within Stage 2, testlets were further classified as "low," "medium," or "high," difficulty (testlet assembly is described more fully below). Testlets were roughly parallel to other testlets of the same classification.

The authors wish to acknowledge the programming support of Jennifer Lawlor. The support of Ronald D. Armstrong of Rutgers University who modified the existing Armstrong-Jones test assembly software to accommodate testlet assembly is also gratefully acknowledged.

This research was collaborative in every respect, and the order of authorship is arbitrary.

Number-right score on Stage 1 served to route examinees to Stage 2 where item difficulty more closely matched examinee ability (e.g., high difficulty testlets for high ability examinees).

In Stage 1, two five-item testlets were randomly selected for each simulated examinee. The simulated examinee's number-right score was calculated and was used to route the simulated examinee to a low, medium, or high Stage 2 level. Simulated examinees with a Stage 1 score between zero and six were routed to a low Stage 2 level, simulated examinees with Stage 1 scores of seven or eight were routed to a medium Stage 2 level, and simulated examinees with a Stage 1 score of nine or ten were routed to a high Stage 2 level. In Stage 2, three testlets were randomly selected at the appropriate level (low, medium, or high, based on the Stage 1 number-right score) for each simulated examinee. Simulated examinees were rerouted to another level within the Stage 2 test if it was determined that they had been misclassified. Since some amount of item overlap between testlets was allowed, the algorithm assured that no simulated examinee was administered the same item twice.

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). See Schnipke and Reese (1997) for a more detailed description of this test design.

Paper-and-Pencil Design

The results obtained with the content-balanced two-stage testlet design were compared with simulated paper-and-pencil tests of both 25 and 51 items. The items in the paper-and-pencil designs were taken from two intact LSAT sections, one consisting of 25 items and the other consisting of 26 items. As a rule, a standard maximum information CAT promises a test length reduction of approximately 50% over a paper-and-pencil test (Weiss, 1982). Foong and Lam (1991) reported a test length reduction for two-stage testing to one third of the length of the original conventional test. Applying a two-stage testlet design and incorporating content constraints will undoubtedly compromise some of this efficiency. However, we expect this design to be an improvement over the full-length paper-and-pencil test.

Testlet-Level Content and Statistical Constraints

In creating a pool of testlets for the content-balanced two-stage testlet design, testlets were assembled to achieve content coverage, as well as to match target testlet information functions. Restrictions on word count were also imposed. Each of these criteria is described more fully in this section.

Content Constraints

Applying reasonable content constraints to the assembly of testlets was a primary objective of this research. In developing the testlet-level content constraints, the goal was to produce a set of testlet constraints that, in combination, would result in a content-balanced total test. Content constraints would function to assure broad coverage of a subject matter or, as in the case of the LSAT, to assure a balanced representation of item subtypes (kinds of questions within an item type). The constraints would also serve to assure that the content or distribution of item subtypes is parallel from examinee to examinee. As a means of achieving these objectives, three different schemes for content balancing were developed, and these schemes were compared to determine which could be most reasonably applied to arrive at a final content-balanced version of the test. These schemes will be referred to as the two testlet type scheme, the three testlet type scheme, and the five testlet type scheme. In each scheme, testlets are comprised of five items each, and a test is comprised of five testlets. A description of how the content requirements for the total test were derived is provided below, as well as a description of each of the content-balanced testlet schemes.

Test-level content constraints. In developing content constraints at the testlet level, the ultimate goal was to produce a content-balanced total test through the selection of an appropriate combination of content-balanced testlets. This study is restricted to the logical reasoning section of the LSAT. A 25-item logical reasoning section represents approximately a 50% reduction over the current full-length paper-and-pencil version of the test. The content constraints of the 25-item test were designed to reflect the complexity of item subtype distribution found in the paper-and-pencil version of the test.

Two testlet type scheme. In this first content balancing scheme, two different sets of testlet-level content constraints were defined. In order to produce a content-balanced total test, each examinee would be administered one Type A testlet and four Type B testlets. This scheme of content balancing would produce an acceptable total test.

Three testlet type scheme. The second content balancing scheme combines three different sets of testlet-level content constraints. By administering one Type A* testlet, one Type B* testlet and three Type C* testlets, a total test with an acceptable balance of content would be produced.

