LSAC RESEARCH REPORT SERIES

Computerized Adaptive Testing with Multiple Form Structures

Ronald D. Armstrong
Faculty of Management
Rutgers University

Douglas H. Jones
Faculty of Management
Rutgers University

Nicole B. Koppel
School of Business
Montclair State University

Peter J. Pashley
Testing and Research
Law School Admission Council

Law School Admission Council
Computerized Testing Report 99-14
May 2006

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

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

Table of Contents

Executive Summary

A multiple form structure (MFS) is an ordered collection or network of testlets (i.e., sets of items). An examinee's progression through the network of testlets is dictated by the correctness of the examinee's answers, thereby adapting the test to the examinee's trait level. The collection of paths through the network yields the set of all possible test forms, allowing test specialists the opportunity to review them before they are administered. Also, limiting the exposure of an individual MFS to a specific period of time can enhance test security. This paper provides an overview of methods that have been developed to generate parallel MFSs. The approach is applied to the assembly of an experimental computerized Law School Admission Test (LSAT).

Introduction

The last decade has seen paper-and-pencil (P&P) tests being replaced by computerized adaptive tests (CATs) within many large-scale standardized testing programs, such as the Graduate Record Examination (GRE). A CAT may yield several advantages relative to a conventional P&P test. A CAT can determine the items to administer in real time, allowing each test form to be tailored to an examinee's trait level (ability). That is, an examinee's responses to items can be assessed sequentially during the test and a regularly updated estimate of the examinee's ability can be maintained. Subsequent items can be chosen that more closely match the true trait level of the examinee. By adapting to an examinee's trait level, a CAT can acquire more information about an examinee while administering fewer items. Other potential advantages of CAT are immediate scoring, more frequent and flexible administrations, the collection and use of item response latencies, and the opportunity to use innovative item formats and types. A general review of CATs can be found in Weiss (1985) and Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, and Thissen (1990).

Although each examinee is usually administered fewer items on a CAT than on an equivalently reliable P&P test, the concerns relating to item exposure in CAT are more severe than with P&P assessments. A typical P&P item can be administered to many examinees at one administration, then disclosed and not used again. CAT administrations are typically spread over longer periods and items are continuously exposed, though not necessarily disclosed. Therefore, high-stakes CAT administrations may be more vulnerable to compromise as compared to their P&P counterparts.

A conventional CAT form is highly personalized because items are chosen from an item pool in real time during the test. Since a CAT form given to an individual examinee might never be administered again, test scores from different examinees are more difficult to equate than P&P scores. Multiple P&P forms for a standardized test are constructed to be parallel, and adjustments for small differences between forms can be made after each administration because the number of examinees responding to each form is large. In addition, the number-correct in a conventional CAT has little meaning and ability estimates can be difficult to explain to examinees. In part to address these latter concerns, Stocking (1996) presented a method for scoring a CAT with an equated number-correct method where a separate equating table was required for each CAT.

Before each P&P test administration, test specialists have the opportunity to review the P&P test form. This would be infeasible with a conventional CAT because the number of possible assessments increases exponentially with the number of items in the pool. For example, even a 10-item test from a 20-item bank would give rise to almost 200,000 possible forms without considering unique item orderings. In a CAT environment, a testing program must rely on a computer item selection algorithm to construct forms. While these algorithms can be very sophisticated, ensuring that large numbers of forms constraint are satisfied, they can still produce forms that would be rejected by an experienced test specialist. A multiple form structure (MFS) provides a means to implement a CAT that alleviates the difficulties described above. An MFS is an ordered collection of testlets (sets of items bundled together; see, e.g., Wainer and Kiely, 1987) that allows for adaptation based on an examinee's ability while exposing a predefined number of items and providing a reasonable number of possible forms. This test structure is a hybrid or compromise between the conventional P&P and CAT formats.

Related Approaches

An MFS can be viewed as a computerized extension of the early attempts at P&P adaptive tests. In Lord's 1971 paper on two-stage testing, a routing test of average difficulty was administered to gauge the examinee's ability. The second stage consisted of three measurement tests developed for low, medium, and high abilities. The number of correct responses on the routing test determined the appropriate test at the second stage. More precise measurement was achieved with the two-stage test than a linear test of the same length. The flexi-level test described in Lord's 1980 book (pp. 115–127) is closely related to the MFS approach. The flexi-level test possesses many of the properties of the MFS. Items are adapted to the

examinees ability, form review is possible, and number-right score can quantify the performance. Although the flexi-level testing described did not use testlets, it could have.

As a precursor to the MFS design, Reese and Schnipke (1996) introduced a computerized multistage testlet design. Testlets at levels within stages were assembled to approximate average difficulty levels for different ability groups. Routing was accomplished with a number-right score. The testlet to administer at each stage/level was chosen from a large set of possible testlets based on exposure control considerations; thus, forms were not easily available for review before they were administered.

A more recent methodology, computer-adaptive sequential testing, or CAST (Luecht & Nungester, 1998), is very similar in theory to the MFS approach, even though it was independently developed. In fact, both CAST and MFS methodologies have employed some of the same terminology to describe their design features. The fundamental advantages and implementation issues for CAST are the same as those for MFS; however, the approaches used to tackle the implementation issues are different. For two reasons the authors of this paper have decided to continue to favor the MFS terminology and methodology that has appeared previously, rather than adopt a CAST framework. First, the CAST approach was primarily developed in the context of a long mastery test, specifically the United States Medical Licensing Examination. The MFS approach described in this paper was developed with a shorter admission test in mind, specifically the Law School Admission Test (LSAT), which requires a broader measurement range. Second, the testlets (or modules) used in the CAST approach were usually assumed to be quite large (between 60 and 90 items in one example given by Luecht and Nungester) compared to the five to eight item testlets assumed in the MFS approach. As such, the CAST approach is closer in nature to multistage testing in which each stage provides fairly reliable measurement results. In comparison, the MFS approach assumes less measurement precision from the much smaller testlets it normally employs, and instead focuses more on the characteristics of the multiple forms that exist within a structure.

The remainder of this paper addresses the various aspects of the MFS approach to CAT. The next section discusses the attributes of a multiple form structure design (MFSD). The potential advantages of an MFS are given after the design considerations are presented. The various considerations that go into creating the components of a complete design are discussed in the next sections. The methods used to assemble MFSs are overviewed and simulation results are then presented.

