DOCUMENT RESUME

ED 385 551 TM 023 967

AUTHOR	Stocking, Martha L.
TITLE	Three Practical Issues for Modern Adaptive Testing Item Pools.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-94-5
PUB DATE	Feb 94
NOTE	45p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Item Banks; Standards; Test Construction; Test Format; *Testing Problems; *Test Items; Test Length
IDENTIFIERS	Large Scale Assessment; *Large Scale Programs; Parallel Test Forms

ABSTRACT

As adaptive testing moves toward operational implementation in large scale testing programs, where it is important that adaptive tests be as parallel as possible to existing linear tests, a number of practical issues arise. This paper concerns three such issues. First, optimum item pool size is difficult to determine in advance of pool construction. Retrospective results are analyzed for five operational pools; these analyses indicate that item pools of the size and quality of six to eight linear tests are adequate to support adaptive tests of roughly half the length of a parallel linear test. Second, item pools may not support sufficiently low exposure rates for items or a sufficiently small amount of test overlap to maintain test security when testing is conducted on a continuous basis. Various simulations suggest that multiple pools which can be chosen randomly before adaptive testing begins provide a satisfactory solution. Finally, over time it will be necessary to refresh or replace operational item pools. Issues that must be considered in refreshing or replacing item pools are discussed and guidance is given for establishing benchmark values from initial pools as standards that must be met for adaptive test from new pools to be considered parallel to adaptive tests from existing pools. (Contains 5 tables and 18 references.) (Author)

RR-94-5

RESEARCH REPORT

THREE PRACTICAL ISSUES FOR MODERN ADAPTIVE TESTING ITEM POOLS

Martha L. Stocking

ETS®
Educational Testing Service
Princeton, New Jersey
February 1994

THREE PRACTICAL ISSUES FOR MODERN ADAPTIVE TESTING ITEM POOLS[1]

Martha L. Stocking
Educational Testing Service
Princeton, New Jersey 08541

[1] This research was supported in part by the Program Research Planning Council of Educational Testing Service.

Abstract

As adaptive testing moves towards operational implementation in large scale testing programs, where it is important that adaptive tests be as parallel as possible to existing linear tests, a number of practical issues arise. This paper concerns three such issues. First, optimum item pool size is difficult to determine in advance of pool construction. Retrospective results are analyzed for five operational pools; these analyses indicate that item pools of the size and quality of six to eight linear tests are adequate to support adaptive tests of roughly half the length of a parallel linear test. Second, item pools may not support sufficiently low exposure rates for items or a sufficiently small amount of test overlap to maintain test security when testing is conducted on a continuous basis. Various simulations suggest that multiple pools which can be chosen randomly before adaptive testing begins provide a satisfactory solution. Finally, over time it will be necessary to refresh or replace operational item pools. Issues that must be considered in refreshing or replacing item pools are discussed, and guidance is given for establishing benchmark values from initial pools as standards that must be met for adaptive tests from new pools to be considered parallel to adaptive tests from existing pools.

Introduction

Recent advances in psychometrics and computing technology have led to the development of a testing paradigm that is very different from linear paper-and-pencil testing -- computerized adaptive testing (CAT), see, for example, Eignor, Way, Stocking, and Steffen (1993), Lord (1977), Schaeffer, Steffen, and Golub-Smith (1993), Stocking and Swanson (1993), and Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, and Thissen (1990). As modern adaptive testing moves towards large-scale operational implementation, particularly in high-stakes testing programs, practical issues arise that have no counterparts in linear paper-and-pencil testing but nevertheless require satisfactory solutions to maintain operational implementation.

In this paper, three related practical issues relevant to the introduction of adaptive testing are discussed. The first issue relates to the required size of an item pool. Older CAT paradigms, such as that used by Lord (1977) and also by Ward (1988), allow some educated guesses in advance about optimum pool sizes based on the nature of the item selection algorithm and the need (or lack thereof) for secure testing. Modern CAT paradigms, such as the paradigm developed by Stocking and Swanson (1993), do not. Yet the answer to the question of pool size is vital for the planning and execution of item writing, pretesting, and pool development efforts. Optimum pools sizes, based on a particular adaptive testing paradigm, are investigated for five different adaptive tests in the first section of this paper.

The second issue relates to the number of pools that should be operational at any given point in time. Although sophisticated exposure control mechanisms, such as the Sympson and Hetter (1985) procedure, are commonly used in modern adaptive testing, such methods may not be adequate to control item exposure in the face of large volumes of test-takers who may test on practically a daily basis. In addition, controlling item exposure only indirectly controls a second aspect of security -- the overlap among adaptive tests drawn from a single pool. A comparison of three different methods of decreasing item exposure and reducing test overlap, including enlarging item pools and increasing the number of pools, is described in the second section of this paper.

The third issue involves consideration of item pool refreshment efforts. Potential approaches range from replacing part of a pool or possibly an entire pool with new items to reconstituting existing pools by redistributing some or all items across some or all existing pools. The issue of the timing of pool refreshment is one of testing program policy, presumably based on concerns for security and the currency of items, and is not discussed in this paper. What is at issue here is a discussion of practical limitations, various approaches, and test design activities that are necessary to establish new adaptive tests that are parallel to previous adaptive tests. These are discussed in the final section of the paper.