Five testlet type scheme. The final scheme requires the assembly of five different testlet types. Each examinee would be administered one testlet each of Type A** through Type E**. An acceptable content balance over the total test would be achieved by applying this scheme.

Word Count Restriction

In addition to the content balancing requirements, maintaining a balanced word count over the total test was also a concern. Each logical reasoning item begins with a short passage of stimulus material, and the length of these stimuli vary. Without monitoring the word count, it is possible that the reading load expected of different examinees could be unbalanced. To prevent unbalanced reading loads, a word count restriction was applied to all content-balancing schemes. The word count was restricted to 275 to 325 words per testlet.

Target Testlet Information Functions

In addition to the content constraints described above, a target testlet information function was defined for Stage 1 and for each of the three Stage 2 levels. This was accomplished through data simulation. Testlets were assembled from a simulated pool of 2,750 items. Table 1 describes the distributions from which the simulated item pool was assembled. In the creation of this item pool, primary emphasis was placed on the assembly of the testlets. For each stage/level, item parameters were generated for one testlet at a time, beginning with the *b*-parameter. *B*-parameter values were generated for a five-item testlet by selecting randomly from the distributions indicated in Table 1 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 1. 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 that were essentially parallel to one another. Once the *b*-parameter values were generated satisfactorily, *a*- and *c*-parameter values were generated as specified in Table1. The *b*-parameter values for items in each testlet at each stage/level (indicated by different markers) are displayed in Figure 1.

TABLE 1
Description of simulated item pool used in establishing target score distributions

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.

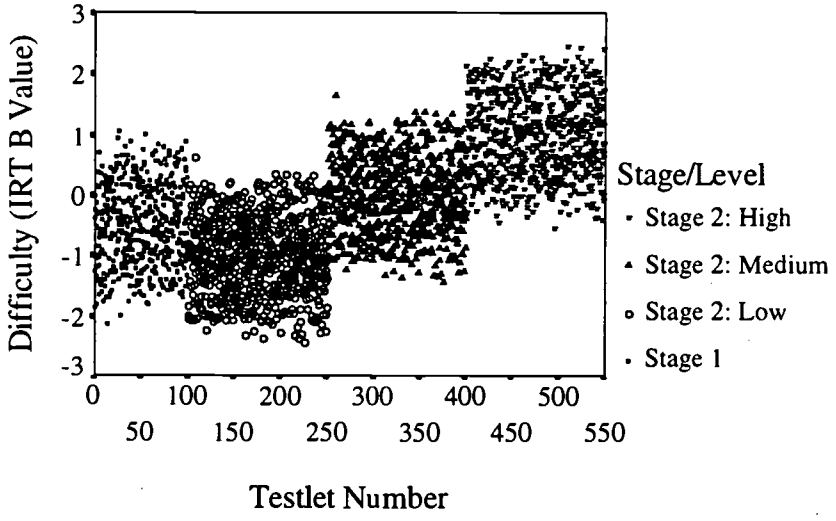

FIGURE 1. *Testlets in simulated item pool*

After all possible testlets were assembled, information functions were calculated for each testlet. This was accomplished by first calculating item information, $I_i(\theta)$, at each of 97 θ values (from -3 to 3 in increments of 0.0625) for each item using the formula

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

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

Testlet information was derived by summing over item information for each item in a testlet. Next, the information functions for each stage/level were averaged. This average information function was treated as the lower bound for the information target. The upper bound was derived by increasing the lower bound by 22.22%. This represents the same proportional difference between the lower and upper information bounds currently utilized for the assembly of the paper-and-pencil forms of the LSAT. The target information functions are presented in Figure 2.

