ACT Research Report Series 2002-3

Controlling Item Allocation in the Automated Assembly of Multiple Test Forms

Judith Spray

Chuan-Ju Lin

Troy T. Chen

ACT

August 2002

Controlling Item Allocation in the Automated Assembly of Multiple Test Forms

Judith Spray
Chuan-Ju Lin
Troy T. Chen

Abstract

Automated test assembly is a technology for producing multiple, equivalent test forms from an item pool. An important consideration for test security in automated test assembly is the inclusion of the same items on these multiple forms. Although it is possible to use item selection as a formal constraint in assembling forms, the number of constraints is often so large to begin with that imposing additional constraints may produce unsatisfactory results. In this paper we propose an alternative method for controlling item allocation that is based on randomization. An example from an actual item pool is presented to illustrate the method.

Controlling Item Allocation in the Automated Assembly of Multiple Test Forms

The automated assembly of multiple test forms for online delivery offers an alternative to a single, computer-administered, fixed test form or even a computerized-adaptive test. The constructed forms are usually assembled according to a set of content and psychometric specifications obtained from a reference test (i.e., a test form that has been administered previously and has exhibited acceptable results in terms of form difficulty, variability, reliability, passing rate or other psychometric considerations). If the constructed tests all meet these reference specifications, by making some assumptions concerning the operating characteristics of the items, the test forms can be thought of as *equivalent* in some sense. For example, if the psychometric specifications refer to the first and second moments of target difficulty and variability for each individual examinee, the constructed test forms would be parallel if all of the psychometric specifications were met across all of the test forms. The result is that a single passing standard or score could be used across forms, eliminating the need for post-administration equating or the establishment of separate passing scores for each form.

The multiple forms may or may not consist of unique test items. Frequently, item pools from which the forms are constructed are small relative to the length and the number of forms required. Consequently, individual items may appear on more than one form. For example, if we were assembling five forms of the same test from a pool of items, each item within the pool would appear on either 0, 1, 2, 3, 4, or 5 forms. The number of items, n_m, that appear on $m = 0, 1, 2, 3, 4, 5$ forms represents the *allocation* of items across the five test forms. We refer to the appearance of items across multiply constructed test forms as *item allocation*.

If enough items appear frequently on many forms, the security of the items and the validity of the test results could be in question. One of the goals of the test assembly or

construction process should be to minimize test-overlap rate, defined as the proportion of items shared between any two forms. One way to do this is to include item usage as a constraint or target in the solution of the assembly problem. However, this may be unnecessary, especially if the form-assembly problem is burdened with numerous other constraints such as multiple levels of content categories and key balancing requirements, in addition to the psychometric requirements of the test forms. And any constraint that forces items onto a test form may end up doing so at the expense of other constraint goals. It may be more efficient to implement a simpler process to control the allocation of items across multiple forms. The purpose of this paper is to illustrate, by example, a simple randomization process that controls item allocation by minimizing the average test-overlap rate between pairs of test forms while producing tests that meet content and psychometric assembly constraints.

Ideal Item Allocation across Multiple Test Forms

What is the most ideal distribution or allocation of test items across multiple, equivalent test forms constructed from the same item pool? Obviously, the most desirable distribution or allocation from a test security standpoint is one in which there are no shared items across the forms. However, the item pool would have to be quite large relative to the length of each test form and the number of forms required to achieve this ideal. In addition, the pool would have to consist of enough "good" items so that all of the psychometric constraints could be met. And obviously if there were content constraints as well, there would have to be a sufficient number of items within each content category to satisfy the assembly goals.

If such an ideal allocation cannot occur, one might ask what is "next-best"? From a test security perspective, we want to minimize the number of times that an item appears on every constructed form or nearly every constructed form. And from a test development perspective, we

do not want the situation where a large proportion of available items in the pool never appears on a single form. The latter situation would appear to be a waste of development time and money. To accomplish this goal, we present a method of controlling item appearances on multiple test forms that is derived from random sampling without replacement. This method can be implemented with any automated test-assembly procedure. It is based on the idea that if one could guarantee that the psychometric constraints would be met, the best way to safeguard overexposure of items would be to select them from the pool or each content category at random without replacement. If this were possible, the resulting allocation of items across forms would be defined as optimal, in the sense that it minimizes average test overlap of the constructed test forms. This claim is substantiated later in this paper.

