Graduate Management Admission Council®

# The Shadow-Test Approach: A Universal Framework for Implementing Adaptive Testing

**Wim J. van der Linden**
**University of Twente, The Netherlands**

*Keynote Address Presented June 8, 2007*

## Editor's Note

**The keynote address presented at the conference was based in large part on the following chapter from Professor van der Linden's book *Linear Models for Optimal Test Design*, published in 2005 by Springer. This chapter is reproduced here with the permission of Springer Science+Business Media, Inc.**

## Acknowledgment

**Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.**



## Citation

**van der Linden, W. J. (2007). The shadow-test approach: A universal framework for implementing adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from** www.psych.umn.edu/psylabs/CATCentral/

## Author Contact

**Wim J. van der Linden, Chief Research Scientist, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, U.S.A. Email: wim_vanderlinden@ctb.com**

# 9
# Models for Adaptive Test Assembly

In adaptive test assembly, items are selected from the pool in a sequential fashion. The response to a new item is used to update the test taker's ability estimate, and the next item is then selected to have maximum accuracy at the updated estimate. Adaptive tests are highly efficient. Extensive computer simulations have shown that to yield equally accurate ability estimates, adaptive tests require only 40–50% of the items needed for a fixed test.

Large-scale adaptive testing became possible in the 1990s when the computer revolution led to PCs that were powerful enough to perform real-time ability estimation and item selection at the testing site. Until then, adaptive testing had been restricted to an occasional experimental program run on a mainframe computer or to the use of approximate, computationally less intensive procedures for ability estimation and item selection.

One approximate procedure was multistage testing in which adaptation took place only at the level of a few alternative subtests, with estimation of $\theta$ replaced by simple number-correct scoring of the subtests. Also, to avoid the use of computers, experiments with paper-and-pencil adaptive testing at the level of the individual items were conducted using a testing format known as flexilevel testing. In this format, the test taker answered the items by scratching an alternative on the answer sheet, whereupon a reference to the next item on the test form became visible. The form was organized such that the difficulty of the items went up after a correct response and down after an incorrect response. The absolute size of the adjustments followed a predetermined sequence of numbers. This item-selection procedure, more

generally known in statistics as a Robbins-Monro procedure, was also tried in the early computerized versions of adaptive testing.

At first, research on adaptive testing was primarily statistical and addressed such topics as the relative efficiency of adaptive tests versus fixed tests, alternative criteria for item selection, simplifying methods for updating ability estimates, the stabilization of ability estimators as a function of test length, and the impact of the choice of the initial item. But the introduction of large-scale adaptive testing in the 1990s marked the beginning of a whole new era of adaptive-testing research.

For one thing, it became clear that it was not enough for adaptive tests to be statistically efficient. In addition, they had to meet the same content specifications as their paper-and-pencil predecessors, which created a dilemma because when items are selected adaptively, each test taker gets a different set of items from the pool. Next, the issue of how to deal with adaptive tests from pools with set-based items quickly became manifest. An even more urgent new problem was item security. By its very nature, an adaptive-testing algorithm tends to capitalize on a small set of high-quality items in the pool; if the item pool is used for some time, those items become vulnerable to security breaches. An equally serious, though less generally understood, problem is that of differential speededness in adaptive testing. Since test takers get different sets of items, some may end up with a set that is very time-intensive, whereas others have ample time to answer their items. Finally, as the practice of releasing the items after the test became no longer affordable for adaptive testing, the question of how to use item content to report test scores in an informative fashion announced itself quickly.

The parallels between these early developments in adaptive test assembly and those after Birnbaum introduced his approach to fixed test assembly in 1968 are conspicuous. As was already pointed out in Section 1.2.8, Birnbaum's approach focused exclusively on the statistical aspects of testing, too. To become realistic, it had to be adjusted to problems with elaborate sets of content specifications and more complicated types of item pools. Moreover, the initial heuristic techniques for item selection developed to implement Birnbaum's method had to be replaced by flexible algorithms with optimal results. Basically, the same challenges were met again when the first large-scale adaptive-testing programs were launched in the 1990s.