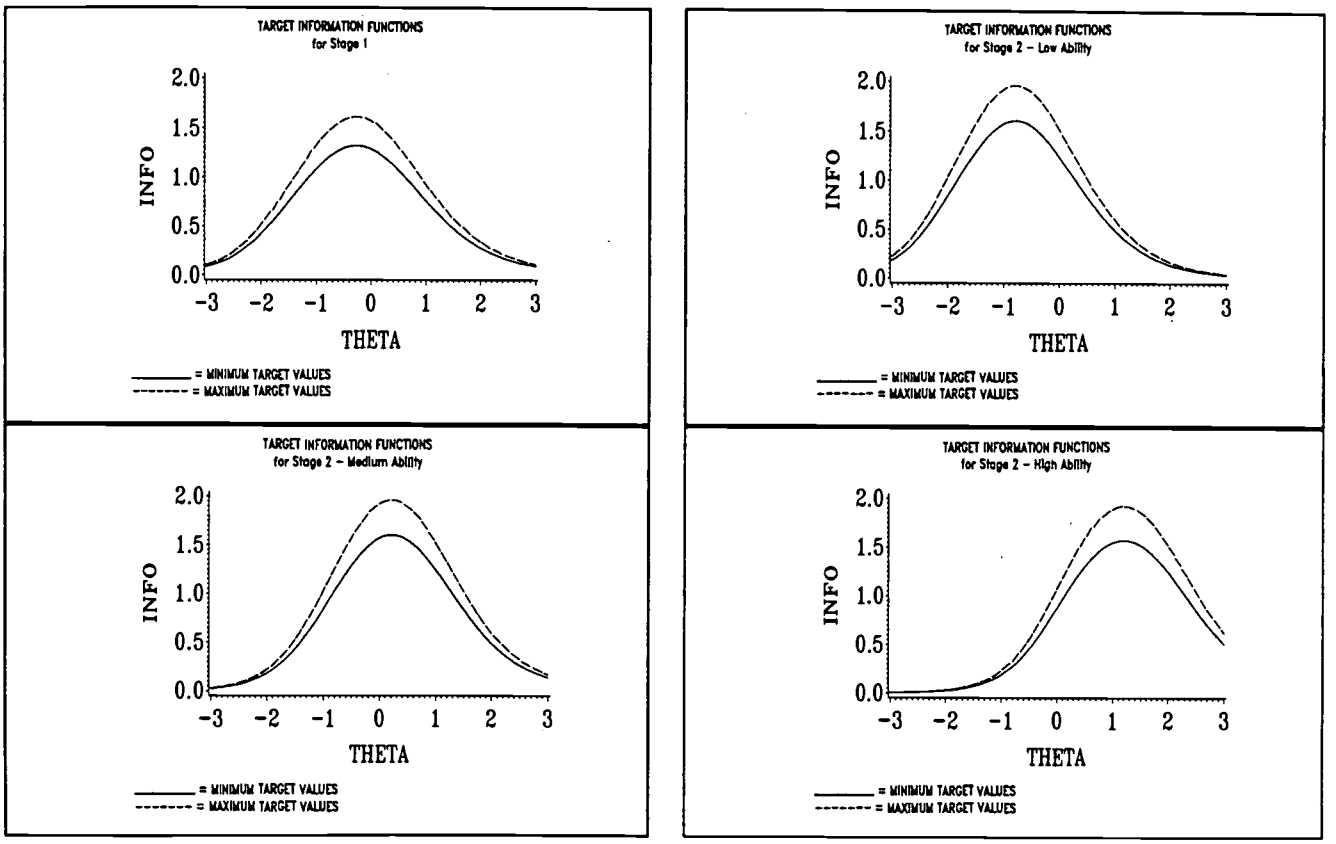

FIGURE 2. *Target testlet information functions*

Data Simulation

Simulations of the content-balanced two-stage testlet design and the paper-and-pencil tests of 25 and 51 items were carried out using a pool of simulated examinees and a pool of real item parameters. These pools are described more fully in this section.

Simulated Examinees

So that the quality of the design could be assessed across the ability range, ability values (θ's) were generated for 1,000 simulated examinees at each θ level from –3 to 3 in increments of 0.25. This resulted in a pool of 25,000 simulated examinees.

Item Pool

Item parameters were taken from the item bank of all usable logical reasoning items for the LSAT. (Usable means that the items met the statistical criteria for use in test assembly.) This initial attempt at CAT content balancing concentrated solely on the logical reasoning item type because this item type is composed of discrete items, as opposed to analytical reasoning and reading comprehension items which are administered in sets. The set-bound nature of the analytical reasoning and reading comprehension item types creates some additional issues to be addressed in testlet assembly, and these item types will be the subject of a follow-up study. Descriptive statistics for the 1,130 logical reasoning items used in this study are presented in Table 2.