Modern Adaptive Testing

The psychometrics underlying the tests studied in this paper are based on the three parameter logistic Item Response Theory (IRT) model (Lord, 1980). Items in the item pools were calibrated and placed on the same metric using the computer program LOGIST (Wingersky, 1983). The item selection process in the adaptive test employs the methodology of the weighted deviations model (WDM) (Stocking and Swanson, 1993, Swanson and Stocking, 1993) with the extended Sympson and Hetter (1985) exposure control methodology (Stocking,

1992) to increase item security. (For details of the test design process, see Eignor, et al., 1993). The goal of the test design process is to have (fixed length) adaptive test scores that are interchangeable with those from companion linear paper-and-pencil tests both in terms of their psychometric properties and the constructs being measured. This is necessary since it is envisioned that both modes of testing must co-exist for some indefinite period of time into the future.

In the WDM approach to adaptive testing, item properties or features are taken into account along with statistical properties in the selection of items. This is to insure that each adaptive test produced from the pool matches a set of test specifications and is therefore as parallel as possible to any other test in terms of content and types of items, while being tailored to an individual examinee in terms of difficulty. The WDM approach also allows for the specification of overlapping items that may not be administered in the same adaptive test. In addition, it is possible to restrict item selection to blocks of items, either because they are associated with a common stimulus or common directions or because they are associated with other item features that test specialists deem important.

In summary, in the weighted deviations model, the next item selected for administration is the item that simultaneously

1) is as informative as possible at a test-taker's estimated ability level, while

2) contributing as much as possible to the satisfaction of all other constraints in addition to the constraints on item information.

At the same time, it is required that the item

3) does not appear in an overlap group containing an item already administered, and

4) is in the current block (if the previous item was in a block), starts a new block, or is in no block.

The Sympson and Hetter exposure control methodology further restricts item selection by determining if the selected item is likely to be overexposed if administered, based on exposure control parameters developed over a series of simulations with a (simulated) typical group of test-takers. If so, this methodology forces the administration of an item that has been administered less frequently. The adaptive test is scored by converting the final (maximum likelihood) estimate of examinee proficiency to an estimated number right true score on a (linear) reference test that was previously scaled to the score reporting metric.

The operational administration of adaptive testing is envisioned as having certain characteristics which also make it very different from linear paper-and-pencil testing. Many relatively small test centers are set up in a variety of convenient locations. Test-takers may reserve a convenient testing time throughout the day. Retesting frequently is likely to become the norm, although this may be restricted somewhat by testing program policy until item and test security concerns can be accommodated. Initial score reporting may be done on site, with follow-up 'official' score reporting done within days of testing.

Issue 1: Pool Size

Previous Methods of Determining Pool Size

Factors in the design of an adaptive test can be expected to influence the number of items required in the item pool. These include the item selection algorithm, constraints on item content, psychometrics, and exposure, stopping rules, overlap restrictions, test scoring, requirements of parallelism with existing paper-and-pencil forms, and so forth. For adaptive testing paradigms developed in simpler contexts it is usually possible to estimate in advance how many items are required to satisfy the demands of the test design.

For example, Lord (1977) in his Broad Range Tailored Test of Verbal Ability was able to use a relatively simple approach to restrictions on test content. The same approach is used in the College Placement Tests (Ward, 1988), and also, with minor modifications, for the NCLEX/CAT (Kingsbury and Zara, 1991). Lord dealt with five mutually exclusive item types, and determined in advance the order of item types for each tailored test drawn from the pool through the mechanism of a table. Once the type of the next item to be administered was determined, the most informative item available of that type was selected. An example of a portion of such a table is given in Table 1.

Insert Table 1 about here

Each row of Table 1 represents the serial position of an item in an adaptive test. Each column represents a particular level of proficiency. The type of the next item selected for a test-taker is found in the row representing the serial position of the item and the column representing a level of proficiency closest to the current estimate of test-taker proficiency. Within the body of the table are letters representing particular item types. If the item pool were completely adequate, all rows of the table would contain the same item type, indicating that all test-takers receive items of the same type in the same order, regardless of proficiency.

The situation illustrated shows that the item pool is not quite adequate. For example, it would be desirable to have more type b items at low and high levels of proficiency, more type e items at low levels of proficiency, and more type d items at middle levels of proficiency. However test-takers with similar levels of proficiency receive similar adaptive tests in terms of content.

If a simple randomization scheme was imposed on this adaptive test design then the number of items in each cell of Table 1 needs to be increased accordingly. For example, if the first item is randomly chosen from a group of five equivalent items, the second from a group of four, and so forth, then each cell in the first row must contain five items and each cell in the second row must contain four items, and so forth. Adding across all cells gives the number of items required for the total pool.

A New Method of Determining Pool Size

This approach of estimating the number of required items in advance is not possible in the more complex context found in modern adaptive testing, for a number of reasons. First, items are classified by content and type according to much more elaborate schemes which do not necessarily result in mutually exclusive groups, as in Table 1. Second, the specification of extensive overlap in which items may overlap with other items for a variety of reasons adds an additional level of complexity. Third, the use of the extended Sympson and Hetter exposure control methodology (Stocking, 1992) adds a population dependent element to item selection that is more complex than the population independent simple randomization scheme outlined above.

What is possible, however, with the adaptive tests studied here which are designed to be as parallel as possible to a preexisting linear test, is to determine the properties of the item pools after completion of the test design simulations and compare these properties to the reference test used for scoring purposes. Since this reference test is typical of the linear test forms administered in paper-and-pencil mode, its properties are well known and such comparisons provide the necessary guidance for future pool-building efforts in familiar terms.