In this chapter, we will show how the optimal test-assembly approach in the previous chapters can also be used to solve the problems in adaptive testing above. We will first model adaptive test assembly from a pool of discrete items as an instance of 0-1 programming. The result will be a modification of the standard test-assembly model for a single fixed test presented in Section 4.1. Although the model changes only a little, a major difference with fixed-form test assembly is that an updated version of the model has to be solved for the selection of each subsequent item in the test. We will then discuss a few alternative objective functions for adaptive

testing and extend our standard model with the different sets of constraints necessary to solve the various new problems discussed above. Some of these constraints are direct generalizations of the ones we used to extend the model for a fixed test in Section 4.1 to those for the special problems in Chapters 5–8, whereas others require a bit more ingenuity.

Because both our basic model for adaptive test assembly and all additional constraints remain linear in the decision variables, the branch-and-bound searches discussed in Section 4.2 can also be used to run an adaptive-testing program. As a matter of fact, thanks to a special feature of adaptive item selection, much faster implementations of these searches than for fixed test assembly are possible, which enables us to execute them in real time.

## 9.1 Shadow-Test Approach

In Section 2.3, we classified the objective used in adaptive testing as a quantitative objective at the item level. The objective was modeled in Section 3.3.1 as

$$\text{maximize} \sum_{i \in R} I_i(\widehat{\theta}) x_i \quad \text{(maximum information)} \tag{9.1}$$

subject to

$$\sum_{i \in R} x_i = 1, \quad \text{(selection of one item)} \tag{9.2}$$

$$x_i \in \{0, 1\}, \quad \text{for all } i, \quad \text{(range of variables)} \tag{9.3}$$

where $R$ denotes the items in the pool that the person has not yet taken and $\widehat{\theta}$ is his or her current ability estimate.

The representation is not yet realistic for the following reasons:

1. For a complete test, a set of these problems has to be solved, one for each of its items.

2. For each of the problems in this set, we have to update the estimate $\widehat{\theta}$ in the objective function.

3. The problems in the set are dependent in the sense that once an item has been administered, it cannot be chosen again in a later problem.

Thus, the model in (9.1)–(9.3) needs to be reformulated. We will do this first for adaptive testing with a random test length and then for the case of a fixed test length.

### 9.1.1 Random Test Length

In early research on adaptive testing, the ideal of a stopping rule based on a common level of accuracy of the final ability estimator $\widehat{\theta}$ for all test takers was advocated. This ideal can easily be realized by estimating the accuracy of $\widehat{\theta}$ online, monitoring the estimates, and stopping when a predetermined threshold is passed.

This case of a random test length can be modeled as follows. Let $g = 1, 2, ...$ denote the items in the adaptive test. We use $\widehat{\theta}^{(g-1)}$ to represent the update of the ability estimate after the first $g-1$ items; this update is thus used to select the $g$th item. Furthermore, we use $R_g$ to denote the set of items in the pool that is still available for administration when the $g$th item for the adaptive test is selected.

The previous model for the selection of the $k$th item can then be reformulated as:

$$\text{maximize} \sum_{i \in R_g} I_i(\widehat{\theta}^{(g-1)})x_i \quad \text{(maximum information)} \tag{9.4}$$

subject to

$$\sum_{i \in R_g} x_i = 1, \quad \text{(selection of one item)} \tag{9.5}$$

$$x_i \in \{0, 1\}, \quad \text{for all } i. \quad \text{(range of variables)} \tag{9.6}$$

This model depicts adaptive test assembly as a process in which after each new response the ability estimate $\widehat{\theta}^{(g-1)}$ and the set of items $R_g$ are updated and the model is run again to select the next item. The only difference with the preceding model is that (9.1) and (9.2) are now formulated explicitly as a *dynamic* objective function and constraint.

### 9.1.2 Fixed Test Length

The ideal of a common level of accuracy for the ability estimator was soon abandoned in favor of adaptive testing with the same fixed number of items for all persons who take the test. Typically, this length is chosen to guarantee a desirable minimum level of accuracy over the ability range for which the test is used. The choice of a fixed test length was necessary to deal with the requirement that each test taker get a test assembled for a common set of content specifications. A more mundane reason was that adaptive-testing sessions cannot be planned efficiently if the time spent on the test is not known in advance. In this chapter, we further focus on the case of adaptive test assembly with a fixed test length.