Traditional Method of Controlling Item Exposure in CAT

Because the method of controlling item inclusion on assembled test forms is very similar to the traditional tactic used to manage computerized adaptive testing (CAT) programs, it is helpful to review that approach. The typical method of controlling for item exposure in CAT situations is to use a conditional approach first suggested by Sympson and Hetter (1985). For this procedure, a maximum expected item-exposure rate, r, is first established. The goal is to find a set of item-exposure-control parameters that govern the administration of items in a CAT item pool in such a way that no single item is ever administered more than r100% of the time, where $0 \leq r \leq 1$.

The approach is called *conditional* because it is formulated within the context of a conditional probability statement. If $P_i(S)$ is the probability that item i is selected for a CAT administration, and $P_i(S,A)$ is the probability that item i is selected <u>and</u> administered (i.e., *exposed*), then an item's exposure control parameter is simply $P_i(A|S)$, the probability of

administering an item, given that it has been selected, or $P_i(A|S) = P_i(S,A) \div P_i(S)$. The purpose of this conditional probability is to allow the item to be administered only if the conditional probability is satisfied, thus controlling for the exposure of that item.

If $P_i(S,A)$ is replaced by the target-exposure rate, r, CAT simulations and an iterative procedure are used to obtain a value of $P_i(A|S)$ for each item in the pool. Simulated examinees, similar in number and ability distribution to the intended CAT examinee population, are *administered* items selected from the CAT item pool. The values of $P_i(S)$ are usually all set to 1.0 at the beginning of a set of simulations. The items are then selected on their ability to satisfy whatever constraints are required (e.g., maximum information at ability estimates, content specifications). However, they are only administered if a uniform random deviate is less than or equal to $r \div P_i(S)$. If it is not, the items are temporarily set aside until all other items have been administered to a particular examinee or the pool has been exhausted. After all N simulated examinees have taken the CAT, and the number of times each item has been selected, S_i, has been counted, $P_i(S)$ is replaced by $(S_i \div N)$ and the process begins again. $P_i(S)$ continues to be refined until such time that the proportion of times that an item has been selected <u>and</u> administered across all examinees, or $(A_i \div N)$, is close to the target value r. The number of iterations of $P_i(S)$ required before $(A_i \div N)$ approaches r is usually fairly small (Sympson & Hetter, 1985). The result is that $P_i(A|S)$ stabilizes, subsequently to be used in real CAT administrations to control item usage or exposure at a rate $\leq r$ across the examinee population. Obtaining $P_i(A|S)$ for each item in the pool is thus the goal of the simulation and iteration process for CAT.

The number of times that an item has been administered or exposed, A_i, can be assumed to be a binomial random variable with parameters $P_i(S,A)$, abbreviated as simply P_i, and N, or $A_i \sim \text{Bin}(P_i, N)$. The variance of $(A_i \div N)$, is small for large N, and therefore $(A_i \div N)$ approaches

P_i. However, the binomial distribution of A_i changes throughout the simulation and iteration process. The use of $P_i(A|S)$ to control when items are administered during the simulations causes P_i to approach r iteratively for the most popular items (i.e., those that have desirable psychometric, content, and other required characteristics), while remaining less than r (i.e., approaching a value less than r) for less desirable items.

How fast and which items converge[1] to r (or a value less than r) somewhat depends on the value of r and its relation to the observed, average item-exposure rate, $\{\Sigma[P_i] \div n\}$. Chen, Ankenmann, and Spray (1999) showed that, regardless of the pool size, n and fixed CAT test length, k, the average item-exposure rate of any fixed-length CAT is equal to $(k \div n)$. Because the target rate, r, is considered to be a maximum allowable rate for any single item, it is obvious that r must be chosen so that $r \geq (k \div n)$. Chen, et al. (1999) further showed that the average test-overlap rate, $\overline{T}$, is a function of P_i. Specifically,

$$\overline{T} = \frac{N\sum_{i=1}^{n}(P_i)^2}{k(N-1)} - \frac{1}{N-1} \ . \tag{1}$$

