A Method for Severely Constrained Item Selection in Adaptive Testing

Martha L. Stocking and Len Swanson

Applied Psychological Measurement 1993; 17; 277
DOI: 10.1177/014662169301700308

A Method for Severely Constrained Item Selection in Adaptive Testing

Martha L. Stocking and Len Swanson

Educational Testing Service

Previous attempts at incorporating expert test construction practices into computerized adaptive testing paradigms are described. A new method is presented for incorporating a large number of constraints on adaptive item selection. The methodology emulates the test construction practices of expert test specialists, which is a necessity if computerized adaptive testing is to compete with conventional tests. Two examples—one for a verbal measure and the other for a quantitative measure—are provided of the successful use of the proposed method in designing adaptive tests. *Index terms: adaptive test design, computerized adaptive testing, constrained adaptive testing, expert systems, test assembly algorithms.*

Conventional tests administered using paper-and-pencil to large numbers of examinees simultaneously have been a fixture of educational testing and measurement for many years. This testing strategy represents vastly reduced unit costs over tests administered individually, which were used during the early part of this century.

However, interest in restoring some of the advantages of individualized testing has never completely disappeared. Turnbull suggested investigations in this direction in 1951 and coined the phrase *tailored testing* to describe this mode of test administration (Lord, 1980, p. 151). Possibilities for constructing individualized tests became likely with the advent of item response theory (IRT; Lord, 1952, 1980). In the 1960s, Lord (1970, 1971a) began to explore this application of IRT by investigating various item selection strategies borrowed from the bioassay field. Later work by Lord (1977, 1980) and Weiss (1976, 1978) laid the foundation for the application of adaptive/tailored testing as an alternative to conventional testing.

Adaptive tests are tests in which items are selected to be appropriate for the examinee—the test *adapts* to the examinee, usually by selecting items of appropriate difficulty. Computerized adaptive testing (CAT) has received increasing attention as a practical alternative to paper-and-pencil tests as the cost of modern computing technology has declined. The Department of Defense has seriously considered using CAT for the Armed Services Vocational Aptitude Battery (CAT-ASVAB; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990); large testing organizations have implemented CAT [e.g., the Computerized Placement Tests program (College Board, 1990)]; and certification and licensure organizations are in the process of implementing CAT as a viable alternative to conventional paper-and-pencil tests (Zara, 1990; Zara, Bosma, & Kaplan, 1987).

The advantages of adaptive testing are well documented (e.g., Wainer et al., 1990, chap. 1). As experience with and knowledge about adaptive testing has accumulated, the psychometric foundations of CAT have become better understood, although reasonable debate still exists concerning the "best" psychometric procedures to employ. The psychometrics of CAT are more tractable now than they have been in the past, and the issues of how best to transfer the nonpsychometric aspects of

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 17, No. 3, September 1993, pp. 277-292

0146-6216/93/030277-16$2.05

good test construction principles to this new environment have received increasing attention.

Conventional Test Construction

To construct tests, items must be written, edited, and assembled into test forms. Naturally, testing organizations seek the most efficient test possible for each measurement purpose. Item selection is usually subject to various rules, called *test specifications*, that constrain the selection of items for test forms.

Test specifications can be classified into four categories: (1) constraints on intrinsic item properties, (2) constraints on item features in relation to all other candidate items (overlap), (3) constraints on item features in relation to a subset of all other candidate items (item sets), and (4) constraints on the statistical properties of items as derived from pretesting.

Constraints on Intrinsic Item Properties

Items have a number of intrinsic properties of interest to the test specialist. Most obvious are the content or subject matter of the item, and also the type or appearance of the item to the examinee. For example, a sentence completion item with science content may have one blank or two blanks, a math item on right triangles may be a quantitative comparisons item or a problem solving item, an item may have four or five answer alternatives, and so forth. Less apparent may be references to certain population subgroups extraneous to the purpose of the test. Finally, the location of the correct answer, the variety engendered by including items from multiple authors, and other similar item features may be important for the test specialist to control. Note that these features are not mutually exclusive; that is, an item may—and probably will—contribute to the satisfaction of several constraints simultaneously.

Overlap Constraints

Items may relate to each other in at least four ways that influence test quality:

1. An item may give away the answer to another item, which is described by Wainer & Kiely (1987) as *cross-information* and also described by Kingsbury & Zara (1991).
2. Items may be redundant—testing the same or nearly the same point but appearing different at a casual glance.
3. Items may have similar incidental features that are not related to the purpose of the test: for example, two reading comprehension passages about science.
4. Items may use relatively uncommon words in incorrect answer alternatives. If such words appear more than once in a test, examinees with more limited vocabularies may be doubly disadvantaged in a manner that is extraneous to the purpose of the test.

Item Set Constraints

Some items are related to each other through their relationship to common stimulus material, as when a number of items are based on a common reading passage in a verbal test, or when a number of items are based on a common graph in a mathematics test. Some items are related to each other through some other feature, such as having common directions. For example, a verbal test might include synonyms and antonyms, which might be confusing to examinees if such items were intermixed; it would be less confusing if all synonym items appeared together and all antonym items appeared together. The implication of an *item set* is that items belonging to a set should not be intermixed with other items not belonging to the set. Moreover, some or all members of the set might be included in the test, but no more than one subset of the set may be included.