To select an adaptive test of fixed length, we could simply run the model in (9.4)–(9.6) and stop after a fixed number of items has been administered. But a much more useful model arises if we adopt an explicit constraint on

the test length in the model. If we do so, we also must have a constraint that sets the variables of the items already administered equal to one.

The result is the following set of models for $g = 1, ..., n$:

$$\text{maximize} \sum_{i=1}^{I} I_i(\widehat{\theta}^{(g-1)})x_i \quad \text{(maximum information)} \tag{9.7}$$

subject to

$$\sum_{i=1}^{I} x_i = n, \quad \text{(test length)} \tag{9.8}$$

$$\sum_{i \in \overline{R}_g} x_i = g - 1, \quad \text{(previous items)} \tag{9.9}$$

$$x_i \in \{0, 1\}, \quad \text{for all } i, \quad \text{(range of variables)} \tag{9.10}$$

where $\overline{R}_g$ is the set of items not in $R_g$; that is, the items already administered. Observe that (9.9) is a single constraint that sets each of the $g - 1$ variables in $\overline{R}_g$ equal to one; this type of constraint was already introduced in (3.28).

Instead of selecting one item, this model selects an entire test of $n$ items with maximum information at $\widehat{\theta}^{(g-1)}$. The item that is administered as the $g$th item in the adaptive test is the one among the $n - g$ free items with the maximum value of $I_i(\widehat{\theta}^{(g-1)})$. Thus, the original objective of an item with maximum information at the ability estimate is now realized through a two-stage optimization procedure in which:

1. a test of length $n$ with maximum information at $\widehat{\theta}^{(g-1)}$ is assembled;
2. the free item in this test with maximum information at $\widehat{\theta}^{(g-1)}$ is selected.

Although at first sight the model in (9.7)–(9.10) may look somewhat overdone, it has an attractive feature: Its constraint set can be extended with whatever other content constraint we find useful. Because the same set of constraints is imposed on each of the $n$ tests assembled for the test takers, the adaptive test automatically satisfies each of its constraints. We document this feature as an explicit principle:

> *Any type of constraint available to give a* fixed *test a certain feature can be inserted in the basic model in (9.7)–(9.10) to give an* adaptive *test the same feature.*

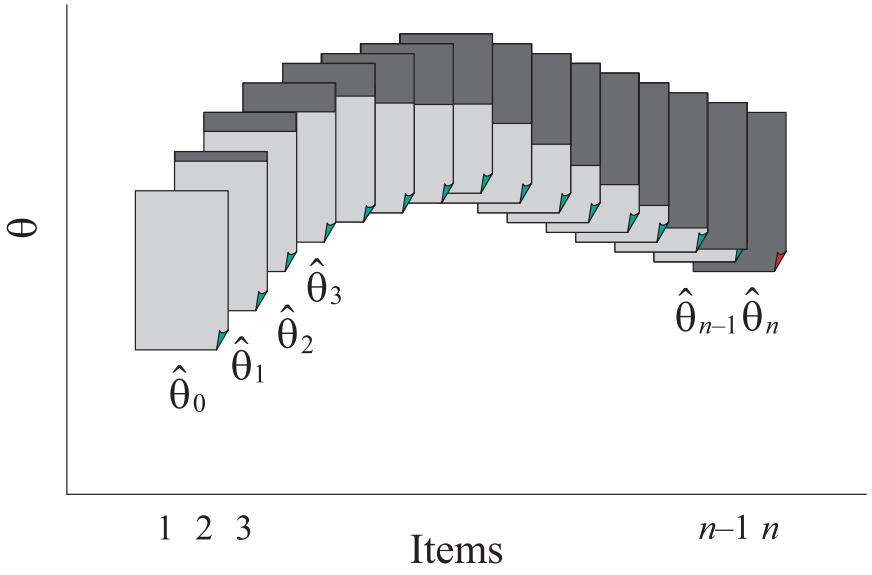


FIGURE 9.1. Graphical representation of the shadow-test approach to adaptive test assembly.