TABLE 2
Descriptive statistics for the pool of logical reasoning items

Variable	*N*	Mean	Standard Deviation	Minimum	Maximum
a	1,130	0.758	0.253	0.257	1.684
b	1,130	−0.019	1.109	−2.924	2.891
c	1,130	0.168	0.106	0.000	0.693

Testlet Assembly

Testlets were assembled from the logical reasoning item pool described above. The mathematical programming model used to describe the problem of assembling a single testlet is the same model found in Armstrong, Jones, and Kunce (1996) for assembling a complete paper-and-pencil test form. Five-item testlets of each content type (A, B, A*, B*, C*, and A** through E**) were assembled to match the target information function for each stage/level. An exact target for the information function was derived by defining the information function that falls between the upper and lower bounds of the target.

The algorithm transforms the content constraints into a network flow representation. A Lagrangian Relaxation (see Fisher, 1981) approach was used to force the testlet information function toward the target function. The word count constraint was enforced through a branch-and-bound procedure (see Nemhauser & Wolsey, 1988). The deviations from the target information function were computed at 21 evenly spaced points on the ability scale between −3.0 and 3.0. The acceptability of the solution was measured based on the sum of weighted squared deviations from the target. The weights used in the study were values from the target information function. The most important target information criteria were that (1) the minimum bound on the information function be met and (2) the peak of the information function occur at the desired point along the ability scale. Weighting the deviations by the target information function places the primary emphasis on the peak of the function. A testlet was considered acceptable if the sum of weighted squared deviations from the information target was less than 0.5. The content specification and word count restriction were met exactly for every testlet. The testlet assembly process continued until a specified number of acceptable testlets were found.

Analyses

The two-stage testlet design described above was simulated using the two testlet type content-balancing scheme. This particular content-balancing scheme was chosen for the simulations because it is the simplest design among the three studied. All other things being equal, the simplest design is preferred. Simulations of the 25- and 51-item paper-and-pencil tests were also carried out. To indicate the amount of error in the ability estimates, the root mean squared error (*RMSE*) was calculated comparing the ability estimates ($\hat{\theta}$) derived for the two-stage design and each of the paper-and-pencil designs to the true ability values (θ). To indicate whether ability is over- or underestimated, the *bias* statistic was also computed and plotted along the θ scale.

RMSE and *bias* were calculated for all simulated examinees at each θ level (from −3 to 3 in increments of .25), and the values were plotted to show how *RMSE* and *bias* vary across θ.

RMSE is given by

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(\theta - \hat{\theta})^2]^{1/2}, \quad (2)$$

where n represents the number of items.

The *bias* statistic was calculated similarly by applying the equation

$$bias = \frac{1}{n}\sum_{i=1}^{n}(\theta - \hat{\theta}). \quad (3)$$

Results

Evaluation of Testlet Assembly

Using the testlet assembly approach described above, 101 testlets were assembled for each content specification (A, B, A*, B*, C*, A** - E**) for each stage/level. The algorithm assured that the content constraints and the word count restriction were always met exactly. Acceptability of the testlet information function was determined by the weighted squared deviation described previously. Testlet information functions were studied for all testlets assembled. Figure 3 displays some typical testlet information functions, along with the corresponding targets. The testlets in the left-hand column of Figure 3, numbered 1 through 5, were considered to be acceptable, while those in the right-hand column, numbered 6 through 10, were considered to be unacceptable. The ten testlets displayed here demonstrate what were considered to be the most important features for an acceptable testlet. The primary concerns were that the minimum of the target information function be met at all points along the ability scale and that the peak of the information curve for the testlet occurs in the desired region of the ability scale. Curves that exceeded the maximum were considered to be acceptable. A peak that was shifted too much to the left or the right was not considered to be acceptable.