Constraints on Statistical Properties

Test constraints typically constrain the selection of items based on their statistical properties in order to construct test forms that have the desired measurement properties. These constraints may take the form of specifying some target aggregation of statistical properties, such as a target test information function (Stocking, Swanson, & Pearlman, 1993; Van der Linden & Boekkooi-Timminga, 1989).

Adaptive Test Construction

Early monte carlo investigations of adaptive testing algorithms concentrated predominantly on the psychometric aspects of test construction (see Lord, 1970, 1971a, 1971b). Such investigations eventually led to IRT-based algorithms that were fast, efficient, and psychometrically sound. A review of the most frequently used algorithms is given in Wainer et al. (1990, chap. 5) and Lord (1980, chap. 9). The fundamental philosophy underlying these algorithms is as follows:

1. An item is selected on some basis and administered to the examinee.
2. The examinee's response to the item is used to update an estimate of trait level.
3. The selection of the next item is based partly on updated trait level estimates. This continues until some stopping criterion is met.
4. The examinee's final score is the trait estimate after all items are administered.

When practical implementation became a possibility, researchers began to address the incorporation of good test construction practices as well as psychometric considerations into the selection of items in adaptive testing.

One of the first to do so was Lord (1977) in his Broad Range Tailored Test of Verbal Ability. The item pool for this adaptive test consisted of five different types of discrete verbal items. To insure that each adaptive test measured the same construct, some mechanism was necessary to insure that each adaptive test contained the same mix of item types. The mechanism employed specified the sequence of item types in advance. Items were selected based on maximum item information for items of the appropriate prespecified type in the sequence at an examinee's (maximum-likelihood-based) estimated trait level.

Specifying the sequence of item types in advance can be extended to handle relatively large numbers of item types, as in the Computerized Placement Tests (Ward, 1988) in which there are 10 to 15 item types. The same kind of control is used in the CAT-ASVAB (Segall, 1987). This type of content control has been called a constrained CAT (C-CAT) by Kingsbury & Zara (1989).

A major disadvantage of this approach is that it assumes that the item features of interest partition the item pool into mutually exclusive subsets. Given the number of item features that may be of interest to test specialists, the number of mutually exclusive partitions can become very large and the number of items in each partition can become quite small. Moreover, incorporating considerations of overlap and item sets requires further partitioning by overlap group and by set, thereby further enlarging the number of mutually exclusive partitions. Wainer & Kiely (1987) hypothesized that the use of testlets could overcome these problems. They suggested that an adaptive test be constructed from testlets by using the testlet rather than an item as the branching point. They hypothesized that this would enable test specialists to enforce constraints on intrinsic item features, overlap, and item sets in the same manner as is currently done with conventional tests.

Kingsbury & Zara (1991) compared the measurement efficiency of the testlet approach to the C-CAT approach. They found that the testlet approach required from 4 to 10 times the test length of the C-CAT approach to achieve the same level of precision. Aside from measurement concerns, the

testlet approach rests on the idea that the pool of available items can be easily subdivided into mutually exclusive subsets (testlets), which is also a disadvantage of the C-CAT approach.

A New Methodology

The foundation of this new methodology for incorporating expert test development practices in the construction of adaptive tests is the application of a weighted deviations model (WDM) and algorithm for item selection (Swanson & Stocking, 1993). This WDM and algorithm were developed in the context of conventional test assembly paradigms that have been proposed in the literature over the last 10 years. Typically, these paradigms employ a combination of IRT, computers, and linear programming models to optimize some aspect of the items selected. Examples of other such paradigms include Theunissen (1985), Van der Linden (1987), Van der Linden & Boekkooi-Timminga (1989), and Ackerman (1989).

The weighted deviations algorithm was developed and investigated in many conventional test construction problems using real item pools (Stocking, Swanson, & Pearlman, 1991, 1993) and found to be satisfactory in its capability of handling constraints on intrinsic item features. The handling of constraints on overlap was added to the weighted deviations algorithm for its application to adaptive testing, although such a feature would be useful in the context of conventional test assembly as well.

In traditional binary programming models applied to linear test construction, test constraints are formulated mathematically as linear constraints. The test construction problem is then cast as a binary programming model, where some objective function is minimized or maximized subject to the test constraints (e.g., Theunissen, 1985; Van der Linden & Boekkooi-Timminga, 1989). However, such models do not always have a feasible solution because one or more of the constraints cannot be satisfied simultaneously with some other (often non-mutually exclusive) constraint(s).

The WDM resolves this problem by treating test *constraints* as *desired properties*, and moving them to the objective function. This is done through a device commonly used in mathematical programming to make an infeasible model feasible (Brooke, Kendrick, & Meeraus, 1988, p. 159). The model also provides a means of weighting the desired properties so that the test specialist can exercise some control over which constraints are most important and which can be relaxed.

A basic version of the WDM has the following mathematical form when used in the context of adaptive testing, in which items are selected for inclusion in the test one at a time. Let $i = 1, \ldots, N$ index the items in the pool, and let x_i denote the decision variable that determines whether item i is included in ($x_i = 1$) or excluded from ($x_i = 0$) the test. Let $j = 1, \ldots, J$ index the item properties associated with the nonpsychometric constraints. Let L_j and U_j be the lower and upper bounds (which may be equal), respectively, on the number of items in the test having each such property, and let a_{ij} be 1 if item i has property j and 0 if it does not.