Multiple Form Structure Design

An MFSD differs from an MFS in that an MFSD is a framework and each MFS is an occurrence (instance) of this framework. The MFSD provides the template for each MFS. Many MFSs can be assembled based on one particular MFSD. All MFSDs are defined by the following attributes:

a) *Bins*. Every MFSD consists of multiple bins. When an MFS is constructed, every bin contains a single testlet. The MFSD defines bins for different ability levels.
b) *Targets*. Every bin in the MFSD has a Target Information Function and a Target Characteristic Curve. Testlets assigned to a bin match the associated target curves within an acceptable deviation limit. These targets can be based on item response theory, or IRT (see, e.g., Hambleton, Swaminathan, & Rogers, 1991; or Lord, 1980).
c) *Number of items per bin*. Every testlet assigned to a bin contains the number of items within the range specified by the MFSD.
d) *Stages*. Every bin in the MFSD is assigned to a stage. A stage ends when the examinee leaves a bin in the stage.
e) *Routing rules*. Decision rules must be employed at the end of every stage to determine the next bin, if any, where the examinee is routed. The rules must be concise and, in some manner, depend on the examinee's ability.
f) *Constraints*. Additional constraints may be included in an MFSD, such as subject content, answer key count, word count, topic coverage, and diversity usage.

Example of an MFSD

MFSDs differ in the number of bins, number of items per bin, the number of stages, the target curves, the routing rules, and the constraints. All MFSs corresponding to a particular MFSD are similar in all of these attributes. Table 1 depicts an example of an MFSD that is being considered as the template for a computerized version of the LSAT. The layout depicts an MFSD having six stages with a total of 14 bins. The bins at a given stage are arranged in levels corresponding to item difficulty/examinee ability classifications or strata. In this example, every bin will have a testlet with six items assigned. Every possible path through the MFS constitutes a test form. The path that a particular examinee may traverse through an MFS contains exactly one testlet from each stage.

TABLE 1
An MFSD with six stages and 14 bins is outlined. Each bin is targeted for a particular population percentile range as indicated by the bracketed values next to each bin number

Stage:	(i)	(ii)	(iii)	(iv)	(v)	(vi)
						14:[87,100]
				7:[75,100]	10:[75,100]	
			4:[50,100]			13:[50,87]
	1:[0,100]	2:[0,100]		6:[25,75]	9:[25,75]	
			3:[0,50]			12:[13,50]
				5:[0,25]	8:[0,25]	
						11:[0,13]

Stages (i) and (ii) each have a single bin which contains testlets targeted for 100% (indicated by "0,100" in Table 1) of the test-taking population. Stage (iii) has two bins. The testlet in Bin 3 targets examinees in the bottom 50 percent (0-50) of this population, and the testlet in Bin 4 is targeted for examinees in the top 50 percent (50-100). Stages (iv) and (v) have three bins designed for testlets that target people at progressively increasing trait levels as indicated in the table. The final stage, Stage (vi), contains four bins, numbers 11, 12, 13, and 14. The testlets targeted for Bin 11 are designed for examinees in the lowest 13 percent (0-13), Bin 12 for the next 37 percent (13-50), Bin 13 for the next 37 percent (50-87), and Bin 14 for the top 13 percent (87-100) of the test-taking population.

All examinees begin at Bin 1 in Stage (i) and proceed to Bin 2 in Stage (ii). That is, the examinee is routed from Bin 1 to Bin 2 regardless of the results in Bin 1. A measure of the examinee's trait level is estimated from the responses to the items in the first two stages. Based on this evaluation, the examinee would advance to one of the two bins in the third stage. A possible method for routing out of Stage (ii) is the following: Proceed to Bin 3 if the total number of correct responses to the items in Bins 1 and 2 is less than 8, and proceed to Bin 4 otherwise. Alternatively, an IRT estimate of ability, rather than number of correct responses, could be used for routing. Whenever a routing decision is to be made, an ability estimate is obtained by considering responses to all previous items.

The number of paths through an MFS can be constrained through routing rules. For example, in the specific MFS described above, an examinee's movement could be restricted to go up or down at most one level. Regardless of the responses in Stage (iii), the option of an examinee advancing from Bin 3 to Bin 7 could be eliminated. Similarly, an examinee administered a testlet from Bin 4 may never be able to move to Bin 5. Further, no adaptation may be allowed from Stage (iv) to Stage (v); that is, those examinees completing Bin 5 (Bin 6, Bin 7) would be routed directly to Bin 8 (Bin 9, Bin 10). Restricting movement between stages in this way limits the number of possible forms that would need to be reviewed by a test specialist, but may decrease the measurement efficiency of the MFS.

The MFSD outlined in Table 1 is representative of a practical design being considered for a possible computerized LSAT. Currently, designs with four to six stages and with five to eight items per testlet are being evaluated. Evidence suggests that a maximum of four levels is desirable at the last stage and three levels may be adequate. All designs considered to date have forced routing from Bin 1 to Bin 2, and two bins at Stage (iii). The number of bins never decreases when progressing through stages. The most complex candidate designs have 12 paths. The number of bins is taken to provide a workable MFS. The maximum number of bins considered arises from a six-stage MFSD with 1-1-2-3-4-4 bins, respectively, at the six stages, yielding 15 bins. The simplest MFSD has four stages with 1-1-2-3 bins. To this point, two types of percentile cutoffs have been considered. The first has evenly spaced percentile cutoffs; that is, the 1-1-2-3-4-4 design would have percentiles [0-50] and [50-100] at Stage (iii), [0-33], [33-67], and [67-100] at Stage (iv), and [0-25], [25-50], [50-75], and [75-100] at Stages (v) and (vi). The second approach is like the one depicted in Table 1. This latter percentile assignment aims for more measurement precision in the tails by assigning the outer percentiles half the weight of the inner percentiles; thus, the percentiles of Stage (v) and (vi) would be [0-13], [13-50], [50-87], and [87-100]. An operational MFSD will have a trade-off between multiple adaptations for scoring accuracy, and simplicity for effective pool usage and review.

Potential Advantages

The multiple-stage format of an MFS is conducive to standardized tests that must adhere to various operational constraints, such as reading passage topics, item conflicts, word count, and answer key distribution. Within each stage, common attributes can be established among the testlets. For instance, all testlets in a specified stage may be of the same subject matter but differ only in difficulty level. Requiring testlets at each stage of an MFS to have specified characteristics ensures that every examinee fulfills all the requirements of the test. While other CAT implementations enforce stated operational constraints, the MFS approach allows for form review. The review may not be as detailed as a P&P review, since multiple forms

exist, but it does provide an opportunity for test specialists to identify subtle violations.