The Five Adaptive Tests

Information about the five adaptive tests at the end of the test design simulations is shown in Table 2. These simulations were conducted using appropriate estimated distributions of true ability for typical populations, developed using the methods of Mislevy (1984). There are two measures of mathematical reasoning, two measures of verbal reasoning, and one measure of analytical reasoning. The first column in Table 2 lists the number of elements in the pool including 'discrete' items, stimuli such as reading passages, and items associated with stimuli, all of which can be subjected to constraints on their selection in the weighted deviations model. The numbers include every element that was used at least once in these simulations.

Insert Table 2 about here

The next two columns give the number of items and the number of stimuli for each pool. In general, the two math pools have the smallest number of sets, followed by the two verbal pools. The analytical reasoning pool has the largest number of sets because the adaptive analytical test has the majority of its items associated with stimuli.

The fourth and fifth columns in Table 2 give the (fixed) adaptive test lengths and reference test lengths for each measure. The adaptive test length is determined by the simulations to be the minimum number of items necessary to provide sufficient constraint and overlap satisfaction, satisfactory item exposure levels, and acceptable conditional standard error of measurement curves and reliability levels for a typical group of test-takers; the latter is computed using the method of Green, Bock, Humphreys, Linn, and Reckase (1984, equation 6). The adaptive tests range from roughly 1/3 to 2/3 the length of the reference test. The analytical measure, because of the restriction that the majority of its items come in sets, is the longest adaptive test both in terms of absolute numbers and in proportion to the length of its companion reference test.

Test development specialists specified constraints on item selection, as shown in the sixth column of Table 2. These constraints insure that each adaptive test built from a pool matches the same test specifications as closely as possible. In addition, the five tests represent two different approaches to item security, as shown in the seventh column. For two tests, a simple randomization scheme was employed with decreasing group size. For Math 2 the randomization groups were {8, 7, 6, 5, 4, 3, 2, 2, 2, 2, 1, . . .}, that is, the selection of the first ten items involved a random process with the first item selected randomly from the eight best candidate items, the second item selected randomly from the seven best candidate items, and so forth. From the eleventh item to the end of the test, items were selected optimally. For Verbal 2, the group sizes were {8, 7, 6, 5, 4, 3, 2, 1, . . .}. For the remaining three tests, the extended Sympson and Hetter exposure control methodology was employed, with the target maximum exposure rate specified to be .2, that is, ideally no element in the pool is to be seen by more than 20% of a typical population.

The final column in Table 2 indicates whether test developers specified that certain items could not be administered in the same adaptive test as other items. For all but one measure, overlap groups were specified, and the degree of overlap was moderately extensive. Across the four tests, the Math 1 adaptive test had the minimum degree of overlap with 57 overlap groups that involved a total of 198 items. For the remaining measure, Math 2, test developers chose to control overlap through the setting of content constraints since their coding of items was sufficiently detailed to support this approach, hence the substantially larger number of constraints for this test.

Pool Size in Terms of Content Considerations

The first four columns of Table 3 display information obtained from analyzing the structure of the pools in terms of the structure of the adaptive tests. The global factor listed in the first column is the ratio of the number of items (not including stimuli) in each pool to the adaptive test length. This ratio is remarkably consistent across the variety represented by the five tests. If one considers the average adaptive test length to be roughly one-half that of the reference test, then these ratios, in terms of the length of the reference test, become equal to roughly six. (That is, the ratio (number of items in pool)/(number of items in CAT) of 12 implies the ratio (number of items in pool)/(number of items in reference test) is 6, since (number of items in CAT) is roughly 1/2*(number of items in reference test).) This leads to the interpretation that the adaptive test pool should contain roughly six times as many items as a reference test if the desired adaptive test is roughly half the length of the reference test.

Insert Table 3 about here

The factors listed in columns two through four of Table 3 attempt to develop parallel information for the more detailed structure of item pools, but clearly with less consistency across tests. The stimulus factor is the ratio of the number of stimuli in the pool to the number of stimuli to be administered to a typical test-taker in an adaptive test, according to the adaptive test specifications. This ratio is represented by single numbers for the two math measures because, in this context, stimuli are not categorized by content. For the two verbal measures and the analytical measure, the ratio is represented by a range because stimuli are typically categorized with regard to length, subject matter, style, and perhaps subgroup reference material. However, a factor of twelve -- that is, the pool should contain twelve times as many sets of any type than are to be administered to a test-taker -- as a rough rule of thumb would be a conservative estimate of this ratio across all types of measures. Using the same logic as before, if one considers the average adaptive test length to be roughly one-half that of the reference test, then this ratio, in terms of the length of the reference test, becomes roughly six.

Most sets of items associated with stimuli in an item pool contain more items than are to be administered from a set to a test-taker in the adaptive test. The third column represents the ratio of the average number of items per set in the pool to the number to be administered in the adaptive test. Factors of 1.0 imply that there should be exactly as many items associated with a set as there will be administered in the adaptive test. However, it is clear that this may impair measurement; once the item selection algorithm has chosen a single item from a set, it must administer the remaining items before it can consider items outside the set, regardless of how appropriate those items are for a particular test-taker. The effect of this is probably not too grave if the number of items to be administered from a set is small, say, two. However, if the number of items is larger, say three or four, measurement properties of the adaptive test may be compromised unless all items in a set are most appropriate at about the same proficiency level. Since it may be difficult, perhaps even impossible, to insure that all items in a set are most appropriate at about the same proficiency level, a conservative approach would be to always have three to four times as many items associated with a set than are to be administered to any test-taker.