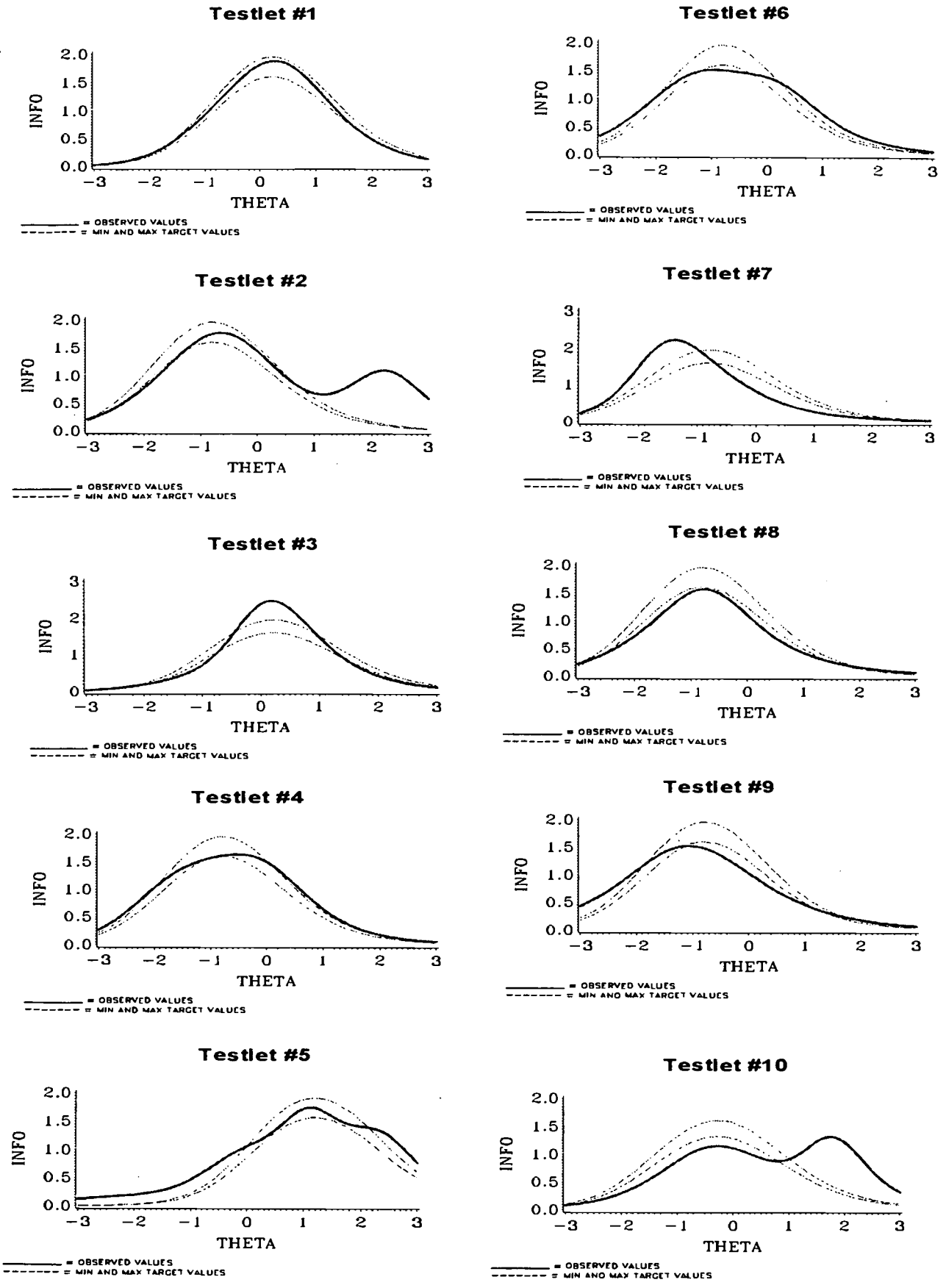

FIGURE 3. *Examples of acceptable (left column) and unacceptable (right column) testlets*

Applying this strict criterion to the evaluation of the testlets, a large proportion of testlets assembled for this study would have been lost. However, since one of the goals of this research was to establish reasonable criteria for assembling content-balanced testlets, no testlets were deleted from the pool based on this subjective evaluation criteria. Testlets consisting of only five items each cannot be expected to meet the same strict assembly criteria as a total test made up of 100 items. After studying the information curves, there was not a content-balancing scheme from among the three studied that allowed the information targets to be met more closely.