For adaptive testing, items are selected that have the largest amount of item information at the current estimate of the examinee's trait level. Let θ be the point on the trait metric of the current trait estimate. Let $I_i(\theta)$ be the item information for item i at θ. Then the model is as follows.

Minimize

$$\sum_{j=1}^{J} w_j d_{L_j} + \sum_{j=1}^{J} w_j d_{U_j} + w_\theta d_\theta \,, \tag{1}$$

where w_j is the weight assigned to constraint j, and w_θ is the weight assigned to the information constraint, subject to

$$\sum_{i=1}^{N} a_{ij} x_i + d_{L_j} - e_{L_j} = L_j, \quad j = 1, \ldots, J \,, \tag{2}$$

$$\sum_{i=1}^{N} a_{ij}x_i - d_{U_j} + e_{U_j} = U_j, \quad j = 1, \ldots, J, \tag{3}$$

$$\sum_{i=1}^{N} I_i(\theta)x_i + d_\theta - e_\theta = \infty, \tag{4}$$

$$d_{L_j}, d_{U_j}, e_{L_j}, e_{U_j} \geq 0, \quad j = 1, \ldots, J, \tag{5}$$

$$d_\theta, e_\theta \geq 0, \tag{6}$$

and

$$x_i \in \{0, 1\}, \; i = 1, \ldots, N. \tag{7}$$

The d_{L_j}, e_{L_j}, d_θ, and e_θ are non-negative *slack* variables. The ds are interpreted as the positive (or zero) deviations from the lower bounds; that is, the differences between the lower bounds and the sums whenever the lower bounds are not met. Similarly, the es represent the differences between the sums and the lower bounds whenever the lower bounds are exceeded, and might be interpreted as the unneeded *surplus* quantity. Note that for a given j and for θ, one or both of these variables must take on the value 0 (that is, the sum cannot both exceed and fail to meet the lower bound). The d_{U_j} and e_{U_j} have a similar interpretation with respect to the upper bounds (Equation 3; see Swanson & Stocking, 1993, Figure 1).

Note that conformance to maximizing item information at θ is expressed by Equation 4. This is simply a special form of Equation 2. For practical purposes, the lower bound in Equation 4 is taken as a large positive number unobtainable in practice.

For convenience of terminology *constraints* are referred to in the sense in which test specialists think of them, recognizing that they are not constraints in the mathematical sense of binary programming. Equations 2–4 reflect the mathematical formalism by which the test specialist's constraints are transferred to the objective function. Extensions of this basic form of the model for handling complications such as item sets were described in Swanson & Stocking (1993).

Swanson & Stocking (1993) also provided an heuristic for solving the WDM. Although the WDM can be solved using standard mixed integer linear programming algorithms, very large test construction problems are difficult to solve with such methods. The heuristic is a *goal seeking* or *greedy* heuristic (Nemhauser & Wolsey, 1988, chap. II.5), similar to those used elsewhere in test construction applications. It essentially consists of three steps:

1. For every item not already in the test, compute the deviation for each of the constraints if the item were added to the test.
2. Sum the weighted deviations across all constraints.
3. Select the item with the smallest weighted sum of deviations.

This describes the general process, but the heuristic also includes provisions for avoiding local optimality, and for incorporating overlap considerations.

Ideally, all test constraints will be met and the adaptive test will be the most informative possible for an examinee. However, there is nothing about the WDM or its algorithm that guarantees this outcome. The only guarantee is that the test produced will come as close as possible to the ideal, given the structure of the constraints and the item pool.

Constraints on Intrinsic Item Properties

The control of intrinsic item features is accomplished through the use of explicit constraints; that is, lower and upper bounds (which may be equal) on the desired number of items that possess a

feature. For example, "2 ≤ number of Type A items ≤ 2" is a specification which denotes that exactly two Type A items are to be included in the test. Positive deviations from each of the bounds are minimized through the objective function.

Overlap Constraints

It is possible to control for overlap among items by coding overlapping items to a sufficient level of detail and then explicitly specifying that only one such item may be included in a test. In practice, it may be difficult to develop a sufficiently detailed item coding scheme so that overlap can be controlled by the imposition of explicit constraints alone. Instead, another mechanism must be employed—that of overlap groups.

An overlap group consists of a list of items that may not appear together in the same adaptive test. Overlap groups do not imply transitivity of overlap. That is, Item A may overlap with Item B, and Item B may overlap with Item C, but that does not imply that Item A overlaps with Item C because the reasons for the overlap may be different.

Item Set Constraints

Theunissen (1986, p. 387) suggested that sets of items could be incorporated into a maximum information adaptive testing paradigm by using a set information function, which is the sum of the item information functions for the items comprising that set. This approach is effective if the tests being constructed are made up entirely of item sets and the number of items to be administered from each set is known in advance.

However, a complication occurs in the more general case when the items to be administered from a set of items are not specified in advance. For example, stimulus material is usually pretested with many more items than would be desirable to include in any single test, and the subset of items administered in a particular adaptive test depends on the current estimate of examinee trait level, although the size of the subset may be specified in advance. In this context, the set of information functions for all possible subsets of a set of items must be computed.

The approach taken here is consistent with Theunissen's (1986) suggestion in that partial sums of item information functions are computed as items (including items from a set) are administered. Each set is assigned a conceptual partition of the item pool (a block); items not belonging to sets are not considered to be in such partitions. Blocks may be designated as re-enterable, with a fixed number of items to be administered at each entry, or not re-enterable, with a fixed number of items to be administered on a single entry.