The discrete factor in the fourth column of Table 3 is the ratio of the number of discrete items in the pool with a particular attribute to the number of items with the same attribute to be administered in the adaptive test. These factors were developed considering only the important (that is, most heavily weighted) constraints on discrete item selection. It would be reasonably conservative to always have roughly twelve times as many discrete items with important attributes in the pool than are to be administered to an individual test-taker. As before, if the adaptive test is roughly one-half the length of the reference test, this ratio, in terms of the length of the reference test, becomes roughly six.

Pool Size in Terms of Statistical Considerations

The above analysis suggests on a global level that if an item pool contains roughly six times the number of items in a reference test then it will support an adaptive tes: that is roughly half the length of the reference test, at least in terms of the content requirements of the adaptive test. So far, however, this conclusion is simplistic at best because we have not taken into account the statistical properties of either the pool, the adaptive test, or the reference test. In this section we incorporate these measurement concerns.

To examine measurement issues, we need to compute typical values of the ratio of the test information function (Lord, 1980, equation 5-6) for the reference test to the test information function for the adaptive test, and also the ratio of the pool information function (computed as if the pool were to be administered as a single linear test) to the information function for the adaptive test. The adaptive tests studied here are fixed length tests, and therefore measure differentially well across the proficiency continuum, although always above some minimal level. (Variable length adaptive tests purport to measure equally well across the proficiency continuum, although they may produce unacceptably large biases in test scores (Stocking, 1987).) Because the tests measure differentially well, we need to consider measurement properties conditional on proficiency level, and then weight these results by the estimated typical distribution of true proficiency to get the results for typical test-takers.

Over the range of proficiency within which 70% to 77% percent of typical test-takers lie (depending upon the test), equally spaced true scores were identified. Two random simulated examinees (simulees) were drawn from each true score level, and the information functions of their adaptive tests in the final simulations were computed and averaged. This treats the adaptive test as if it were a fixed linear test for each examinee, which, while not an appropriate surrogate for the information function of an adaptive test (because it does not take into account the nature of adaptive item selection), does not impair the intended comparisons.

The information function for the reference test was computed at these same true scores, and the ratio of these to the (average) information functions for two random simulees was taken. An average ratio was computed using as weights the typical distribution of proficiency. For all tests, the weighted ratio of the information function for the reference test to the information function for the adaptive test was close to 1.0. This is not surprising -- the adaptive tests were carefully designed to be as parallel as possible to the companion reference tests since the linear paper-and-pencil tests and the adaptive tests are expected to co-exist for some time.

To get the ratio of the pool test information to the adaptiv test information, the pool information function was computed at the same true scores, and the ratios computed and weighted as before by the estimated distribution of proficiency to produce a typical value. This weighted average ratio is presented as the statistical global factor in column five of Table 3. Since the adaptive test and the reference test can be viewed as interchangeable from a measurement perspective, this global factor can be interpreted to mean that the pool must contain about six to seven times as much information as the reference test in order to support an adaptive test that is roughly one-half the reference test length. This conclusion parallels the conclusion reached for content considerations in the previous section.

As a check on this procedure, the average statistical global factor of 6.6 was multiplied by the reference test length to get a predicted pool size for each of the five tests. The ratio of the actual pool size in terms of items only (not stimuli) to the predicted pool size is given in the final column of Table 3. For adaptive tests with lengths less than or equal to half the reference test length, our methodology has been conservative -- actual pool sizes are smaller than we would predict. For the one measure for which the adaptive test length is greater than half of the reference test length the actual pool size exceeded our prediction by about 20%. For tests of this nature, in which item selection is substantially restricted because of the preponderance of set based items, it would be better to aim for an item pool that is the equivalent of seven or eight linear test forms.

Conclusions

The context for this study is one in which adaptive tests use the weighted deviations model for item selection, are designed to be as parallel as possible to existing linear forms, and use a typical linear form as a reference test for score reporting. In this context, both in terms of content and in terms of measurement properties, the data suggest that a pool composed of six to eight typical linear forms will support adaptive testing where the (fixed) length of the adaptive test is roughly one-half that of the linear forms. This rule of thumb seems to hold across two different randomization methods, three different types of measures, and two different approaches to the specification of overlap.

Issue 2: The Number of Pools[1]

Item Exposure and Test Overlap

Ever since adaptive testing has been viewed as being operationally feasible, concern has been focussed on controlling the exposure of items even at the expense of the measurement efficiency of the adaptive test. The earliest schemes were similar to the randomization approach outlined above in which items are selected randomly from groups of items that are approximately equivalent to each other. However, it is difficult, if not impossible, to determine the optimum group size and selection sequence with this approach except by time consuming trial and error.

The Sympson and Hetter approach as extended by Stocking (1992) represents a model in which exposure control parameters are developed for individual elements in the pool through a sequence of adaptive test simulations with simulees drawn from a typical distribution of proficiency. The exposure control parameters for items that might be frequently chosen for administration based on their content and/or statistical properties tend to be lowered, implying that these items will only be administered some fraction of the times they are selected as being optimal. The exposure control parameters for less popular items tend to be raised, in comparison, so that less desirable items are administered a higher proportion of the times that they are chosen as being optimal.