### *9.1.3 Definition of Shadow Tests*

Tests calculated by the model in (9.7)–(9.10) are known as *shadow tests* in adaptive testing. Their name reminds us of the big-shadow-test method for multiple-test assembly in Section 6.3. In this method, shadow tests were assembled to keep a sequence of test-assembly problems feasible with respect to their constraints. The same principle is implemented at the item level by (9.7)–(9.10). Items in a shadow test that are not selected for administration are returned to the pool; they are made available again when the shadow test is reassembled at the next update of $\widehat{\theta}$. Their only purpose is to keep the adaptive test feasible with respect to its constraints when an individual item is selected.

A graphical representation of the shadow-test approach (STA) to adaptive testing is given in Figure 9.1. The horizontal axis represents the sequence of items in the adaptive test and the vertical axis the ability measured by the item pool. The vertical position of the shadow tests corresponds with the current ability estimate; the higher the test, the larger the estimate. Also, a higher position indicates that the preceding response was correct and a lower position that it was incorrect. The smaller differences between the vertical positions of the shadow tests toward the end of the test reflect the stabilization of the ability estimates. The darker portions of the shadow test represent the items already administered and the lighter portions the parts that are reassembled at the new ability estimate. The

last test is the actual adaptive test that has been administered; it automatically meets the common set of content constraints imposed on each of the individual shadow tests.

### *9.1.4 Standard Model for a Shadow Test*

For a pool of discrete items, the standard model for a shadow test follows directly from the model for a fixed test introduced in Section 4.1. For the selection of item $g = 1, ..., n$, the model is

$$\text{maximize} \sum_{i=1}^{I} I_i(\widehat{\theta}^{(g-1)})x_i \quad \text{(objective)} \tag{9.11}$$

subject to possible constraints at the following levels:

*Test Level*

$$\sum_{i=1}^{I} x_i = n, \quad \text{(test length)} \tag{9.12}$$

$$\sum_{i \in V_c} x_i \gtreqless n_c, \quad \text{for all } c, \quad \text{(categorical attributes)} \tag{9.13}$$

$$\sum_{i=1}^{I} q_i x_i \gtreqless b_q; \quad \text{(quantitative attributes)} \tag{9.14}$$

*Subtest Level*

$$\sum_{i \in \overline{R}_g} x_i = g - 1; \quad \text{(previous items)} \tag{9.15}$$

*Item Level*

$$\sum_{i \in V_1} x_i = n_1, \quad \text{(categorical attributes)} \tag{9.16}$$

$$\sum_{i \in V_0} x_i = 0, \quad \text{(categorical attributes)} \tag{9.17}$$

$$q_i x_i \leq b_q^{\max}, \quad \text{for all } i, \quad \text{(quantitative attributes)} \tag{9.18}$$

$$b_q^{\min} x_i \leq q_i, \quad \text{for all } i, \quad \text{(quantitative attributes)} \tag{9.19}$$

$$\sum_{i \in V_e} x_i \leq 1, \quad \text{for all } e; \quad \text{(enemies)} \tag{9.20}$$

*Definition of Variables*

$$x_i \in \{0,1\}, \quad \text{for all } i. \quad \text{(range of variables)} \tag{9.21}$$

Observe that, except for the objective function in (9.11) and the extra constraint in (9.15), the model is identical to that for a single fixed test with discrete items. This fact illustrates our earlier observation that an adaptive test can be given any feature possible for a fixed test.

The only dynamic quantities in the model are the objective function and constraint in (9.11) and (9.15). The updates of the objective function make the selection of the items adaptive with respect to the interim ability estimates and, in doing so, give the test its favorable statistical features for the estimation of $\theta$. The updates of the constraint force the new items in the shadow test to have attributes that complement the attributes of the items already administered with respect to the other constraints in (9.9)–(9.20) (Exercise 9.1).