The procedure described above for assembling testlets did not control for overlap between testlets or exposure of items, although restrictions will be added to the testlet assembly algorithm to control for overlap and item exposure. For pairs of testlets with overlap of four items or more, one testlet from the pair was randomly selected and removed from the pool. This resulted in the testlet pools described in Table 3. The number of testlets remaining in the pool for the various content types at the various stages/levels ranged from 61 to 77. Again, no particular content-balancing scheme resulted in less item overlap among testlets. Note that overlap was studied within each stage/level for each content type and not across stages and levels or across content types within a content-balancing scheme. When item overlap controls are incorporated into the testlet assembly, this issue will be explored more thoroughly.

Analyses were also run to determine the degree of item exposure for each content type at each stage/level. Note from Table 3 that overall, the highest item exposure occurs for low and high testlets for every content type. This is not surprising, since there are fewer items in the item pool at these extremes. Testlet Types B, C*, and E**, all representing identical content, have the lowest overall item exposure. Testlet Type A displays the most extreme item exposure, with one item contributing to 29 testlets. Note that only one item had this extreme amount of exposure, and the item with the next highest exposure appeared in 16 testlets. Again, this analysis did not serve to eliminate any of the content-balancing schemes from the study.

TABLE 3

Number of testlets (and maximum item exposure) by testlet type and stage/level

Stage/Level	Testlet Type A	Testlet Type B
Stage 1	68 (8)	66 (7)
Stage 2—Low	69 (29)	66 (9)
Stage 2—Medium	69 (13)	71 (6)
Stage 2—High	68 (16)	75 (7)

	Testlet Type A*	Testlet Type B*	Testlet Type C*
Stage 1	64 (8)	65 (6)	66 (7)
Stage 2—Low	69 (20)	70 (17)	66 (9)
Stage 2—Medium	76 (10)	70 (6)	71 (6)
Stage 2—High	72 (12)	74 (12)	75 (7)

	Testlet Type A**	Testlet Type B**	Testlet Type C**	Testlet Type D**	Testlet Type E**
Stage 1	71 (6)	66 (6)	66 (8)	61 (4)	66 (7)
Stage 2—Low	72 (16)	71 (18)	75 (15)	64 (9)	66 (9)
Stage 2—Medium	72 (8)	73 (6)	72 (7)	75 (7)	71 (6)
Stage 2—High	74 (14)	74 (13)	77 (10)	77 (15)	75 (7)

Note. The first number in each cell represents the number of testlets with acceptable overlap. The numbers in parentheses represent the highest item exposure rate for a group of testlets.

With regard to pool efficiency, Table 4 provides summary statistics of the item parameters for the 58 items in the pool that were not used in any testlets. Comparing this table to Table 2 does not reveal any particular statistical characteristics of these unused items. Table 5 summarizes the distribution across item subtypes for the unused items. (Note that there are nine item subtypes for the logical reasoning item type, and they are

numbered from 01 through 09 for this table.) There were 28 items of subtype 03 and 15 items of subtype 08 that were not used in any testlets. The number of unused items for the other subtypes was insignificant. In reviewing the item pool along with the content constraints for the various testlets, the larger number of unused items from the 03 and 08 item subtypes seems to simply reflect a greater surplus of these items in the pool.

TABLE 4
Descriptive statistics of the items not included in any testlets

Variable	*N*	Mean	Standard Deviation	Minimum	Maximum
a	58	0.553	0.168	0.324	1.007
b	58	−0.525	1.008	−2.924	1.921
c	58	0.160	0.096	0.029	0.451

TABLE 5
Distribution of unused items by item subtype code

Item Subtype Code	Number of Unused Items
01	4
02	2
03	28
04	1
05	1
06	0
07	7
08	15
09	0

In order to further evaluate the characteristics of the unused items, the IRT item parameters and word count were studied for each. Nearly every item was determined to have a very low *b*-parameter value (–2.0 or less), a lower *a*-parameter value (0.4 or less), a high *c*-parameter value (0.4 or greater), or a very high word count (80 or greater). These observations seem to indicate that the characteristics of these items are such that they are not very useful for testlet assembly. A certain number of such items are expected in any item pool, and the number observed here is not unreasonably high.