[1] Len Swanson suggested many of the ideas contained in this section, as well as providing the computer programming necessary to accomplish the analyses.

For the three measures discussed in the previous section that employed the extended Sympson and Hetter methodology, the starting item pools were constituted from all the available items at hand and these pools were larger than the final pool sizes listed in Table 2. The extended Sympson and Hetter iterations were conducted on these larger pools with the final target expected maximum exposure rate specified as .20, that is, no more than 20% of a typical population of test-takers would see even the most popular item. Lower target extpected maximum exposure rates were attempted, but with unsuccessful results, that is, the sequence of simulations failed to converge as required by the procedure. The actual observed maximum exposure rates may be uncomfortably high, however, for a testing program expected to have a large yearly volume.

In addition, neither approach to item security considers directly a second important security issue -- the amount of overlap between adaptive tests constructed for different test-takers. Overlap is not unrelated to item exposure. In principle, the lower the exposure rate, the lower the amount of test overlap. If the amount of overlap is small, then information that is subsequently shared by individuals who have taken adaptive tests is not very useful. If the amount of overlap is large, then some test-takers may be advantaged by receiving information from other test-takers who have taken adaptive tests before them. This is, of course, the reason why many linear testing programs with a few fixed administrations throughout a testing year may choose never to repeat a test form. In this section we describe a series of investigations into improving item security and test overlap.

Three Approaches

We considered three different approaches to improving item security and decreasing the amount of test overlap, as follows:

1) Pool doubling

If the extended Sympson and Hetter iterations fail to converge for specified maximum exposure rates that are comfortably low, then the pool must be constituted in such a fashion as to not support the lower rates. Perhaps doubling the size of the pool by adding elements exactly like those already in the pool will halve the exposure rates. A Monte Carlo experiment was conducted to investigate this proposed solution.

2) The "item by item" approach

Consider an item pool for which one has developed exposure control parameters. Suppose one could construct, for each element in the pool, an exact twin element in terms of all of the determinants of items selection such as content, statistical properties, overlap, and so forth. If this were possible, then when an item is chosen for administration one could administer randomly either the item or its twin, thus halving the exposure rate of either item. One could also apply the same principle to triplets, quadruplets, and so forth. Item twinning is theoretically identical to doubling the pool, but will be shown below to have different properties than the pool doubling approach.

3) The "whole pool" approach

Suppose one had two nonoverlapping pools, and before testing, randomly selected the pool from which to test. This is similar to the concept of twinning individual items and could be extended to three or more pools, but part of the random selection takes place at the pool level, not the item level. Regardless of what the maximum observed exposure rate was for a single pool, that exposure rate is automatically halved if there are two pools, automatically reduced to one-third if there are three pools, and so forth. This approach appears to hold promise, both in terms of item security and the reduction of test overlap, and was studied further in the simulations descri^hed below.

The Three Tests

Item security and test overlap issues were addressed using three different adaptive tests. These tests are parallel to, but not the same as, some of those presented in the last section. Table 4 shows the results after the end of the extended Sympson and Hetter test design simulations. The desired maximum expected exposure rate for all three tests was specified as .2. As anticipated, the maximum observed (not expected) exposure rates were typically slightlv higher than this desired expectation. Information about the resultant test overlap is presented as part of Table 5 below.

Insert Table 3, 4, and 5 about here

Pool Doubling

This simple approach was investigated first. The Math 3 pool above was doubled by the duplication of every element in the pool. The current overlap specifications detail overlap only within halves of this doubled pool. Therefore overlap across halves was artificially constructed by randomly selecting items from the first half and randomly pairing them with items from the second half to form additional overlap groups of the same size as the median overlap group size for a single half-pool. The number of overlap groups across the two half-pools was about the same as the number of overlap groups within each half-pool.

Extended Sympson and Hetter iterations were performed on this doubled pool with the maximum desirable exposure rate set at .1, half of the previous value. After the eleventh iteration, the procedure began to diverge and produce increasingly large exposure rates, especially for stimuli and items associated with them. At first it was thought that this divergence might be due to some flaw in the creation of the artificial overlap. The overlap between the two halves of the pool was removed and the iterations repeated, again with a maximum desirable exposure rate of .1. These iterations also failed to converge, again starting in the eleventh iteration. An explanation for this phenomenon is given below in connection with the resul s for twinning items.

Item and Pool Twinning

The number of simulees (from a rectangular ability distribution) in the simulations to establish the test designs for the three tests was 1650 for Math 3, 1300 for Verbal 3, and 1170 for Analytical 2. Using the estimated distribution of true ability in a typical population (a different distribution for each measure) these simulees were sampled to represent typical distributions of examinees, producing 879 simulees for Math 3, 511 for Verbal 3, and 518 for Analytical 2.

Separately for each measure, all possible pairs of tests were then formed and compared. That is, the test for the first simulee was compared with the test fo the second, third, . . ., nth simulee. The test for the second simulee was compared with the test for the third, fourth, . . ., nth simulee, and so forth. This resulted in 385,881 unique pairs of tests for the

Math 3 measure, 130,305 for the Verbal 3 measure and 133,903 f(r the Analytical 2 m<asure. The comparisons were made without regard to the order in which items were actually administered.