An examinee's path through the network of testlets adapts to the examinee's trait level. The collection of paths through the network yields the set of all possible test forms. In the above example with the mentioned forced routings, there are eight (1-2-4-7-10-14, 1-2-4-7-10-13, 1-2-4-6-9-13, 1-2-4-6-9-12, 1-2-3-6-9-13, 1-2-3-6-9-12, 1-2-3-5-8-12, and 1-2-3-5-8-11) possible paths (where each number represents a bin in the MFSD). The forms contain common items and each stage contains testlets with similar characteristics; for example, the testlets for Stage (iii) in the example might all be based on the same passage. The MFS design and assembly should be constructed to facilitate reviews by test specialists. Test security may be enhanced because a single MFS can be delivered in a testing session. The 14 testlets from the above example may be exposed to several hundred people over a few hours, although any given examinee would only be exposed to exactly six testlets. A different parallel MFS (based on the same MFSD) could be given at the next session. Through careful use of the item pool and limitation of the number of testing sessions, concerns about item exposure are substantially reduced. The most practical way to administer MFSs is not clear at this time. Ideally, the administration would emulate the P&P administrations where a large number of examinees are examined simultaneously. However, similar to the conventional CAT, an MFS is taken on a computer and so the large-scale administration of an MFS to many examinees simultaneously may not be practical with facilities currently available. Testlets are a natural way to administer a set of items related to a single passage, where a passage (or scenario) is given, followed by a set of related items. Packaging items in testlets allows examinees the opportunity to observe all the items before answering any of them and to review all of their answers, changing them if need be, before moving to the next testlet. This is true with any testlet, but the benefit is more pronounced if all items refer to a single passage.

While IRT can be used to equate CATs, large-scale operational testing programs have found that in practice the underlying IRT assumptions, such as item invariance, were often sufficiently violated to cause scale drift over time (see, e.g., Bock, Muraki, & Pfeiffenberger, 1988). Intact forms, where item positions are held constant, can be more robustly equated than standard CATs. Testing programs typically apply item position rules while assembling operational forms (i.e., rules that restrict the movement of items position-wise from pretest to operational forms) to guard against pretest item calibrations from being adversely affected by context effects, differential speededness, fatigue effects, etc. An MFS attempts to control these factors by maintaining item positions from the pretest to the operational step, thus ensuring more stable IRT calibrations and equatings. Parallel MFSs would be created for each administration, and every path in an MFS could be equated to its associated path in another parallel MFS. Using the number correct along a path to produce a scaled score is a reasonable alternative to an IRT ability estimate commonly used with CAT.

Application of Study

The Law School Admission Council (LSAC) currently administers the Law School Admission Test (LSAT) in a P&P format. It is given at four main administrations a year (June, September or October, December, and February) at designated centers throughout the world (though most centers are located in the United States and Canada). In 2002–2003, over 140,000 LSATs were administered. Most law schools in the United States and Canada use LSAT results as part of their admission process. The LSAT is composed of five sections, one of which is used for pretesting and preequating future items and form sections. The four scored sections are composed of two in logical reasoning (LR), one in analytical reasoning (AR), and one in reading comprehension (RC). More information on the LSAT can be found at *www.LSAC.org*.

In December 2000, LSAC finished an initial project to computerize the assembly of the LSAT P&P test, redesign the database and computer displays, and study possible implementations of computerized testing. LSAC is carrying out research to determine which MFSD might be most appropriate for a computerized LSAT. Similar to the P&P test administration, the proposed computerized LSAT includes the same general sections of LR, AR, and RC. Each of these sections can be represented by a specific MFSD. Each section has unique characteristics and constraints that must be considered while the MFSs are constructed. Both the RC and the AR sections have set based items; that is, a passage is provided and items related to the passage follow. The passage and its items become a testlet. The LR testlets primarily contain discrete items with a one-to-one relationship between a passage and an item. However, there are a few LR passages with two associated items and both items appear in a testlet when the passage is present. All testlets are composed of between five and eight items each.

Each section of the P&P tests has constraints on cognitive content coverage, answer key distribution, and total word count. A computerized LSAT might eventually replace word count constraints with timing constraints obtained from statistics gathered during the pretesting and preequating. The AR and RC sections have passage topic restrictions. These constraints are similar to those of most standardized tests.

Targets and Routing

Unlike most CAT implementations, MFSs utilize target testlet information functions (TTIFs) and target testlet characteristic curves (TTCCs). Target curves are common in P&P tests because they facilitate the construction of parallel forms. As with P&P targets, the MFS targets were chosen to produce reliable tests while utilizing the item pool effectively. However, in the case of an MFS the targets for each bin and the routing rules are interrelated. That is, routing rules must depend on the targets and vice versa.

A good target for a bin is based on the subpopulation visiting the bin. Reasonable MFSs can be assembled from targets created with a trial-and-error approach. However, the MFSDs of this study utilize a formal method to determine targets reflecting the general population and available item pool. The distribution of examinee's IRT ability over recent years was estimated. It was found to be approximately normal, with a mean close to zero and a standard deviation close to one. Given this assumed population, an algorithm determining targets and routing rules was developed. A detailed description of the method is given in Armstrong and Roussos (2001) and Armstrong (2002). An overview is given here.

The target creation process required an item pool that accurately reflected the characteristics of *future* item pools to be used by the testing program. This study utilized an existing pool. Next, an MFSD was created without targets and routing rules but with all the other components listed above. The administration of a test based on the MFSD was simulated using an "omniscient" testing implementation. Omniscient testing is a modification of the shadow CAT proposed by van der Linden and Reese (1998) where an active test, satisfying all constraints, is maintained and items to deliver are chosen from this active test. Once an item is delivered, it is fixed on the active test. The abilities of the examinees are drawn from a normal distribution. The omniscient testing simulation knows the true ability of each examinee before the administration and uses the true ability to choose items to place on the active test and administer. Simulation results from the omniscient testing were used to develop targets for the MFSD. The motivation was to produce the best average testlet information functions (TIFs) and testlet characteristic curves (TCCs) that the pool would support subject to exposure control and other constraints.