### *9.1.5 Calculating Shadow Tests*

Shadow tests in adaptive testing have to be calculated in real time. It is therefore important to use a fast implementation of the branch-and-bound search discussed in Section 4.2. Fortunately, such implementations are possible because of a special feature of the set of $n$ models for the shadow tests in (9.11)–(9.21).

The differences between the models for two subsequent shadow tests reside in the updates of (9.11) and (9.15). Generally, the collection of feasible tests in a test-assembly problem is determined only by the constraints in the model. The changes in the objective function in (9.11) thus do not have any impact on this set. On the other hand, the update of the constraint in (9.15) does have an impact. But since the update consists only of the fixing of one more decision variable, and all other constraints in (9.11)–(9.21) remain the same, the collection of feasible tests for the next problem is always a subset of the collection for the preceding problem. More importantly, since the variable is fixed at the value found in the solution of the preceding solution, this solution remains feasible for the next problem. Finally, because the value of the objective function changes gradually between subsequent shadow tests, new shadow tests tend to be found in the neighborhood of their predecessor in the feasible collection.

This argument shows that the preceding shadow test is a good initial solution in a branch-and-bound algorithm for the calculation of a shadow test. In all our applications, this choice has dramatically sped up the search for shadow tests. For problems with constraint sets and item pools comparable to those in the empirical examples in the next section, the current integer solver in *CPLEX* (Section 4.2.5) finds the shadow tests in a split second.

When selecting the first item in the adaptive test, no previous shadow test is available. However, an initial solution can easily be calculated before the adaptive testing program is operational. All we have to do is choose a typical value for $\widehat{\theta}$ in the objective function in (9.11) and calculate a solution. Because the content constraints remain identical for all test takers, the same initial solution can be used for each of them.

As an aside, we observe that the same argument shows that the presence of the constraints on the previous items in (9.15) can never be a reason for a shadow test to become infeasible. If an initial shadow test can be assembled, all later problems are feasible. Because the shadow tests for all test takers are subject to a common set of content constraints, it follows that if the item pool has at least one feasible test, the STA will never run into feasibility problems.

### *9.1.6 Empirical Example*

Computer simulations of a 50-item adaptive version of the *Law School Admission Test* (LSAT) were conducted. The item pool was the same pool of 753 items used in the previous examples for the LSAT. The full set of specifications for the test was used, including the specifications needed to deal with the item-set structure of two of the three sections in the test. (The topic of how to model adaptive tests with item sets will be addressed in Section 9.3.) The only difference with the earlier paper-and-pencil version of the LSAT was a proportional reduction of the bounds in the constraints to account for the reduction of the test length to 50 items. The number of variables and constraints needed to model the three sections were: 232 variables and 179 constraints for Section SA, 264 variables and 218 constraints for Section SB, and 305 variables and 30 constraints for Section SC.

The order of the sections in the adaptive test was (i) SC, (ii) SA, and (iii) SB. This order allowed the ability estimator in the objective function to stabilize somewhat before the more severely constrained sections were introduced. The test administrations were replicated for 100 test takers at $\theta = -2.0, -1.5, ..., 2.0$. The first item in the test was selected at a common value $\widehat{\theta} = 0$ for all test takers. The updates of the ability estimates were calculated using the expected a posteriori (EAP) estimator with a uniform prior distribution.

The study was repeated a second time without any of the content constraints on the items. The differences between the results for these two studies enable us to evaluate the efficiency of the STA in the presence of large numbers of constraints on the test.

Because the test takers were simulated, we knew their true ability levels and were able to estimate the bias and mean-squared error (MSE) in their ability estimates. Figure 9.2 shows the estimated bias as a function of the true values in the simulation study after 10, 20, 30, and 40 items were administered. The differences between the functions for adaptive testing