The comparisons between pairs of tests were made in two different ways in order to simulate the twinning (or higher order replication) of items and the twinning (or higher order replication) of pools. In the first method, each item in each of two tests being compared was randomly assigned to two different pools and also to three different pools. This "item by item" method gives the results for test overlap if we consider twinning or tripling each item in a pool and retaining enlarged pools as opposed to separate pools.

In the second method, each test was randomly assigned as originating from one of two equivalent pools, giving results for test overlap as if one had two separate nonoverlapping pools of items, while halving the exposure rates. This was also done to simulate three different pools and four different pools. This whole test method would be implemented by randomly selecting the pool before the test begins.

Results

Table 5 presents the results for the two methods tried (and also the original pool) for all three measures and for two nonoverlapping pools. Within the body of the table are the cumulative percentages over the entire distribution of relevant pairs. Each row represents a different amount of overlap. The columns labeled "one pool" represent the results for the current CAT pools presented in Table 4. With two pools, regardless of whether or .ot they are obtained by twinning individual items or twinning the pool, the observed maximum exposure rates shown in Table 4 are reduced by half.

With the original pool, 39% of a typical distribution of examinees taking the Math 3 measure would receive tests that overlap 10% or less with other examinees; 57% would receive tests that overlap 15% or less with other examinees, and 16% would receive tests that overlap more than 30% with other examinees. If there were two separate Math 3 pools, 70% would find that the overlap was 10% or less (vs 39%); if we had twinned items, 63% would find that the overlap was 10% or less. Thus the item by item method is not as good as two separate pools at the 10% or less overlap level.

Not shown in Table 5 are indications of overlap at the median of a typical population for these measures. For the single Math 3 pool, 50% of the people receive tests that overlap 14% or less. For the two pools approach, 50% of the people receive tests that overlap 1% or less, while for the item by item method, 50% of the people receive tests that overlap 7% or less. The percentages overlap at the 50th percentile for the Verbal 3 measure are identical. The median percentages are slightly different for the Analytical 2 measure, where with the current pool 50% receive tests that overlap 11% or less, for the whole pool approach 50% receive tests that overlap 1% or less, while for the item by item approach, 50% receive tests that overlap 6% or less.

The whole test approach has more people at lower levels of overlap (than the item by item approach) because many examinees have no overlap at all. Therefore the overlap percentages for the test-takers at the median are lower. The item by item approach has fewer examinees with large (>30%) levels of overlap than the whole test approach but the median percentages of overlap are higher because they have fewer tests with no overlap.

Also not shown in the Table 5 are the results for larger numbers of pools. For the percentage of people with 10% overlap or less only, for the whole test approach 80% and 85% of the examinees taking the Math 3 measure have 10% or less overlap for three and four pools respectively. For the whole test approach and the Verbal 3 measure, 82% and 86% have overlap of 10% or less for three and four pools respectively. For the Analytical 2 measure, the figures are 83% and 87% with 10% or less overlap with three and four pools respectively.

For the item by item approach, for three pools (four pools was not simulated) 77% of the Math 3 test-takers, 87% of the Verbal 3 test-takers, and 86% of the Analytical 2 test-takers had overlap of 10% or less.

Pool Doubling Revisited

Given a pool with a certain observed maximum exposure rate, one can guarantee to halve all exposure rates with two identical pools constructed using either of the two methods described above. It was originally thought that doubling an existing pool would produce the same results. After further thought and review of the simulations described above, the explanation for why this does not occur is now clear.

When the pool is simply doubled (as opposed to twinning each item), the selection of an item for administration is not guaranteed to be equivalent to selecting randomly between the two equivalent members of a pair of items on each and every item selection, as it is for item twinning. Rather, because of the randomness introduced into the item selection by the extended Sympson and Hetter procedure, one may be randomly choosing among unpaired items. When this happens, one is guaranteed not to do as well as when one has two identical pools and using either the item by item approach or the whole pool approach, in terms of exposure rates of items. For the Math 3 pool on which these simulations were conducted, this randomness produces an increasingly chaotic situation as the iterations to stabilize exposure rates progress. It is more manifest for those pool entities, like sets, that are in short supply. Thus, one is lead to the conclusion that a doubled pool will not support halved exposure rates, but rather something more than half.

To test the hypothesis in the previous paragraph for the Math 3 measure, the double-pool iterations were run again, this time with a maximum desirable exposure rate of .15, half way between the two previous attempts. These iterations performed completely satisfactorily in terms of the convergence of the individual exposure rates. The maximum observed exposure rates were .19 for discrete items, .15 for stimuli, and .15 for items associated with stimuli, as opposed to .22, .18, and .18 for the single (real) item pool.

Conclusions

The whole pool approach performs better in terms of test overlap than the item by item approach for both the Math 3 measure and the Analytical 2 measure; both approaches, however, produce the same reduction in item exposure rates. The whole pool approach performs slightly less well than the item by item approach for the Verbal 3 measure. It is clear that the improvements of two pools over a single pool are large for both methods.

Based on these results for a variety of measures, the whole pool approach, that is, the production of two or more equivalent pools and the random assignment of pools to test-takers before testing begins, seems to be the recommended mechanism of choice to minimize overlap among forms administered and to reduce the exposure rate of elements in any single pool.