The assembly problem can be modeled as a mixed integer programming (MIP) problem to minimize an objective function subject to linear constraints. All test assembly constraints were satisfied by the active test. The constraints on the optimization problem were those usually associated with automated test assembly and mentioned previously. The constraints are not stated explicitly here as they vary and are not the focus of this paper. A formal statement of typical test assembly constraints can be found in Armstrong, Jones, and Kunce (1998); Boekkooi-Timminga (1990); Theunissen (1985); and van der Linden (1998). The objective function provided a trade-off between exposure control and information. The cost coefficient for the i^{th} item is given by

$$\alpha_1 r_i + \alpha_2 \tilde{u}_i - \alpha_3 I_i(\theta) \tag{1}$$

where r_i is a uniform random number between 0 and 1, $\tilde{u}_i$ is the empirical exposure rate of the i^{th} item derived from all previous assemblies, and θ is the true ability of the examinee. The function $I_i(\theta)$ denotes the information derived from the IRT model for the i^{th} item. The parameters $\alpha_1, \alpha_2, \alpha_3$ were assigned positive values determining the weight for each term. The cost coefficients of all decision variables not associated with the item assignments were zero.

The omniscient test simulation was administered as it would be administered for an MFS. Any stipulated sequencing of the administration was enforced. For example, the MFS may start with items covering specific topics. The CAT administered a testlet (stage) at a time. If more than one testlet can be administered in a stage, the testlet to administer was chosen randomly with equal probability as the other eligible testlets. No adaptation was necessary since the examinee's ability was known. In fact, for purposes of developing targets, there was no reason to simulate the administration of the items. This was done for summary information used in later comparisons. The administration of the omniscient test was simulated to a large number of examinees (4,000, for example). The first half of the simulated examinees was used to stabilize the exposure rates of items, but the results were not used to determine the targets. The information function and characteristic curve values for each testlet administered to every examinee in the second half were saved in the ability range from –3.0 to +3.0 in steps of 0.3.

The probability of an examinee with a known true ability, denoted by θ, visiting any bin was approximated by sequentially considering the bins one at a time. All examinees visited the first bin with probability one. The average of all the TIFs and TCCs administered in Stage (i) created the initial targets for this stage. When considering bins after Stage (i), the TTCCs of all previous bins and the routing rules provided an estimate of the probability of an examinee with a given trait level visiting a bin. These probabilities were used to yield a weighted average of the TIFs and TCCs administered during the current stage. This became the TTIFs and TTCCs for the current bin.

The routing rules for the MFSD were determined concurrently with the targets. The forced routings mentioned previously determined the routing without further labor. Consider the case where forced routing was not present. The routing out of a bin at stage s was dependent on the path taken to arrive at the bin. All bins of the MFSD were intended to attract a specific percentile group. First, the expected fraction of the entire population visiting the bin by traversing a given path was obtained. This was accomplished by numerical integration over the population. The quadrature points used in the integration provided θ values. Conditioning on a θ value, the probability of visiting a specific bin was estimated using the established TTCC as the difficulty of testlets assigned to a bin. This probability was multiplied times its associated integration weight. Summing the result over the quadrature points provided an estimate of the fraction of the entire population visiting a bin by traversing a given path. Next, each percentile group corresponding to the bins at stage $s + 1$ was considered. Using numerical integration (similar to before), the fraction of the examinees visiting the specified bin at stage s by traversing the path and whose ability came from each percentile group was estimated. If any fraction was small (less than .05), the fraction was allocated proportionally to the other percentile group fractions and the small fraction was set to zero. Examinees were routed out of the bin based on these fractions and the population's cumulative score distribution conditioned on the path. For example, suppose that half the examinees arrived at a bin via the path came from one percentile group and half came from another. Examinees scoring in the lower half of the score distribution were routed to the lower percentile group, and those scoring in the top half of the score distribution were routed to the higher percentile group. All the necessary probabilities were estimated from the TTCCs of the current bin and the bins of prior stages.

Once an MFS was assembled, the routing rules for the MFS were computed using the IRT parameters for the items found in the testlets. These rules are more precise than those computed with the TTCCs. However, since the testlets are assembled to match the TTCCs and the TTIFs, the routing rules for the MFS are similar to the routing rules for the corresponding MFSD.

MFS Assembly

The assembly of a single MFS is a large-scale MIP problem. This section describes a two-phase process of developing parallel MFSs once the MFSD has been specified. In the first phase, a testlet pool was created from the item pool. In the second phase, parallel structures were created from the testlet pool. Special purpose solution methods were utilized in this study, but it is possible to perform the assembly with a commercial MIP package. The MFS assembly method did not assure the optimal solution to the stated mathematical programming problems. The true objective when assembling MFSs was the same as the usual objective when assembling parallel P&P tests. The goal was to assemble as many "acceptable" tests as possible from the item pool. MFSs were assembled in sequence, and randomization and exposure control were employed to increase the number of parallel MFSs to be derived from the pool.

Phase I

From an item pool with estimated IRT parameters, a pool of testlets was created. Each testlet in the pool can potentially be assigned to a bin in the MFS. An individual TIF and TCC matched the TTCC and TTIF, associated with a particular bin, to a specified degree of accuracy. Also, the testlets had to satisfy other practical constraints as well, such as total word count or answer key count. Testlets were categorized according to the specifications that were related to a bin, or possibly bins, in an MFS. Items can appear in multiple testlets within the testlet pool, but not within multiple testlets on a single path through an assembled MFS.

The TTCCs and TTIFs were determined across a range of ability. The mathematical model proposed here took values for the TIF and TCC evaluated at discrete levels, θ_k, $k = 1, ..., K$. The TIF and TCC were smoothed for a testlet with only five or six items; and, with $K > 15$, the deviations between the targeted and observed curves at the discrete points were an excellent indicator of the overall deviation. The deviation of a TCC and TIF from the targets was determined by summing the weighted absolute deviations at the K discrete ability values; that is, a weighted norm was used. The weights depended on the population group for which the testlet was aimed. For example, a testlet designed for a high-ability group would have small or even zero weights assigned to low θ values.

The deviation limits for matching the testlets to the targets could be loose because the MFS assembly (Phase II) had the deviation of path target curves as a criterion. Positive and negative deviations among testlets combine to provide an acceptable deviation over the path. A passage with 12 items yields 924 possible 6-item testlets, and many of these may be acceptable for use in an MFS. Certain single item passages may be eligible for inclusion in even more testlets. There was no reason to have this many testlets from one passage in the testlet pool and the algorithm did not include all of them.

The testlet pool was generated before assembling the MFSs for computational reasons. The complete testlet pool was stored temporarily and only those testlets used in MFSs were kept in the database. Every testlet in the testlet pool was classified according to the target and constraint set used to construct the testlet. For example, a testlet may be classified as targeting a low-ability level, using a passage with a scientific topic, and satisfying specified cognitive skill constraints.