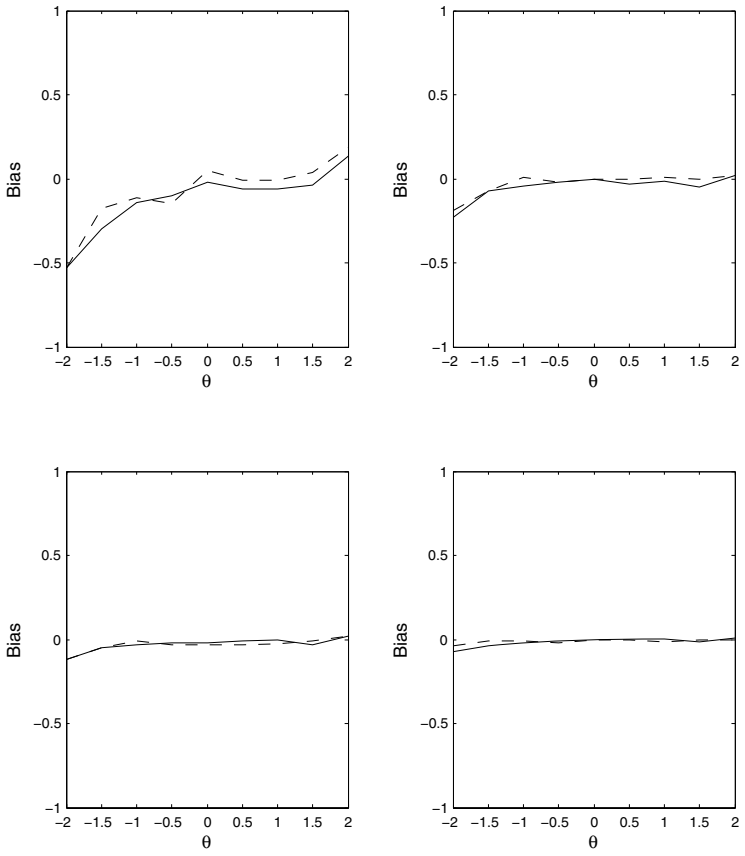


FIGURE 9.2. Bias functions for $n = 10$ and 20 (top) and $n = 30$ and 40 (bottom) for adaptive testing with (solid line) and without (dashed line) content constraints.

with and without the constraints were already small after $n = 10$ items but, for all practical purposes, disappeared with the increase in test length. Essentially the same results were observed for the MSE functions in Figure 9.3.

The main conclusion from this study is that the presence of large sets of content constraints on the tests did not have any noticeable impact on the quality of the ability estimation, and the STA appeared to be an efficient way to impose these constraints.

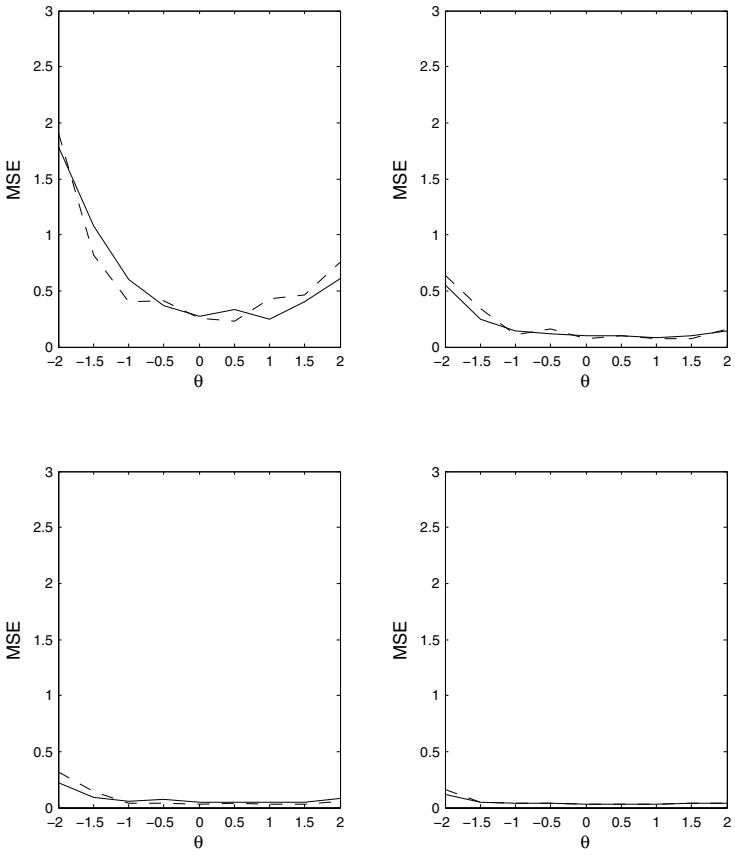


FIGURE 9.3. MSE functions for $n = 10$ and 20 (top) and $n = 30$ and 40 (bottom) for adaptive testing with (solid line) and without (dashed line) content constraints.