By completing the square in equation (1) above, they then showed that

$$\overline{T} = \frac{N\sum_{i=1}^{n}\left[\left(P_i - \frac{k}{n}\right)^2 + 2P_i\frac{k}{n} - \frac{k^2}{n^2}\right]}{k(N-1)} - \frac{1}{(N-1)} \ . \tag{2}$$

This simplifies to

$$\overline{T} = \frac{N\sum_{i=1}^{n}\left(P_i - \frac{k}{n}\right)^2 + N\frac{k^2}{n}}{k(N-1)} - \frac{1}{(N-1)} \tag{3}$$

[1] We note that the term, *convergence*, as used in this paper, describes the iterative process whereby the rates with which items in the pool are administered change after each iteration. Because the sum of these rates must always equal the length of the test, k, only variance of these rates can change; it decreases iteratively until it stabilizes. Thus, the term does not connote a statistical *convergence*, say in distribution or probability.

or

$$\overline{T} = \frac{\frac{N}{n}\sum_{i=1}^{n}\left(P_i - \frac{k}{n}\right)^2 + N\frac{k^2}{n^2}}{\frac{k}{n}(N-1)} - \frac{1}{n(N-1)} \tag{4}$$

which is equivalent to

$$\overline{T} = \frac{Var(P_i) + \frac{k^2}{n^2}}{\frac{k}{n}\frac{(N-1)}{N}} - \frac{1}{n(N-1)}. \tag{5}$$

Because the Chen, et al. (1999) paper was concerned with CAT where N is typically very large, they used a large-sample approximation for average test-overlap rate or

$$\overline{T} \doteq \frac{Var(P_i) + \frac{k^2}{n^2}}{\frac{k}{n}}. \tag{6}$$

The average item exposure, ($k \div n$), is also the probability of drawing k items from an item pool of size n randomly without replacement (see Appendix). In fact Chen, et al. (1999) showed that when $P_i = (k \div n)$, for all i, $\overline{T}$ reaches its minimum value of ($k \div n$) (i.e., when the variance of P_i is zero, the minimum value of $\overline{T}$ occurs). This suggests that perhaps the target rate, r, could be set to ($k \div n$) to minimize test overlap. However, because items are selected based on their psychometric and other characteristics and are not actually drawn at random, $r = (k \div n)$ is not a realistic target (Chen, et al., 1999). Still, a target value slightly higher than ($k \div n$) might be quite realistic and would produce a lower test overlap if this target could be reached by a majority of the items during the simulation-iteration process described earlier.

Controlling Item Allocation across a Small Number of Test Forms

In the CAT situation, N represents the number of tests that are to be given, or in this case, the examinee-population size. However, when multiple test forms are constructed for administration via computer at a later time, N represents the number of forms to be assembled. In this situation, N may be fairly small. This difference in definition and, hence, size, results in a slightly different interpretation of the goal of the Sympson-Hetter procedure. Because N is small, $(A_i \div N)$ will not converge to P_i. However, the behavior of A_i can only be described by its probability density function or pdf, $A_i \sim \text{Bin}(P_i, N)$, or

$$\text{Prob}(\text{A}_i = m_i) = \binom{N}{m_i} P_i^{m_i} \left(1 - P_i\right)^{N - m_i}, m_i = 0, 1, \ldots, N . \tag{7}$$

The *allocation of n items across N forms* is the sum of these pdfs or

$$\sum_{i=1}^{n} \left[\text{Prob}\left(A_i = m_i\right)\right] = \sum_{i=1}^{n} \left\{ \binom{N}{m_i} P_i^{m_i} \left(1 - P_i\right)^{N - m_i} \right\}, m_i = 0, 1, \ldots, N . \tag{8}$$

Likewise, each P_i will not converge closely to the target rate, r, when N is small. With only $N + 1$ possible values for the estimates of P_i to assume, it is even difficult to obtain a large degree of stability of the estimates. However, the variance of the estimates of P_i will stabilize, even after a small number of iterations.