Blocks are entered (or possibly re-entered) by the selection of an item in that block that contributes the most to the satisfaction of all other constraints and does not appear in an overlap group containing an item already administered. Once within a block, items continue to be selected adaptively for administration based on their contribution to the satisfaction of all constraints and overlap, until the number of items to be administered at that entry into the block is reached.

Summary

Using an augmented WDM and algorithm, a mechanism for selecting items in adaptive testing that mirrors as closely as possible the considerations that govern the assembly of conventional tests was developed. The next item administered in an adaptive test is the item that simultaneously is the most informative item at an examinee's estimated θ level, and contributes the most to the satisfaction of all other constraints, in addition to the constraint on item information. At the same time, it is required that the item does not appear in an overlap group containing an item already administered,

and is in the current block (if the procedure is in a block), starts a new block, or is in no block.

Example 1: A Verbal Adaptive Test

A verbal adaptive test was examined using a monte carlo simulation study to determine test properties. The goal was to construct a test as parallel as possible in terms of content to an existing conventional 85-item paper-and-pencil test and to achieve an estimated reliability of .91, the average reliability of the most recent 10 editions of the conventional test, in the shortest possible (fixed) adaptive test length.

The Process

Several test design decisions were made. The difficulty of the first item was selected to be approximately 1 standard deviation (SD) below the mean of the θ distribution. This was to simulate providing an initial "success experience" to examinees, similar to that provided by typical conventional tests when the items are ordered from easy to difficult. The score reporting metric for this adaptive test was the number-correct "true score" on a reference set of 85 items. The final (maximum likelihood) θ estimate was converted to this metric using the test response function (Lord, 1980, Equation 4-9). This reference set of items was actually an intact conventional paper-and-pencil edition of the parent form, calibrated and placed on the same IRT metric as the item pool.

The optimum (fixed) adaptive test length and the relative weights to be given to each specification to achieve the desired measurement and content goals were unknown at the beginning of this study. Test specialists constructed constraints on item features appropriate for a range of different adaptive test lengths from 25 items to 35 items in steps of one item.

The shortest test length was selected and a number of simulations were tried with different relative weights on all constraints in successive attempts to satisfy content and measurement goals at that test length. Because satisfaction could not be obtained, test length was increased, constraints and weights were changed by test specialists, and the process was repeated. This iterative process continued until a test design was found that satisfied both measurement and content goals, at which point design parameters were fixed and the adaptive test was ready for live administration. The final test design was a fixed-length test of 27 items.

The Item Pool

The item pool contained 518 verbal items, 197 of which were associated with 51 reading passages, giving a total of 569 elements (items plus stimuli) in the pool. All items had been calibrated on large samples (>2,000) from the current testing population using the three-parameter logistic item response model and the computer program LOGIST (Wingersky, 1983). The mean estimated discrimination (a) value for the items in the pool was .86, with SD = .28, and a range of .22 to 1.83. The mean estimated difficulty (b) was .17, with SD = 1.31, and a range from -3.68 to 3.32. The mean estimated pseudo-guessing parameter (c) value was .17 with SD = .09, and a range of 0.00 to .50.

The Content Constraints

Items and passages in the pool were classified according to 41 different features. Table 1 shows the 41 constraints developed by test specialists for a 27-item adaptive test. The number of passages or items in the pool that were identified as having each specific property also is listed (n). The WDM actually employs a single constraint for every feature that has equal lower and upper bounds, and two constraints for every feature with unequal lower and upper bounds. Thus, from the perspective of the WDM, the constraints represented a total of 71 constraints [11 + (2 × 30)]. However, for ease

Table 1
Content Constraints (L_j and U_j), Weights (W), and n Elements in the Pool for the Adaptive Verbal Test

Specification	Description	L_j	U_j	W	n
1	Type 1 Passages	2	2	20	26
2	Type 2 Passages	1	1	20	25
3	Science Passages	1	1	20	11
4	Humanities Passages	0	1	2	8
5	Social Science Passages	0	1	1	13
6	Argumentative Passages	0	1	1	10
7	Narrative Passages	0	1	1	9
8	Type A Passages	0	1	20	16
9	Type B Passages	0	1	1	2
10	Type I Passages	1	1	20	14
11	Reading Comprehension Items (RCMP)	8	8	20	197
12	RCMP Type 1 Items	1	4	1	35
13	RCMP Type 2 Items	1	4	1	52
14	RCMP Type 3 Items	2	5	1	58
15	RCMP Type 4 Items	1	4	1	52
16	Sentence Completion Items (SNCP)	5	5	20	95
17	SNCP Aesthetic/Philosophical Items	1	2	3	25
18	SNCP Practical Affairs Items	1	2	3	27
19	SNCP Science Items	1	2	3	21
20	SNCP Human Relations Items	1	2	3	22
21	SNCP Type 1 Items	2	2	20	40
22	SNCP Type 2 Items	3	3	20	55
23	SNCP Type A Items	0	1	20	16
24	SNCP Type B Items	2	2	20	25
25	Analogies	6	6	20	85
26	Aesthetic/Philosophical Analogies	1	2	1	20
27	Practical Affairs Analogies	1	2	1	21
28	Science Analogies	1	2	1	26
29	Human Relations Analogies	1	2	1	18
30	Type 1 Analogies	1	3	1	24
31	Type 2 Analogies	1	3	1	30
32	Type 3 Analogies	0	1	1	15
33	Type 4 Analogies	0	1	1	11
34	Antonyms (ANTM)	8	8	20	141
35	Aesthetic/Philosophical Antonyms	1	2	1	30
36	Practical Affairs Antonyms	1	2	3	30
37	Science Antonyms	1	2	3	38
38	Human Relations Antonyms	1	2	1	43
39	Type 1 Antonyms	1	4	1	36
40	Type 2 Antonyms	1	4	1	32
41	Type 3 Antonyms	1	4	1	73