Issue 3: The Refreshment of Pools

Concerns about individual item exposure rates and whole pool exposure lead to reflection about how best to replace or refresh pools before test security is breached. Such concerns are driven by volume. If an item is administered to no more than 10% of a test-taker population, but there are 1,000,000 test-takers in a year, then perhaps administering an item to 100,000 test-takers will compromise its security. Likewise, if one has three pools from which to randomly select before starting an adaptive test, it might be argued that the exposure of a pool to 300,000 test-takers, where no item is exposed to more than 30,000 test-takers may still compromise pool and item security. Decisions about the timing of pool refreshment or replacement must be the result of test sponsor policy and will not be discussed further in this paper. What we focus on instead are practical issues to consider when contemplating pool replacement or refreshment, some approaches that might be taken, and the test design activities required to insure the continuing parallelism of adaptive tests across multiple pools and across time.

This practical issue differs from the two previously discussed in two important respects. First, at this point in time, typically only initial item pools exist for the measures at hand, although this situation is expected to change rapidly. Therefore simulation studies to explore certain concepts in this section are not currently possible but will be shortly and such studies could and should be conducted when data become available. Second, the predominant focus in the available literature is on the initiation of CAT. Since pool refreshment is a maintenance rather than initiation activity, there are few available sources of guidance. Therefore the content of this section should be viewed as informed advice that needs to be confirmed in future studies.

Practical Issues

A limiting factor in pool refreshment or replacement is the production and pretesting of new items. It is possible to develop and pretest items for adaptive pools as part of a paper-and-pencil testing program, as long as it can be shown that individual item parameter estimates do not change when the mode of administration changes. To avoid this issue, items can be pretested as part of an adaptive testing program, either by scattering the items throughout an adaptive test or placing them in a separate test section (Stocking, 1988a).

We have discovered that an adequate adaptive testing pool can be constructed from the equivalent of six to eight linear test forms. The initial number of items necessary to produce six to eight linear test forms is likely to be more than double the number of items that survive the extensive reviews for final-form quality items. Thus the item production effort, even for a single adaptive testing pool, may be high. To reduce test overlap, it is recommended that there be at least two pools, which further increases the item production effort. To maintain the equivalent of at least two pools over time will require careful planning to insure sufficient volumes of candidate items are pretested with a sufficient volume of test-takers.

Approaches

The use of the weighted deviations model (WDM) in adaptive testing should facilitate the refreshment or replacement of pools in at least some circumstances. New or partially new pools may need to only follow the rough guidelines of being equivalent to six to eight final forms, without extensive checking for strict pool parallelism. The weighted deviations model itself will attempt to build the most parallel possible adaptive tests from the new pools, as will be discussed.

There are two approaches to pool replacement or refreshment that require new items. An obvious approach is replace an entire item pool at once. This has the disadvantage of requiring a large pretesting effort, as described above, but provides clear-cut protection of pool security.

A more conservative approach in terms of the requirement of new items might be to consider removing a small number of the most heavily used items and replacing them by equivalent ones. This approach may be more difficult to implement in practice than expected since the items are not removed at random and the WDM will be constrained by the items remaining in the pool. This implies that the new items must be reasonably close matches to the items removed -- any substantial deviations will impair the efforts of the weighted deviations model in attempting to build parallel adaptive tests from the modified pool. If the number of items replaced is large, and the replacement items are lower in statistical quality when compared to the removed items, it may be that the quality of the pool is sufficiently degraded that either a longer adaptive test is required, or no adaptive testing is possible from the pool (Stocking, 1988b).

There are a number of alternative approaches to pool refreshment or replacement that could be structured so as to limit the required number of new items while still enhancing item and pool security. The idea behind these approaches is that of reconstituted pools. A pool is considered to be reconstituted if it is constructed from items from a number of different pools. Suppose that we have two non-overlapping item pools. We can form a third (reconstituted) pool by randomly assigning half the items from each of the pools to be members of the third pool. In terms of test overlap discussed in the previous section, as well as item exposure rates, we would expect that the reconstituted pool would not be as good as having three non-overlapping pools, but would probably be better than having two non-overlapping pools.

A variant on this approach is to take all available (non-overlapping) pools and randomly assign all items to an equal number of new pools. The random assignment will guarantee that the new pools are roughly equivalent in terms of content and measurement properties, while insuring that items originally from the same pool are randomly placed in different pools.

The use of the WDM model for adaptive test item selection makes it possible to consider new sources of items that might have been more difficult to consider with different models. It may be possible, for example, to consider using disclosed items, or items from current paper-and-pencil forms in adaptive test pools. Constraints could then be written to limit the number of disclosed items or current paper-and-pencil items that could appear in any adaptive test. It might be worth allowing every test to contain one or two disclosed items and/or current paper-and-pencil items in terms of the savings in item writing efforts that would accrue.

Test Design Activities

Regardless of the approach taken to refresh or replace item pools, the adaptive tests administered must remain parallel to each other over time and over different pools. From the perspective of the test user this principle implies that the distribution of observed scores must remain the same across pools and time. In linear testing, this is the reason why test forms are built to the same extensive test specifications, and why equating is performed.

To compare adaptive tests from new or reconstituted pools to adaptive tests from the original adaptive testing pool requires the establishment of benchmark values of important features from the original pool. One approach to accomplishing this would be to repeat the final satisfactory adaptive test simulation on the original pool a number of times (perhaps 100), using a different random number seed for each repetition. This would provide model-based approximate empirical null distributions for any statistic of interest, including maximum and average exposure rates, content constraint violations, conditional means and standard errors of measurement, adaptive test reliability, predicted distributions of observed scores, and the mean and standard deviation of predicted observed scores.