The overall test constraints were divided into sets of testlet requirements. These requirements provided constraints that were enforced during testlet construction. A constraint set was then associated with each stage in an MFS. Once testlets were assigned to bins of an MFS, the overall test constraints were satisfied on every path. These constraints were developed by test specialists to achieve a valid test. For example, a six-stage MFSD had test constraints represented by six testlet constraint sets, where each set was assigned once on a path. Multiple testlets were constructed for every constraint set and target combination. Every bin at a randomly chosen stage was designated to receive a testlet satisfying a specific constraint set and fit the target curves for the bin. Multiple testlet constraint sets with different path assignment stipulations were necessary to fully utilize the item pool with set-based items. The following discussion assumes a single constraint specification for each MFSD.

There were multiple objectives when creating the testlet pool. First, uniform usage of items in the pool was desirable. That is, ideally, every item appeared in approximately the same number of testlets. This objective promoted uniform item exposure and optimal usage of the item pool. Second, a degree of randomness in the allocation of items to testlets was desirable, so that no recognizable patterns exist in item usage. Of course, the various constraints that were imposed on the construction of a testlet precluded meeting exactly the usage and randomness goals. Third, a testlet came moderately close to the TTIFs and TTCCs to which it was associated.

Consider the assembly of a new testlet to be added to an existing testlet pool where some of the testlets may have been used in existing MFSs. A testlet is classified according to the bin target and constraint set used in the assembly. The symbolic representation of the testlet pool is denoted by L. Let x_i , $i = 1, 2, ..., m$ denote zero-one decision variables for testlet item selection. The testlet assembly process had $x_i = 1$ if the i^{th} item was present on the testlet and $x_i = 0$ when it was not present. A cost was assigned to each item. This cost was the value of a function $G_i(r_i , \overline{u}_i , L_i)$:

$$G_i(r_i , \overline{u}_i , L_i) = \beta_1 r_i + \beta_2 \overline{u}_i + \beta_3 L_i \tag{2}$$

where r_i was a random number, $\overline{u}_i$ was the fraction of existing MFSs that contain item i, and L_i represented the fraction of testlets in pool L that contain item i. The parameters $\beta_1 , \beta_2 , \beta_3$ were assigned positive values determining the weight for each term.

The objective function included a measure of the weighted deviation of the TIF and TCC from a bin's TTIF and TTCC (denoted by "deviations"). The mathematical programming problem was the following:

$$\text{Minimize} \sum_{i \in D} G_i x_i + \sum_{k=1}^{K} |(deviations)| , \qquad \text{subject to } x \in X \tag{3}$$

where X was the set of values for x = (x_1, x_2, ..., x_m) that satisfied the constraints. The weights for the deviation terms were normalized to sum to one for each target. The weights were chosen to approximate the proportion of examinees with the θ of the target point visiting the bin. The TCC and TIF have different metrics, and this resulted in the TIF deviation having a greater impact than the TCC deviation.

In the LSAT case, the AR and RC sections had between 6 and 12 potential items associated with each passage. Since all items in a testlet must come from a single passage, complete enumeration of all candidate testlets was possible. The items from some passages generated several hundred acceptable testlets. The number of testlets to add to the pool relating to a single passage was reduced, and the testlets were spread over various constraint sets and targets.

Lagrangian relaxation was used to match the TTIFs and TTCCs, and to obtain approximate solutions to the MIP problem when the testlet had discrete items. The reader is referred to Armstrong, Jones, and Kunce (1998) for details on the algorithmic process. A general review of Lagrangian relaxation can be found in Fisher (1981, 1985) and Nemhauser and Wolsey (1988).

The generation of a new testlet, in the discrete item case, included one special constraint that attempted to force the usage of all items. An item was forced to be present on the testlet by setting the associated decision variable equal to one. This item had a low utilization in the testlet pool and was not found in the already assembled MFSs. A testlet was assembled considering all other eligible items in the pool. Testlets were repeatedly assembled with various constraint sets and target combinations. Thus, approximate

solutions to a large number of MIP problems were attained where each problem forced a different item in a testlet. These problems had many fewer constraints than the associated problem of assembling a complete P&P test form, and the solution process terminated after a couple of seconds.

Once a solution was obtained, the associated testlet was accepted into the pool or rejected. If the testlet sum of weighted absolute deviations from the associated TTIF and TTCC was greater than a prespecified deviation limit, then the testlet was rejected. This deviation limit varied depending on the number of testlets in the pool using a specific passage or item. The prespecified deviation tolerance limits affected the size of the testlet pool. The lower the tolerance, the fewer testlets were included in the pool, and the higher the limit, the more testlets were included. The process chose the tolerance limits large enough to provide an adequate number of testlets in all categories.

Phase II

This phase created parallel MFSs. As mentioned in the Phase I discussion, the testlet pool was organized according to characteristics that will be associated with a bin when a single MFS is created. Each bin must be assigned exactly one testlet out of the set of testlets that are eligible. Multiple MFSs were assembled sequentially and the content constraints associated with a bin changed with each assembly. In this implementation, certain characteristics were randomly associated with stages. The target path information function (TPIF) and target path characteristic curve (TPCC) were the sum of the TTIFs and TTCCs used when assembling the testlets for the bins on the path. Thus, any feasible testlet-to-bin assignment provided a reasonable match to the path targets. However, a Lagrangian relaxation approach was used to balance the positive and negative deviations from the testlet targets along a path. This process made the form associated with a path through an MFS parallel to the form associated with the same path through another MFS.

Let N represent the number of testlets in the pool and $j = 1, 2, ..., N$ be the index ranging over all the testlets. Let V be the number of testlet categories where the constraint set and targets used in the assembly determined the category. Let E_v, $v = 1, 2, ..., V$ be the index set for those testlets in category v. The decision variable $y_j = 1$ if the j^{th} testlet is included on the MFS currently being assembled, and $y_j = 0$ otherwise. Let Q be the pool of MFSs that have already been assembled and may be administered in the future. Costs were assigned, in a similar manner to Phase I, to all eligible testlets. The cost function for the j^{th} testlet was $H_j(r_j, u'_j, Q_j)$ where $r_j$was a random number, u'_j was the testlet usage rate in the current assembly, and Q_j was the usage rate of the testlet in the current pool of MFSs:

$$H_j(r_j, u'_j, Q_j) = \omega_1 r_j + \omega_2 u'_j + \omega_3 Q_j .$$

Most of the constraints on the MFS were automatically satisfied from the constraints imposed when the testlets were formed and assigned a particular testlet category to a designated bin. However, it was not reasonable to account for all constraints by categorizing the testlets, and some constraints were enforced by checks during the MFS assembly. Restrictions related to item reuse along a path and stimulus similarities across stages were examples of such constraints.