### *9.1.7 Discussion*

There exists a fundamental dilemma between sequential and simultaneous item selection in adaptive testing: To realize the objective of maximum information, items have to be selected sequentially with an update of the ability estimate in the objective function after each item. But to realize the constraints on the test, they have to be selected simultaneously. If they were selected sequentially, we could easily run into the problems illustrated in Table 4.1: After a few items, it would become impossible to select a new

item without violating a constraint and/or making a suboptimal choice. The STA solves the dilemma by treating adaptive test assembly as a *sequence of n simultaneous optimization problems*. The importance of this principle was already explained in Section 6.3 when we introduced the big-shadow-test method for the assembly of multiple tests.

Another way to view the STA is as a *projection* method. At each ability update, the STA projects the remaining part of the adaptive test for the test taker and picks the best item from this projection.

It is instructive to compare this interpretation of the STA with an application of the Swanson-Stocking heuristic in adaptive testing (Section 4.4.3) known as the weighted-deviations method (WDM) for adaptive testing. The WDM selects the items using the criterion of a weighted sum of projections of the deviations of the contributions of the items from the bounds in the constraints of the model. Suppose $g-1$ items have been administered. The weighted sum was formulated in (4.20) as

$$\sum_{h=1}^{H} w_h \left| \pi_{i_g h} - b_h \right|, \tag{9.22}$$

where $h = 1, ..., H$ denote the constraints on the test, $b_h$ their bounds, and $\pi_{i_g h}$ the prediction of the contribution of item $i$ to constraint $h$ when selected as the $g$th item in the test. The prediction is calculated according to (4.19) as

$$\pi_{i_g h} = \sum_{j=1}^{g-1} a_{i_j h} + a_{i_g h} + (n-g) \frac{\sum_{i \in R_g \setminus \{i_g\}} a_{ih}}{(I-g)}. \tag{9.23}$$

The first term in this prediction is the sum of the attribute values $a_{i_g h}$ in constraint $h$ for the first $g-1$ items in the test, the second term is the attribute value of candidate item $i$, and the last term equals $n-g$ times the average attribute value calculated over all remaining items in the pool.

The last term of (9.23) is the actual projection of the remaining part of the adaptive test for each constraint $h$ by the WDM. But it is unlikely that the pool contains a feasible subset of $n-g$ items with average values for $H$ attributes equal to these projections. Therefore, the WDM is vulnerable to constraint violation. The STA uses the *best feasible subset* of $n-g$ items in the pool at the ability estimate as a projection and does not suffer from this problem.

## 9.2 Alternative Objective Functions

The standard model for a shadow test in (9.11)–(9.21) offers a menu of options from which a choice has to be made to accommodate our applications. Several of these options will be discussed later in this chapter. We first

discuss a few alternative choices for the objective function in (9.11). The discussion will be rather concise; for a more elaborate treatment of these objectives, we refer to the literature on the statistical aspects of adaptive testing at the end of this chapter.

### *9.2.1 Kullback-Leibler Information*

The objective function in (9.11) is based on Fisher's information measure in (1.18). A useful alternative measure for adaptive testing is *Kullback-Leibler information*. For item $i$, this information is defined as

$$K_i(\widehat{\theta},\theta) = \mathcal{E}\left[\ln \frac{p_i(\widehat{\theta})^{U_i}[1-p_i(\widehat{\theta})]^{1-U_i}}{p_i(\theta)^{U_i}[1-p_i(\theta)]^{1-U_i}}\right], \tag{9.24}$$

where $p_i(.)$ is the response function for the 3PL model in (1.16), $U_i$ is the random response on item $i$ for a test taker with true ability $\theta$, and $\widehat{\theta}$ is the estimate of $\theta$. The expectation is taken with respect to the distribution of $U_i$ at $\widehat{\theta}$.