Results of Content-Balanced Two-Stage Test Simulation

The two-stage testlet design was simulated using the two testlet type content-balancing scheme. Figures 4 and 5 present the *RMSE* and *bias* analyses for this simulation, plotted along with the *RMSE* and *bias* for the two paper-and-pencil tests. The *RMSE* analysis indicates that the content-balanced two-stage test is more precise than either of the paper-and-pencil tests in the middle of the ability scale. For ability values above 1.5 and below –2.0, the precision of the 51-item paper-and-pencil test begins to exceed that of the content-balanced two-stage test, and for ability values below –2.5 and above 2.5, both paper-and-pencil tests are more precise than the content-balanced two-stage test. With regard to the *bias* statistic, the content-balanced two-stage testlet performs similarly to the 25-item paper-and-pencil test for the entire range of the ability scale.

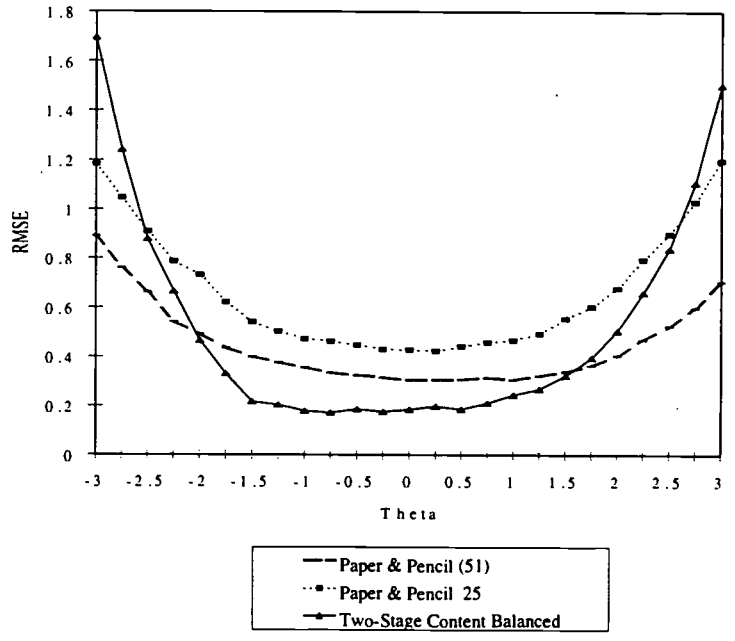

FIGURE 4. *Comparing test designs: RMSE*

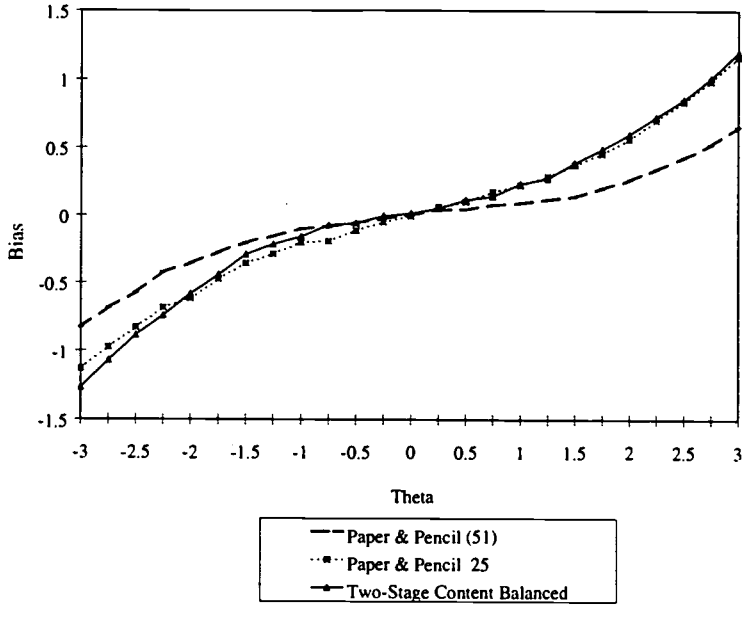

FIGURE 5. *Comparison of test designs: Bias*

Discussion and Future Directions

This study evaluated three different content-balancing schemes for applying content constraints to a testlet-based two-stage testing design. Evaluation of these content-balancing schemes occurred on two levels. First, the assembly of the testlets was evaluated in terms of statistical criteria and in terms of overlap and item exposure. Next, a simulation of a two-stage testing design was carried out for the simplest content-balancing scheme, and the precision of the final ability estimate was evaluated.