The following is a generic statement of the mathematical programming problem for the assembly of an MFS:

$$\sum_{v=1}^{V} \sum_{j \in E_v} H_j(r_j, u'_j, Q_j)\, y_j + |deviations|, \qquad \text{subject to } y \in Y \tag{4}$$

where Y is the set of values for $y = (y_1, y_2, \ldots, y_N)$ that satisfy the constraints.

Numerical Comparison with Paper-and-Pencil and CAT

This section gives the results of simulations comparing the MFS to both a P&P version of the test and a variation of the conventional CAT approach. The MFSD network for each section was based on Table 1. The results reported in this section were consistent with results from other MFSDs studied. An item pool supplied by LSAC was utilized with the number of items and passages given in Table 2. Twenty-two of the LR items were in 11 sets with two items per set, while 1,315 of the LR items were discrete items. If one of the LR set based items was used on a testlet, both of the items were used. With the RC and AR sections, if a passage was used, then exactly 6 items from the passage were used per testlet. Associated with every AR and RC passage there were between 6 and 12 potential items.

TABLE 2
Item pool dimensions for reading comprehension (RC), analytical reasoning (AR), and logical reasoning (LR) sections

	Passages	Items
RC	108	1,021
AR	110	951
LR	1,326	1,337

The MFSDs and MFSs used in the comparison were generated as outlined in this paper. The MFSD network was the one given in Table 1. The population of examinees was drawn from a normal distribution. The omniscient test simulation used to create the targets had $\alpha_1 = 5, \alpha_2 = 30$, and $\alpha_3 = 120$ as the cost coefficients from Equation (1) for all items in both the RC and the AR sections. Because of the increased flexibility in assigning items to testlets with the LR section, the objective function parameters were $\alpha_1 = 5, \alpha_2 = 120$, and $\alpha_3 = 120$. The number of simulated examinees was set at 4,000 for all runs, with the first 2,000 used to establish exposure rates, and were not used for the targets. The second 2,000 examinees provided maximum empirical exposure rates of .209, .237, and .167, and median exposure rates of .013, .019, and .012 for the AR, RC, and LR sections, respectively. The exposure rate for an item was calculated as the number of tests where an item appeared divided by the total number of tests assembled—2,000 in this case. Without the controls for exposure, $\alpha_1 = 0$ and $\alpha_2 = 0$, the maximum exposure rates were over .5 and the median exposure rates were always 0. This means that over half the pool was never used when exposure controls were omitted. The parameters chosen to create the targets provide "reasonable" parameters; that is, they established targets that yielded an acceptable usage of the item pool. A detailed study and a separate report are required to consider the various factors that affect pool usage.

The constraints for MFSDs and the omniscient testing were derived from the constraints for the LSAT P&P sections. There were constraints for content, topic, answer key, and word count, and in the RC section there was an additional requirement for passages with diversity. The information derived from the MFSD design was compared to the current LSAT P&P design. The AR and RC sections of the P&P had four passages per section, and between 22 and 27 items assigned to a section. An LR section contained between 24 and 26 items. The MFSD proposed here had exactly 36 items in each MFS. Since the MFS approach is adaptive, the MFS and two P&P sections were considered to place both testing methods on a comparable level. The two AR (RC, LR) sections chosen for comparison had a total of 46 (52, 52) items. There were no common items in sections. Figures 1, 2, and 3 give an indication of the change in information when the 36-item MFS and two P&P sections are compared. The P&P information across the ability axis was obtained by doubling the target information of the associated LSAT section. The MFSD information was obtained by taking a weighted average of the bin targets. The weight for a bin at ability, θ, was the probability that an examinee with ability θ would visit the each bin.

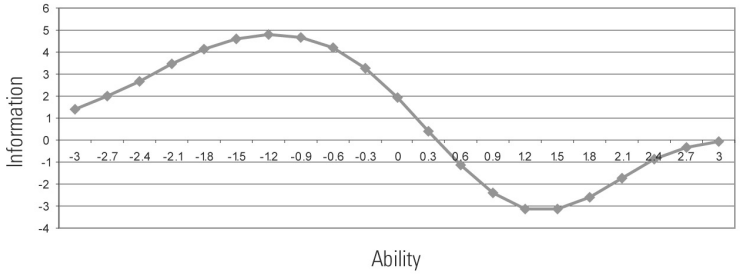

FIGURE 1. *The curve shows the difference between the weighted sum of target information functions for the MFS paths and the target information function of a 46-item Analytical Reasoning P&P Test. The weights used to sum the MFS path targets are the probabilities of traveling the path conditioned on ability*

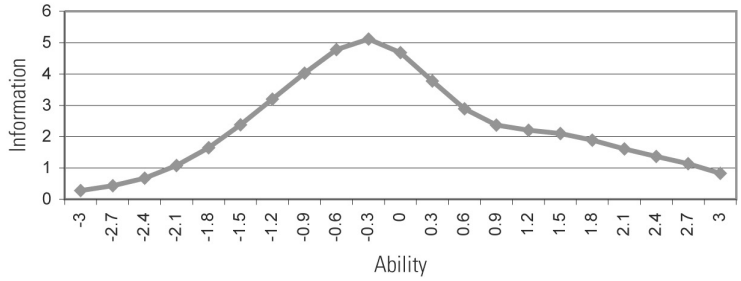

FIGURE 2. *The curve shows the difference between the weighted sum of target information functions for the MFS paths and the target information function of a 52-item Logical Reasoning P&P Test. The weights used to sum the MFS path targets are the probabilities of traveling the path conditioned on ability*

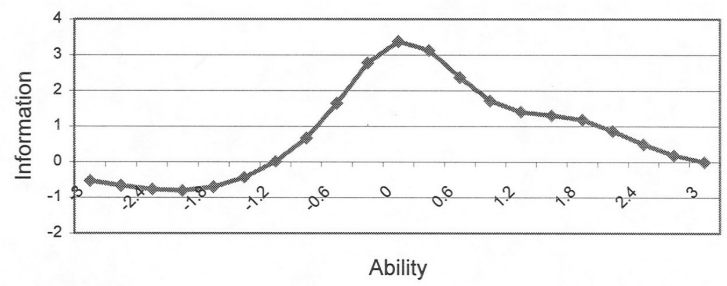

FIGURE 3. *The curve shows the difference between the weighted sum of target information functions for the MFS paths and the target information function of a 52-item Reading Comprehension P&P Test. The weights used to sum the MFS path targets are the probabilities of traveling the path conditioned on ability*

Generally, the MFSD target information for a given ability was more than, or about the same as, the information target for a two-section P&P test. The exception was the high ability group with the AR section. The AR P&P target peaked with a relatively high information target value (greater than 10) for the two combined sections. The MFS AR generated target, based on the item pool and assumed normal distribution of the population, did not achieve information to surpass the two section target at most high θ points. A shift occurred in the AR pool over time to include less difficult items than at the time the P&P AR target was created. This shift was the cause of the exception.