All test design simulation activities should be completed for each new pool, no matter how small the difference between the new pool and the original pool in terms of numbers of items. Current research shows, for example, that the removal of only two items from a pool can effect the parallelism of adaptive tests constructed for the original and reduced pools (Potenza and Stocking, 1993). With current technology, this is relatively simple to do. Features of interest for the new pool can then be compared to the empirical distributions from the original pool, and judgements made as tɔ whether the adaptive tests from the new pool are sufficiently parallel to those from the original pool.

If adaptive tests from the new pool are not sufficiently parallel, it may be possible to change certain aspects of the test design to increase parallelism. For example, it may be necessary to change the weights on certain constraints, to accommodate the strengths or weaknesses, in the new pool in order to achieve constraint violations similar to those from the original pool. Or, it may be necessary to accept slightly higher item exposure rates to achieve content goals. Likewise, it may be imperative to lengthen or shorten the adaptive tests from the new pool in order to match the measurement properties of adaptive tests from the original pool.

Conclusions

In spite of the strong drive for secure items and pools in adaptive testing, the practical limitations of item production efforts will limit the level of security that can be realistically obtained. While there are many creative approaches to pool refreshment and/or replacement, including reconstituting pools and the limited use of disclosed items or items from companion paper-and-pencil forms, adaptive tests from all new pools must be carefully compared to benchmark distributions of important features from the original pool. If adaptive tests from new pools are sufficiently different, it may be possible to increase similarity through the adjustment of test design parameters for the new pools.

References

Eignor, D. R., Way, W. D., Stocking, M. L., and Steffen, M. (1993). Case studies in computer adaptive test design through simulation (Research Report 93-56). 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. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. Applied Measurement in Education, 4, 241-261.

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.

Potenza, M. & Stocking, M. L. (1993, in progress). Flawed items in computerized adaptive testing. Princeton, NJ: Educational Testing Service.

Schaeffer, G., Steffen, M., Golub-Smith, M. (1993). Introduction of a computer adaptive GRE general test (Research Report XX-XX). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An International Review, 36, 3/4, 263-277.

Stocking, M. L. (1988a). Scale drift in on-line calibration (Research Report 88-28-ONR, ONR Contract No. N00014-83-K-0457). Princeton, NJ: Educational Testing Service.

Stocking,, M. L. (1988b). Some considerations in maintaining adaptive test item pools (Research Report 88-33-ONR, ONR Contract No N00014-85-K-0683). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1992). Controlling item exposure rates in a realistic adaptive testing paradigm (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L., and Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. Applied Psychological Measurement, 17, 277-292.

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

Sympson J. B., and 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.

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

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

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. Vancouver, BC: Educational Research Institute of British Columbia.

Table 1: An example of controlling item type delivery in a simple adaptive testing paradigm.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
Item 1	a	a	b	b	b	b	a	a
Item 2	c	c	c	c	c	c	c	c
Item 3	d	d	d	e	e	e	e	e
Item 4	a	a	b	b	b	a	a	a
Item 5	d	d	d	e	e	e	e	d
Item 6	a	a	a	a	a	a	a	a
. . .								
Item 25	d	d	d	e	e	e	d	d

Table 2: The Five Adaptive Tests

Test	Elements in Pool	Number of items	Number of stimuli	CAT length	Reference test length	Number of constraints	Item security	Overlap
Math 1	351	333	18	28	60	24	ESH, .2	Yes
Verbal 1	372	340	32	27	76	35	ESH, .2	Yes
Analytical	469	411	58	35	50	23	ESH, .2	Yes
Math 2	241	234	7	20	60	75	Random	No
Verbal 2	330	303	27	27	85	41	Random	Yes

Table 3: Statistics for comparison of item pools and adaptive tests (see text).

Test	Content: global factor	Content: stimulus factor	Content: items associated with stimuli factor	Content: Discrete factor	Statistical: Global Factor	Statistical: Actual/Predicted
Math 1	12	9	3-4	9-11	6	.8
Verbal 1	13	10-13	2-3	8	7	.7
Analytical	12	9-10	1	8	8	1.2
Math 2	12	7	1	9-16	6	.6
Verbal 2	11	8-13	1-2	9-18	6	.5
Average	12	8.4-10.4	1.6-2.2	8.6-12.2	6.6	.8

Table 4: Maximum observed exposure rates for the final test design simulations for three measures.

Measure	Pool Size	Test Length	Maximum Observed Exposure Rate		
			Discrete	Stimulus	Stim items
Math 3	348	28	.22	.18	.18
Verbal 3	381	30	.24	.21	.19
Analytical 2	512	35	.24	.22	.22

Table 5: Proportions of test takers with different amounts of test overlap for three measures.

Percentage of overlapping items	Math 3			Verbal 3			Analytical 2		
	One pool	Two pools	Twinned items	One pool	Two pools	Twinned items	One pool	Two pools	Twinned items
10% or less	39	70	63	45	72	74	49	75	73
15% or less	57	79	84	54	77	84	64	83	88
20% or less	65	83	90	71	85	94	77	88	95
30% or less	84	92	99	88	94	99	89	94	99
Over 30%	16	8	1	12	6	1	11	6	1
# items in test	28			30			35		