In theory, if we set $r = (k \div n)$ we should get the item allocation that one would achieve with the random sampling of k items from a pool of n items without replacement. This would also lead to the minimum average test-overlap rate, $\overline{T}$, as in the CAT situation. However, once again, achieving the minimum test-overlap rate while meeting test-assembly specifications may not be possible, and a target that is slightly higher than $(k \div n)$ will probably need to be used.

Except for the size of N, the iterative process for CAT and for the assembly of multiple test forms is the same. A good stopping rule for the CAT iterations is to stop the process when the maximum exposure rate observed in the CAT item pool is "nearly" r, where "nearly" must be defined. For the multiple forms assembly, $\overline{T}$ can be used to stop the process. We select the item allocation that results when $\overline{T}$ is a minimum and all assembly constraints have been satisfied. Therefore, a number of iterations are specified arbitrarily and the chosen item allocation across forms is the one that produces the minimum value of $\overline{T}$ from these iterations while meeting all assembly requirements or constraints. Usually only a few iterations are necessary, as in the CAT situation.

Example

We have illustrated this procedure using a sample pool containing 247 items. Tests were constructed to be 75 items in length, and eight test forms were assembled to have the same average difficulty level (in terms of number-correct score) and variability (in terms of the standard deviation of observed test scores) as a reference form. We used the heuristic procedure developed by Swanson and Stocking (1993) using their weighted deviations model or WDM. When assembled without item-exposure control[2], the observed test-overlap rate for the construction of eight forms was .41. This meant that, on average, 41% of the items on each form were also on another form. The allocation of items without exposure control is given in Table 1 in the second column.

If 75 items were drawn completely at random without replacement from the pool with probability (75 ÷ 247) to create eight forms without regard to psychometric requirements, the

[2] In order to assemble multiple forms without item-exposure control, the first item included on a form is selected randomly. Thereafter, items are selected for inclusion based on the WDM criteria. Without random selection for the first item, all eight forms would be identical.

item allocation across the eight forms can be obtained from equation (8) using $P_i = (k \div n) \doteq .30$. Note that this is also the value of $\overline{T}$. These results appear in the fourth column of Table 1. Although unattainable in practice, we used this ideal allocation as a baseline against which to compare the item allocation that we achieved following the Sympson-Hetter iterations.

In this example, we increased the value of r on successive computer runs until a value of $r = .36$ produced eight forms that met all psychometric constraints and yielded a minimum value for $\overline{T}$. These results are given in the third column of Table 1. Thus, our results fell somewhere between the item allocation observed with no exposure control (the second column) and the random or ideal allocation (the fourth column). The use of the Sympson-Hetter procedure to find the item allocation with the smallest average test-overlap rate, $\overline{T}$, with all psychometric constraints or requirements satisfied reduced the value of $\overline{T}$ from .41 to .31. The number of items that appeared on all eight forms was reduced from 13 to 0, while the number of items that never appeared on a single form was reduced from 25 to 14.

TABLE 1

Item Allocations from the Sample Item Pool

# of Test Forms (m)	Without Item-Exposure Control (# of Items)	With Item-Exposure Control $r = .36$ (# of Items)	Random Distribution $r = (k \div n)$ (# of Items)
0	25	14	14
1	55	50	48
2	74	75	73
3	47	56	63
4	19	32	34
5	9	17	12
6	4	2	3
7	1	1	0
8	13	0	0
Test-Overlap Rate	.41	.31	.30

Item Allocations and Test Assembly under Content Constraints

The previous discussion centered on a simple assembly problem in which only psychometric constraints had to be met. However, in most multiple-form assembly problems, additional conditions or constraints involving content requirements also must be satisfied. In this situation there are J content categories, $j = 1, 2, \ldots, J$, so that the item pool of size n is stratified into $n_1, n_2, \ldots, n_J$ mutually exclusive partitions. The test-assembly specifications require that $k_1, k_2, \ldots, k_J$ items from each of these content categories appear on each assembled form, in addition to psychometric constraints.