The testlet pool and several MFSs for each section were assembled as outlined in this paper. A single path constraint specification was used for each section. The deviation cutoff limit to make a testlet eligible for the testlet pool was 1.0 in all cases. About 90% of the AR and RC pool passages were included in the testlet pools that had 9,105 and 9,946 testlets, respectively. This was not equivalent to saying that 90% of the passages could be used in MFSs. Additional testlet constraint sets and path requirements achieving the overall desired form constraints should be considered. Usage of the item pool was an important issue. The time for assembling each of the AR and RC testlet pools was about 30 seconds on a 2.0 GHz desktop computer with Windows XP. Every LR item in the pool could have been used in a testlet satisfying at least one of the five constraint sets supplied by the test specialists. This would result in an extremely large testlet pool, and the LR testlet pool was curtailed at 12,880 testlets. This required about 45 minutes of computing time. It appears that additional constraint sets were not necessary for the LR section. Each testlet pool was used to assemble six MFSs. All MFSs provided a fit for the path curves within 15% of the targets across the ability scale. The assembly of the six MFSs took about two minutes. There was no significant difference in assembly time across the sections.

The shadow CAT assembly approach proposed by van der Linden and Reese (1998) was used to compare the MFS approach to a traditional CAT. The shadow CAT assembles a complete "active" test before an item is administered, and the items to administer next are chosen from this active test. Items that had been administered to the current examinee were forced on the active test and could not be chosen for administration again. All test assembly constraints were satisfied by the active test exactly as they were for

the omniscient testing. The only difference between the shadow CAT and omniscient testing simulations came in the way the costs were determined and how items were chosen for administration. The shadow CAT assumed a $N(0, 1)$ prior distribution on ability for all examinees, and this gave an initial θ ability estimate of $\hat{\theta} = 0$. The ability distribution of the examinee was updated before each adaptation based on standard Bayesian rules. The expectation of the updated distribution became the new $\hat{\theta}$ for the examinee. The shadow CAT simulations used expression (1) with $\hat{\theta}$ in place of θ when calculating the cost coefficient. The testlet to administer was chosen with the maximum information evaluated at $\hat{\theta}$. Adaptations in the shadow CAT occurred at the same point as adaptations in the MFSD. As a side observation, the solution times required to solve the problems at each adaptation of the shadow CAT were well below a time limit necessary in an operational setting.

TABLE 3
Summary results of the RMSE and Fidelity for each testing format for Reading Comprehension, Analytical Reasoning, and Logical Reasoning sections

	Analytical Reasoning (AR)		Reading Comprehension (RC)		Logical Reasoning (LR)	
	RMSE	Fidelity	RMSE	Fidelity	RMSE	Fidelity
Paper-and-pencil	45.71	0.902	38.78	0.930	38.32	0.929
Omniscient testing	32.58	0.945	29.33	0.955	25.88	0.964
Shadow CAT	35.69	0.937	33.05	0.946	29.39	0.954
Multiple form structure (MFS)	37.60	0.929	35.20	0.938	31.66	0.948

Table 3 shows the summary results of the simulation of 2,000 examinees drawn from a normal distribution. The results were based on a scaled score ranging from 200 to 800. The scoring method was an extension of the current LSAT scoring utilizing IRT equating (Kolen & Brennan, 1995). The root mean squared error (RMSE) was the square root of the arithmetic mean of the squared difference between the true scores and the scaled scores of the simulated examinees. A fidelity coefficient was the correlation between the true score calculated from the true ability and the scaled score calculated from the number of correct responses of the simulated examinee. Based on these summary statistics, the MFSs out perform the two-section P&P tests, and the Shadow CAT outperforms the MFSs.

To determine the scoring accuracy of the MFS and P&P across the ability range, a simulation was conducted at discrete trait levels. Quadrature points along the θ axis range from –3.0 to +3.0, in increments of 0.3. There were 200 replications at each point. Figures 4, 5, and 6 show the RMSE plot across the range for the AR, RC, and LR sections, respectively. The results are summarized numerically in Table 4. The results were consistent with those already noted for the target information across the ability range.

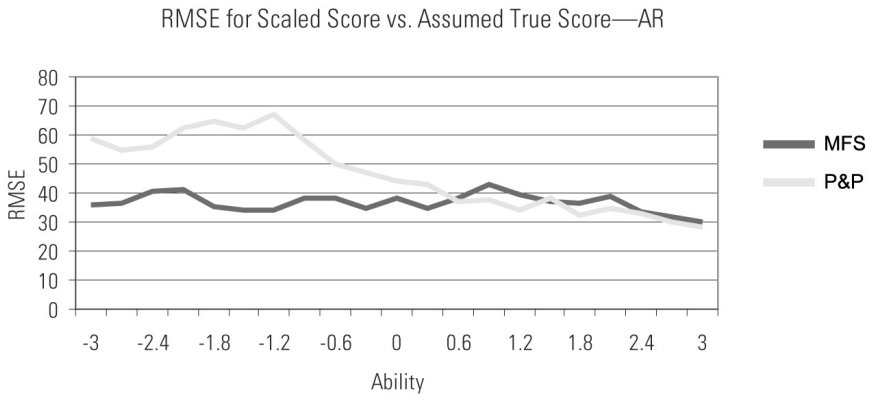

FIGURE 4. *The root mean squared error of scaled scores for a 36-item Analytical Reasoning MFS compared with a 46-item P&P test*