of discussion the test specialists' perspective of 41 constraints on item features was adopted. Also shown in Table 1 are the relative weights given to the satisfaction of each specification in the final test design. The weight given the constraint on item information (not shown in Table 1) was 15. These relative weights were arrived at through the iterative process described above. Constraints with the highest weight (20) were so important that they could not be violated or the resultant adaptive test would not be judged acceptable.

Overlap Constraints

Table 2 gives a portion of the set of overlap groups constructed by test specialists after careful examination of the pool. Items could be indicated as overlapping with other items and/or with passages. For this pool of 518 items and 51 passages, there were 528 such groups with 1,358 entries. For example, the table shows that Group 1 contains Items 232, 22, 242, and 103, so that at most one of these four items could be selected.

Table 2
A Portion of the Overlap Groups for the Adaptive Verbal Test

Group	Number in Group	Items/Passages in Group
1	4	232, 22, 242, 103
2	3	232, 218, 79
3	3	232, 298, 307
.	.	.
.	.	.
.	.	.
250	3	321, 284, 281
251	4	321, 305, 281, 308
252	3	38, 240, 142
.	.	.
.	.	.
.	.	.
526	2	449, 550
527	2	518, 556
528	2	518, 565

Item Sets

Table 3 displays a portion of the list of blocks of pool elements that were to be considered as sets. For this example, none of the blocks were re-enterable and every item appeared in a block. For this pool, there were a total of 54 logical blocks.

Table 3
A Portion of the List of Blocks for the Adaptive Verbal Test

Block	Number to Select	Starting Position	Ending Position	Classification
1	5	1	95	SNCP
2	6	96	180	Analogies
3	8	181	321	ANTM
4	3	322	327	Type 1 Passage
5	3	328	333	Type 1 Passage
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
52	2	556	559	Type 2 Passage
53	2	560	564	Type 2 Passage
54	3	565	569	Type 1 Passage

The Simulations

The simulation was performed for 200 simulees at each of 15 values on the reported score metric ranging from just above the chance level to just below a perfect score. These 15 values were nearly equally spaced on the score metric, and unequally spaced on the trait metric. A randomization scheme was imposed to improve item security, in which the first item was randomly selected from a list of

the eight best items, the second item was randomly selected from a list of the seven best items, and so forth. The eighth and subsequent items were selected to be optimal based on the WDM.

Results

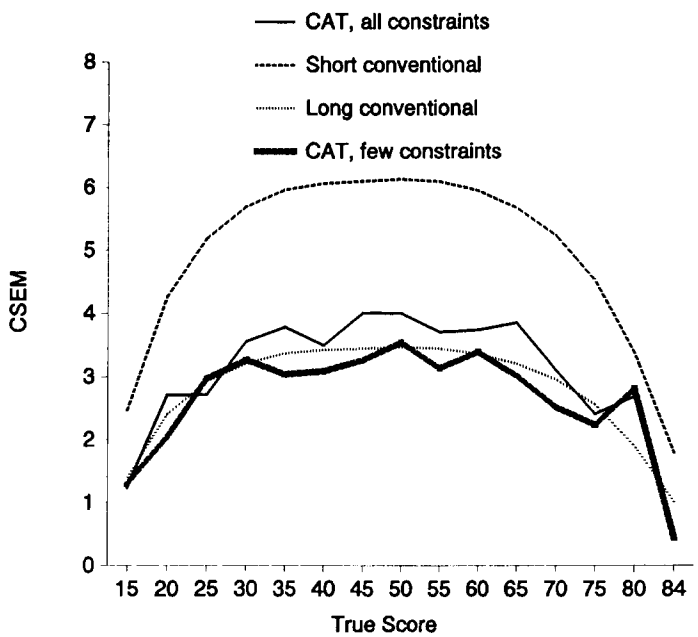
The results of the simulation were evaluated both conditional on score level and unconditionally. To perform the unconditional evaluations, the item parameters and item responses from a group of over 6,000 examinees who took an edition of the 85-item parent form were used to compute an estimated distribution of true θ using the method of Mislevy (1984). Proportional values of this distribution were applied to the conditional results to yield an estimate of the unconditional results in a "typical" group of test takers.

The estimated reliability of the 27-item adaptive test was .91, computed using the method of Green, Bock, Humphreys, Linn, & Reckase (1984, Equation 6). This reliability was achieved by using 295 items (out of 518) and 28 reading passages (out of 51) at least once across 3,000 simulated adaptive tests. The reliability of the conventional reference test used for scoring purposes was .93, making the conventional reference test more reliable than the average conventional test.

Figure 1 displays more detail about the measurement properties of the tests. Four conditional standard error of measurement (CSEM) functions are displayed for (1) the 85-item reference test (Lord, 1980, Equation 4-8), referred to as *long conventional*; (2) the reference test scaled to a length of 27 items, referred to as *short conventional*; (3) the adaptive test with constraints on content, overlap, and sets, referred to as *CAT, all constraints*; and (4) an adaptive test with no constraints on content or overlap, and only the reading passage constraints on sets, referred to as *CAT, few constraints*.