The average test-overlap rate increases with additional content constraints because the required number of items must be drawn from smaller pools of size n_j rather than from n. Therefore, more overlap is expected, especially from those content categories where k_j is large relative to n_j. We can compute the minimal test-overlap rate, $\overline{T}_{\min}$, that would result if each test form were assembled by drawing k_j items randomly from categories of size n_j without replacement. Even though the average item-exposure rate will remain equal to $(k \div n)$, the random sampling would be stratified so that the value of P_i would depend upon the content category for that item. For stratified random sampling without replacement, the probability of an item being selected from content category j is $(k_j \div n_j)$. Thus, from equation (5), the variance of P_i would not be zero and $\overline{T}$ would increase. However, the computation of $\overline{T}$ from equation (5) under stratified random sampling would still yield a baseline test-overlap rate to use as a reference, along with an expected item allocation from equations (7) and (8).

In our sample pool, items were categorized by one of 37 mutually exclusive categories. One of the categories had only a single item represented in the pool. The test specifications called for exactly one item from this category; therefore, it was expected that this item had to

appear on all eight forms. The expected item allocation across eight forms from stratified random sampling appears in Table 2 in the fourth column. The item allocation without exposure control appears in the second column of this table.

Using $r_j = (k_j \div n_j)$ as the ideal target, we again experimented by adding a small constant, δ, to the ideal and found the smallest value of δ that would result in a minimal value of $\overline{T}$ and still meet all assembly constraints, both psychometric and content[3]. This value was $\delta = .05$. The results showed that this reduced the value of $\overline{T}$ from .49 to .36.

TABLE 2

Item Allocations from the Sample Item Pool with Content Constraints

# of Test Forms (m)	Without Item-Exposure Control (# of Items)	With Item-Exposure Control $r_j = (k_j \div n_j) + .05$ (# of Items)	Random Distribution $r_j = (k_j \div n_j)$ (# of Items)
0	41	25	24
1	59	61	51
2	57	50	61
3	32	51	52
4	24	28	34
5	9	20	17
6	4	11	6
7	2	0	1
8	19	1	1
Test-Overlap Rate	.49	.36	.35

[3] There is probably an ideal constant, δ_j, for each content category, that would produce a slightly better allocation of items. The time required to find J such values, however, may not justify the small benefit in this example. There may be other situations in which the determination of J distinct values of δ would be worthwhile.

Summary

Our results indicated that we could control the overall allocation of items across multiple test forms assembled via automated assembly methods using the same procedure that is used to control for item exposure in CAT situations. The iterative procedure was programmed directly into the form-assembly code. Thus, no "pre-assembly" work had to be done, as is done in CAT to obtain the values of $P_i(A|S)$ for later testing. In this case the iterations were a part of the assembly process, and the goal was to produce the desired item allocation across forms, rather than to obtain exposure-control parameters for each item.

References

Chen, S., Ankenmann, R. D., & Spray, J. A. (1999). *Exploring the relationship between exposure rate and test overlap rate in computerized adaptive testing*. ACT Research Report Series (99-5). ACT, Inc. Iowa City, IA.

Swanson, L., & Stocking, M. I. (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). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the 17th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Appendix

We desire the probability that one of k items will be drawn without replacement from an item pool containing n items. The easiest way to approach the problem is to compute the probability that an item will <u>not</u> be drawn, even after k attempts. Our desired probability is then the complement of this probability.

The probability that an item will not be drawn on the first attempt is $[(n - 1) \div n]$. The probability that the item will not be drawn without replacement on the second attempt is $[(n - 2) \div (n - 1)]$. For the third attempt, it is $[(n - 3) \div (n - 2)]$. For the kth and last attempt, it is $[(n - k) \div (n - k + 1)]$. Because these are independent draws, the probability that the item will not be drawn after all k attempts is their product, or

$$\frac{(n-1)}{n}\frac{(n-2)}{(n-1)}\frac{(n-3)}{(n-2)}\cdots\frac{(n-k)}{(n-k+1)} = \prod_{i=1}^{k}\frac{(n-i)}{(n-i+1)},$$

which, after cancellation, simplifies to $(n - k) \div n$. Therefore, the probability that an item <u>will be</u> selected without replacement is $1 - (n - k) \div n$ or $(k \div n)$.