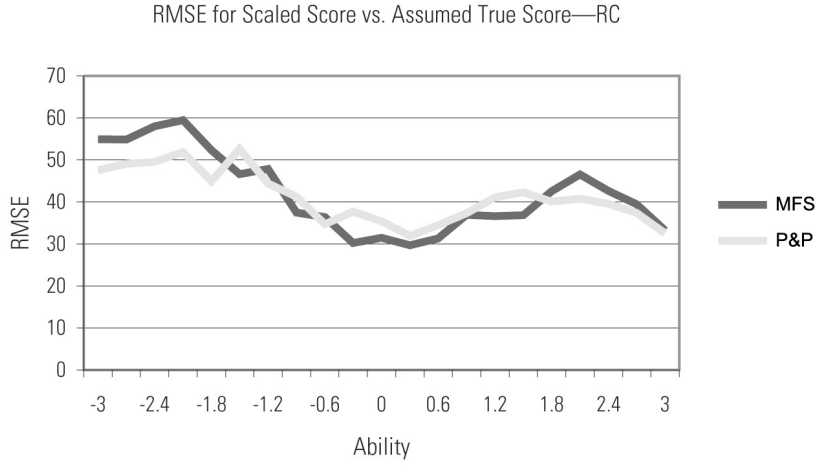

FIGURE 5. *The root mean squared error of scaled scores for a 36-item Reading Comprehension MFS compared with a 52-item P&P test*

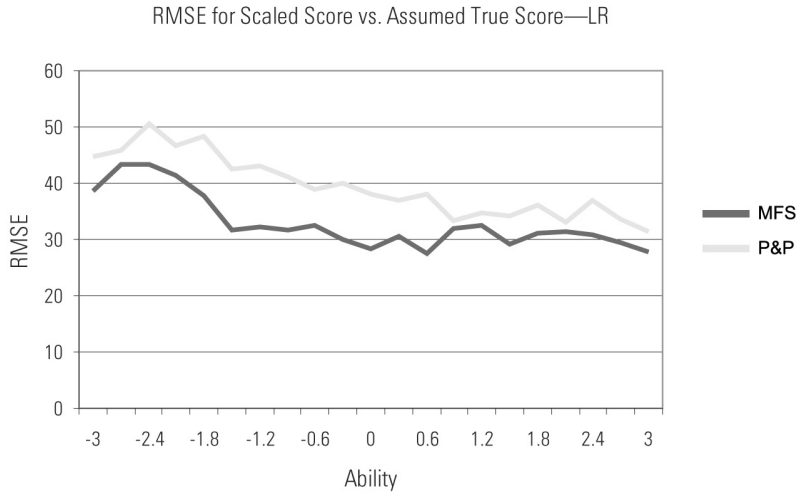

FIGURE 6. *The root mean squared error of scaled score for a 36-item Logical Reasoning MFS compared with a 52-item P&P test*

TABLE 4
RMSE of the scaled score for an MFS and P&P at each of the quadrature points from –3.0 to +3.0 for all three sections—AR, RC, and LR

	Analytical Reasoning (AR)		Reading Comprehension (RC)		Logical Reasoning (LR)	
Ability	MFS	PNP	MFS	PNP	MFS	PNP
–3.0	36.14	59.08	54.91	47.54	38.71	44.63
–2.7	36.67	54.45	54.83	49.05	43.38	45.82
–2.4	40.80	56.07	57.99	49.56	43.35	50.47
–2.1	41.08	62.31	59.47	51.88	41.27	46.75
–1.8	35.30	64.45	52.29	44.95	37.65	48.27
–1.5	34.04	62.61	46.60	52.61	31.60	42.39
–1.2	33.93	66.95	47.88	44.41	32.29	43.09
–0.9	38.21	58.50	37.46	41.19	31.55	40.97
–0.6	38.44	50.15	36.37	34.70	32.54	38.82
–0.3	34.53	47.10	30.21	37.68	29.92	40.10
0.0	38.29	44.03	31.46	35.41	28.32	38.10
0.3	34.83	42.92	29.66	31.93	30.54	36.91
0.6	38.24	36.96	31.32	34.49	27.47	38.05
0.9	42.96	37.65	36.91	37.29	31.82	33.47
1.2	39.34	34.23	36.58	41.11	32.47	34.65
1.5	36.78	38.18	36.88	42.34	29.06	34.30
1.8	36.20	32.13	42.52	40.10	31.25	36.10
2.1	39.08	34.82	46.55	40.77	31.38	33.12
2.4	33.59	32.68	42.61	39.52	30.77	36.87
2.7	31.55	30.17	39.45	37.41	29.47	33.59
3.0	30.10	28.25	33.24	32.55	27.90	31.25

Discussion

A standard CAT chooses items from an item pool as the test is delivered. This paper discusses a hybrid CAT that maintains benefits from traditional paper-and-pencil exams by preselecting items and grouping them into testlets. The ordered collection of testlets comprising the potential items to deliver is called a multiple form structure (MFS). MFSs can be assembled using the mathematical programming techniques outlined here. The test adapts to the ability of examinees by allowing several paths through the MFS, providing better measurement efficiency than for a similar P&P instrument. However, by delivering an MFS over a limited period of time to a large number of examinees, the benefits of the P&P test, such as exposure control and test equating, can be approached.

The use of MFSs was illustrated with data from the Law School Admission Council (LSAC) and enforcing constraints found in the Law School Admission Test (LSAT). The MFS is easily implemented once the design has been determined. The forms are predetermined and items to deliver are chosen based on simple rules. The simulations show that the MFS approach is a viable method to implement CATs. An MFS may achieve comparable reliability with a corresponding standard CAT when exposure controls require almost uniform item usage in the CAT. This was found to be the case with simulations where the exposure control parameter was significantly larger than the value used to create the displayed results.

This paper provided the fundamentals for creating the MFS designs and assembling MFSs. The constraints on testlets and paths were provided by test specialists. The targets for each MFSD bin were created by considering the population percentile designated for the bin and the characteristics of the item pool. Routing rules were designed to direct examinees from a percentile group to the bin designated for the group.

References

Armstrong, R. D. (2003). *Routing rules for multiple form structures* (Research Report CT-02-08). Newtown, PA: Law School Admission Council, Inc.

Armstrong, R. D., Jones, D. H., & Kunce, C. S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement, 22*, 237–247.

Armstrong, R. D., & Roussos, L. (2003). *A method to determine targets for multistage adaptive tests* (Research Report CT-02-07). Newtown, PA: Law School Admission Council, Inc.

Bock, R., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275–285.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item pools. *Journal of Educational Statistics, 15*, 129–145.

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

Fisher, M. L. (1985). An applications oriented guide to Lagrangian relaxation. *Interfaces, 15*, 10–21.

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

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. New York: Springer.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*(3), 229–249.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36,* 227-242.

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

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

Reese, L. M., & Schnipke, D. L. (1996, June). *An evaluation of a two-stage testlet design for computerized adaptive testing*. Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365–389.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50*, 411–420.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195–211.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

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

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