The CAT with few constraints represents the best that can be done in terms of measurement alone from this pool when the only block structure preserved is that of items associated with reading

Figure 1
CSEM for the 85-Item Conventional Verbal Test, the 27-Item Conventional Verbal Test, the 27-Item Adaptive Verbal Test With All Constraints, and the 27-Item Adaptive Verbal Test With Few Constraints



passages. The reliability of this test was .94 and was attained using 144 out of the 518 items and 16 out of the 51 reading passages at least once across 3,000 simulated adaptive tests. However, this test was unsatisfactory from a content perspective. All of a typical group received adaptive tests that violated the constraint that exactly six analogy items were to be administered; the average number of analogies was .9. Approximately 91.1% of the group violated the constraint that exactly five sentence completion items were to be administered; the average number administered was 9.5. 79.3% of the group violated the specification that exactly eight antonyms were to be administered; the average number was 8.7. 12% and 14% of the group violated the constraints that exactly one Type 1 passage and two Type 2 passages were to be included, with averages of 1.9 and 1.3, respectively.

The difference between the two adaptive test CSEM functions, particularly throughout the middle of the true score range, represents the penalty imposed, in terms of measurement, by the attempt to satisfy content constraints, overlap constraints, and set constraints simultaneously with psychometric constraints. The adaptive test with full constraints was as good as it was because it used more than twice the number of items and nearly twice the number of passages in the pool in order to satisfy as many of the test constraints as possible.

The adaptive test with full constraints specified achieved its CSEM without violating overlap constraints or set constraints. However, some content constraints were violated. Table 4 displays, for each constraint that had some violation, the percent of a typical population that could be expected to experience such violations (%) and the typical extent of such violations. The number of

Table 4
Content Constraints (L_j and U_j), Weights (W), Number of Elements (n), Percent Content Constraint Violations in Typical Group (%), and Average Number of Items (AI) for the Adaptive Verbal Test

Specification	Description	L_j	U_j	W	n	%	AI
4	Humanities Passages	0	1	2	8	3.9	.49
7	Narrative Passages	0	1	1	9	2.3	.63
12	RCMP Type 1 Items	1	4	1	35	1.8	1.2
13	RCMP Type 2 Items	1	4	1	52	25.9	1.6
14	RCMP Type 3 Items	2	5	1	58	6.4	2.3
15	RCMP Type 4 Items	1	4	1	52	10.8	3.0
17	SNCP Aesthetic/Philosophical Items	1	2	3	25	13.8	1.3
18	SNCP Practical Affairs Items	1	2	3	27	12.5	1.2
19	SNCP Science Items	1	2	3	21	11.4	1.4
20	SNCP Human Relations Items	1	2	3	22	13.6	1.1
26	Aesthetic/Philosophical Analogies	1	2	1	20	30.2	1.8
27	Practical Affairs Analogies	1	2	1	21	1.7	1.2
28	Science Analogies	1	2	1	26	5.2	1.4
29	Human Relations Analogies	1	2	1	18	40.0	1.6
30	Type 1 Analogies	1	3	1	24	1.5	2.1
31	Type 2 Analogies	1	3	1	30	9.8	1.8
32	Type 3 Analogies	0	1	1	15	4.9	.6
35	Aesthetic/Philosophical Antonyms	1	2	1	30	23.8	1.5
36	Practical Affairs Antonyms	1	2	1	30	6.4	1.8
37	Science Antonyms	1	2	1	38	28.7	2.0
38	Human Relations Antonyms	1	2	1	43	64.7	2.6
39	Type 1 Antonyms	1	4	1	36	.6	2.4
40	Type 2 Antonyms	1	4	1	32	4.0	2.0
41	Type 3 Antonyms	1	4	1	73	18.4	3.6

elements administered for each constraint, averaged over the typical distribution (AI), rarely violated the constraint.

However, the conditional average number of items at each trait level (not displayed in Table 4) showed that constraint violations tended to occur when there was a relationship between items with a particular feature and the appropriateness of the item for a particular trait level. For example, 30.2% of the typical population had adaptive tests that violated the constraint that between 1 and 2 analogy items on the aesthetic/philosophical topics were to be included. A substantial proportion of simulees with below average true trait levels were administered three such items. Likewise, 64.7% of the typical population had adaptive tests that violated the constraint that between 1 and 2 human relations antonym items were to be included. A substantial proportion of simulees with above average true trait levels were administered three or four such items.

These constraint violations would be reduced, and possibly disappear, if it were possible to obtain items appropriate for all trait levels that also had all of the features of interest. However, this may not be economically feasible.

Example 2: A Quantitative Adaptive Test

In this example, the methodology was applied to the design of a quantitative adaptive test. This effort had three major goals: (1) the adaptive test must be as short as possible and as parallel as possible in terms of test content to existing 60-item conventional paper-and-pencil forms, (2) the adaptive test must have approximately the same reliability as the paper-and-pencil forms, and (3) no item was to be exposed or administered to more than 20% of a typical population of test takers. Although the first two goals were similar to those for the verbal test design example, the concerns for item security were different in the quantitative context.