The results indicated that a great deal more work is needed on the assembly of the testlets. The target testlet information functions were not met to an acceptable degree for many of the testlets. Modifications to better meet this criterion will include adjusting the upper bound on the target information function. Since it is the lower bound that we are most concerned with, the upper bound will be made more liberal by increasing the distance from the lower bound to the upper bound and by deriving an upper bound that represents a uniform increase over the lower bound across the entire ability scale. The lower bound will also be treated as a true minimum that must be met by all testlets. In this way, the desired properties of the testlet information functions should be met more closely. In addition, a target score distribution will be applied to the assembly of the testlets. The algorithm being applied is already set up to handle a target score distribution, and this additional statistical specification should lead to more uniformity in the testlet assembly.

More strict constraints on item overlap and item exposure will also be applied in future testlet assembly. Testlets will not be permitted to overlap by more than two items within a content type. Restrictions on overlap between content types within a content-balancing scheme will also be defined and imposed. With regard to item exposure, more controls will be implemented to assure that the most desirable items are not over-exposed. Item exposure was highest for the Stage 2 low and high difficulty testlets for all content types. If exposure rates cannot be imposed in these regions for the current item pools, perhaps an effort needs to be made to produce more items in these ranges.

Recall that the content constraints were met exactly in the assembly of all testlets. This is true in spite of the fact that the content constraints defined for this study were very strict and inflexible. In reviewing the results observed here, LSAT specialists have indicated that more flexible content constraints can be developed that will still result in an acceptable balance over the total two-stage test. Relaxing the content constraints may allow for more efficient use of the item pool and more adherence to the statistical constraints.

Even though the statistical constraints established for the testlet assembly were not met exactly, the two-stage testlet design with the two testlet-type content-balancing scheme performed quite well in the middle of the ability scale. Less precision was observed at the extremes of the ability scale. These results are very encouraging. The target information curves for the Stage 2 low, medium, and high difficulty testlets were centered at θ values of -1.0, 0.0, and 1.0. In the region of the ability scale covered by these testlets, the content-balanced two-stage design performed quite well. Less precision was observed at the extremes of the ability scale. Since many law schools are making admissions decisions among highly ranked candidates, we are concerned with improving the precision at the upper end of the ability scale. Schnipke and Reese (1999) reported on a multi-stage design that included a third stage with a very high and a very low difficulty level. Perhaps a content-balanced version of this design would result in greater precision at the extremes of the ability scale. In addition, the modifications to the testlet assembly described above will also result in greater precision.

This study focused on only one of the item types currently included in the LSAT. An extension of this research is currently underway to incorporate the analytical reasoning and reading comprehension item types. Since these item types are set-bound, some additional issues will have to be considered in both the testlet assembly and the test administration.

References

Armstrong, R., Jones, D., & Kunce, C. (1996). *IRT test development applications using network flow methods.* Working paper, Rutgers Center for Operational Research, New Brunswick, NJ.

Fisher, M. L. (1981). The Lagrangian relaxation method for solving integer programming problems. *Management Science, 27* (1), 1-18.

Foong, Y., Lam, T. (1991, April). *Development and evaluation of hierarchical testlets in two-stage tests using integer linear programming.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Thousand Oaks, 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.

Nemhauser, G., & Wolsey, L. (1988). *Integer and combinatorial optimization*. New York: John Wiley & Sons.

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

Reese, L. M., & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized adaptive testing* (Computerized Testing Report 96-04). Newtown, PA: Law School Admission Council.

Schnipke, D. L., & Reese, L. M. (1997, March). *A comparison of testlet-based designs for computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Stocking, M. L. (1996). *Revising answers to items in computerized adaptive tests: A comparison of three models* (Research Report RR-96-12). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Swanson, L. (1992). *A method for severely constrained item selection in adaptive testing* (Research Report RR-92-31). Princeton, NJ: Educational Testing Service.

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

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").