The Process

As for the verbal test, the difficulty of the first item was selected to be 1 SD below the mean of the θ distribution. The score reporting metric for the quantitative adaptive test was the number-correct true score metric for a 60-item conventional paper-and-pencil edition of the parent form, calibrated and placed on the IRT scale of the item pool. The same kind of process that was used for the verbal test was used for the quantitative test to find a minimum test length and a set of relative constraint weights that lead to the satisfaction of the measurement and content goals of this example.

The Item Pool

There were 496 quantitative items, 154 of which were associated with 22 stimuli, such as tables or graphs, for a total of 518 elements (items plus stimuli) in the pool. The mean estimated a for the items in the pool was .82 with SD = .33, and a range of .23 to 1.84. The mean estimated b was .04 with SD = 1.24, and a range from -4.64 to 2.65. The mean estimated c was .14 with SD = .11, and a range of 0.00 to .50.

The Content Constraints

Items and sets were classified according to 24 different features, as shown in Table 5. Items not in sets are of two different major types—problem solving (PS) and quantitative comparisons (QC). This table also shows the number of items and stimuli in the pool (n), as well as the desired number of elements with each feature in a 25-item adaptive test (L_j and U_j). Note that, in contrast to the verbal test, stimuli do not have additional features that must be controlled; they are simply identified as data interpretation (DI) stimuli.

Table 5
Content Constraints (L_j and U_j), Weights (W), Number of Elements (n), Percent Content Constraint Violations in Typical Group (%), and Average Number of Items (AI) for the Adaptive Quantitative Test

Specification	Description	L_j	U_j	W	n	%	AI
1	Data Interpretation Stimuli	2	2	11	22		
2	QC 1	5	5	10	62		
3	QC 2	4	4	10	73		
4	QC 3	4	4	10	69		
5	PS 1	3	3	10	53		
6	PS 2	3	3	10	44		
7	PS 3	2	2	10	41		
8	DI Item Type 1	4	4	1	154		
9	Type 4	8	8	1	246		
10	QC Type 4	2	2	10	36		
11	PS Type 4	2	2	10	56		
12	QC 1 Type 4	0	1	1	12	7.1	.75
13	QC 2 Type 4	0	1	1	11	1.0	.52
14	QC 3 Type 4	0	1	1	13	4.2	.72
15	PS 1 Type 4	0	1	1	28	1.8	.54
16	PS 2 Type 4	0	1	1	16	17.5	1.03
17	PS 3 Type 4	0	1	1	12	.1	.43
18	Type 5	0	1	1	11	.3	.26
19	Type 6	0	1	1	8		
20	QC Type 7	1	10	1	53	1.3	2.83
21	QC Type 8	1	10	1	58	.1	3.09
22	QC Type 9	1	10	1	54		
23	QC Type 10	1	10	1	39	.2	3.38
24	Type 11	1	12	1	63	.3	3.48

Overlap and Item Blocks

The overlap structure for the quantitative pool had 88 overlap groups with 318 entries, in contrast with 528 groups with 1,358 entries for the verbal pool. In addition, the only block structure specified for the pool was for the item sets. All other items were considered as not belonging to any block.

The Approach to Item Security

The randomization approach described previously was judged unsatisfactory for the quantitative adaptive test because of a greater requirement for item security. Nothing in the previous approach directly inhibits the administration of items to a large proportion of examinees. To more directly control the exposure rate for each item, the approach suggested by Sympson & Hetter (1985) was employed.

Sympson & Hetter (1985) outlined a method of developing exposure control parameters for each item in an adaptive test item pool in such a way that a specified expected maximum exposure rate is not exceeded. This method was extended by Stocking (1993) to apply to pools containing items as well as stimuli. The exposure control parameters specify the proportion of time an item or stimulus is administered given that it is selected to be optimum. The parameters are developed over a series of iterations, each iteration consisting of an adaptive test simulation to a large number (over 1,000 is recommended) of simulated examinees randomly drawn from a trait level distribution typical of the real examinee population. The exposure control parameters are adjusted by the results of each iteration so that new exposure control parameters are used for the next iteration. The parameters

eventually converge so that the observed (not expected) maximum exposure rate is slightly above the specified expected maximum exposure rate.

The Simulations

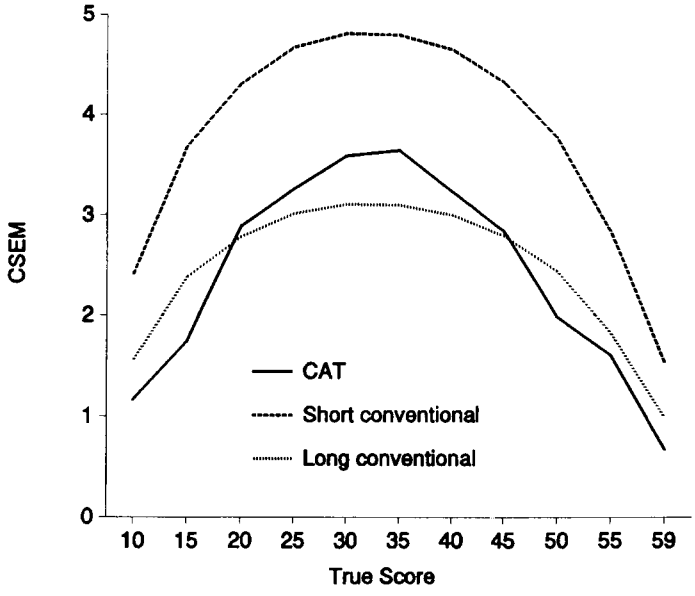
The simulations were performed for 100 simulees at each of 11 values on the reported score metric ranging from just above chance level to just below a perfect score. 16 Sympson-Hetter iterations were performed, with an expected maximum exposure rate specified at .20. To obtain the unconditional evaluations necessary for the Sympson-Hetter development of exposure control parameters for each item, an estimated distribution of true θ was computed using the method of Mislevy (1984) and the item parameters and responses from a group of over 5,000 real examinees who took an edition of the 60-item parent form.

Results

The estimated reliability of the 25-item adaptive test was .92. This reliability was achieved using 256 items (out of 496) and 19 item sets (out of 22) at least once across 1,100 adaptive tests, and with the maximum observed exposure rate for any item of .23. The reliability of the paper-and-pencil test used for scoring purposes was .93. Figure 2 displays the CSEM functions for the adaptive test (*CAT*), the reference test (*long conventional*), and the reference test shortened to the length of the adaptive test (*short conventional*).

The adaptive test CSEM function was more peaked than expected, an aspect that is a consequence of the Sympson-Hetter exposure control methodology. The Sympson-Hetter method controls item exposure with respect to a typical distribution of examinee trait level. Thus, optimum items for the most numerous examinees are given low exposure control parameters, forcing the administration of

Figure 2
CSEM for the 60-Item Conventional Quantitative Test, the 25-Item Conventional Quantitative Test, and the 25-Item Adaptive Quantitative Test



less optimum items. This happens less at the extremes of the θ distribution because there are fewer examinees.

The adaptive test achieved its measurement and item exposure properties without violating overlap or set constraints, although some content constraints were violated, as also shown in Table 5. Only in constraint 16 did the number of items administered averaged over the typical distribution violate the constraint, but only by a small amount. In contrast to the verbal adaptive test, an examination of the conditional average number of items failed to show a relationship between items with a particular feature and the appropriateness of the items for a particular trait level.

Discussion

Previous methodologies for adaptive testing have not been able to take into account the number and complexity of constraints on item selection found in expert test development practice. The examples differed from each other in terms of the content (verbal, quantitative), number of constraints (large, small), the amount of overlap (large, small), block structure of the pool (more complex, less complex), and the methodology employed to deal with item security (simple randomization, the Sympson and Hetter method). The two contrasting examples showed that the proposed methodology can be successfully applied under a wide variety of circumstances.

The success of this new method rests on the fact that it can incorporate content, overlap, and set constraints on the sequential selection of items as desired properties of the resultant adaptive tests. The extent to which restrictions on item selection are not satisfied is then the result of deficiencies in the item pool and the structure of the constraints. With this method, adaptive test construction can incorporate the same standards already established for conventional tests.

References

Ackerman, T. (1989, March). *An alternative methodology for creating parallel test forms using the IRT information function.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Brooke, A., Kendrick, D., & Meeraus, A. (1988). *GAMS: A user's guide.* Redwood City CA: The Scientific Press.

College Board. (1990). *Coordinator's notebook for the computerized placement tests.* Princeton NJ: Educational Testing Service.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

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

Kingsbury, G. G., & Zara, A. R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241–261.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monograph,* No. 7.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139–183) New York: Harper and Row.

Lord, F. M. (1971a). Robbins-Munro procedures for tailored testing. *Educational and Psychological Measurement, 31,* 3–31.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement, 8,* 147–151.

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

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

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

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

Segall, D. O. (1987). *ACAP item pools: Analysis and recommendations.* San Diego CA: Navy Personnel Research and Development Center.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report 93-2). Princeton NJ: Educational Testing Service.

Stocking, M. L., Swanson, L., & Pearlman, M. (1991). *Automatic item selection (AIS) methods in the ETS*

testing environment (Research Memorandum 91-5). Princeton NJ: Educational Testing Service.

Stocking, M. L., Swanson, L., & Pearlman, M. (1993). The application of an automated item selection method to real data. *Applied Psychological Measurement, 17,* 167–176.

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

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego CA: Navy Personnel Research and Development Center.

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

Theunissen, T. J. J. M. (1986). Some applications of optimization algorithms in test design and adaptive testing. *Applied Psychological Measurement, 10,* 381–389.

Van der Linden, W. J. (1987). Automated test construction using minimax programming. In W. J. Van der Linden (Ed.), *IRT-based test construction* (pp. 1–16). Enschede, The Netherlands: Department of Education, University of Twente.

Van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237–248.

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: Erlbaum.

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

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

Weiss, D. J. (1976). Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (pp. 24–35). Washington DC: United States Civil Service Commission.

Weiss, D. J. (Ed.) (1978). *Proceedings of the 1977 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver BC: Educational Research Institute of British Columbia.

Zara, A. R. (1990). A research proposal for field testing CAT for nursing licensure examinations. In *Delegate Assembly Book of Reports 1989.* Chicago: National Council of State Boards of Nursing.

Zara, A. R., Bosma, J., & Kaplan, R. (1987). *Functional and design specifications for the National Council of State Boards of Nursing adaptive testing system.* Unpublished manuscript.

Author's Address

Send requests for reprints or further information to Martha L. Stocking, Educational Testing Service, Princeton NJ 08541, U.S.A. Internet: mstocking@rosedale.org.