Since $\theta$ is unknown, it should be integrated out of (9.24), preferably over its posterior distribution. Let $f(\theta \mid u_{i_1},...,u_{i_{g-1}})$ denote the density of the posterior distribution of $\theta$ after the responses $(u_{i_1},...,u_{i_{g-1}})$ to the first $g-1$ items. The $i$th item is selected such that

$$K_i(\widehat{\theta}) = \int K_i(\widehat{\theta},\theta) f(\theta \mid u_{i_1},...,u_{i_{g-1}}) d\theta \tag{9.25}$$

is maximal.

This criterion can easily be adopted in the model for the shadow test. The only thing we have to do is replace the objective function in (9.11) by

$$\text{maximize} \sum_{i=1}^{I} K_i(\widehat{\theta}^{(g-1)}) x_i. \tag{9.26}$$

The $g$th item in the adaptive test is then the one in the shadow test with a maximum value for (9.25).

An attractive feature of Kullback-Leibler item selection is its robustness to the uncertainty on $\widehat{\theta}$ in the beginning of the test. Another attractive feature is that the criterion for item selection in (9.25) generalizes easily to multidimensional response models. We will show this generalization in an application of the STA to multidimensional adaptive testing in Section 9.7.

### *9.2.2 Bayesian Item-Selection Criteria*

Although the use of information measures dominates, item selection in adaptive testing is a natural area for the application of Bayesian criteria.

One obvious criterion is to select the items that minimize the expected posterior variance of $\theta$.

Suppose we evaluate the candidacy of item $i$ in the pool for administration as the $g$th item in the test. If item $i$ is selected, the posterior variance of $\theta$ can be expected to be

$$\mathcal{E}[\text{Var}(\theta \mid u_{i_1}, ..., u_{i_{g-1}}, U_i)], \tag{9.27}$$

where the expectation is taken with respect to the posterior predictive distribution of $U_i$ given the responses $(u_{i_1}, ..., u_{i_{g-1}})$. This expected value is a useful criterion for item selection; it tells us what reduction of posterior variance of $\theta$ to expect if item $i$ is administered.

A shadow-test approach with this criterion has the objective function

$$\text{minimize} \sum_{i=1}^{I} \mathcal{E}\left[\text{Var}(\theta \mid u_{i_1}, ..., u_{i_{g-1}}, U_i)\right] x_i. \tag{9.28}$$

This function results in the selection of a test with the $n-g$ items with the largest reductions of the expected posterior variance of $\theta$ possible given all constraints on the test. The item that is administered is the one with the smallest value for (9.27).

Other Bayesian criteria are possible, all of which share the fact that they are defined for the posterior distribution of $\theta$ given the preceding item responses. In fact, if the Kullback-Leibler measure is integrated over the posterior distribution, as we suggested for (9.25), it can also be considered as a Bayesian criterion for item selection.

## 9.3 Adaptive Testing with Item Sets

If the item pool contains set-based items, the model for the shadow test in (9.11)–(9.21) should be adjusted to the model for the simultaneous selection of items and stimuli in (7.3)–(7.24). Obviously, we retain the objective function in (9.11) as well as the constraint on the previous items in (9.15). The only addition necessary is a constraint on the previous stimuli in the test comparable to (9.15).

Let $l = 1, ..., m$ represent the stimuli in the adaptive test, and suppose that the test taker has already seen the first $l-1$ stimuli. We use $R_l$ to denote the set of stimuli in the pool available for selection as the $l$th stimulus in the test. Thus, $\overline{R}_l$ is the set of $l-1$ stimuli already administered. Analogously to (9.15), the following constraint has to be added to the model:

$$\sum_{s \in \overline{R}_l} z_s = l - 1. \tag{9.29}$$

This constraint must be updated every time a new stimulus is administered.

Typically, if a new stimulus becomes active, we want to continue administering items for it until the bounds on the number of items in (7.12) are satisfied. This practice is supported by the updates of (9.29). Once the first item for a stimulus is selected, (9.29) sets the decision variable for the stimulus equal to one. Consequently, the bounds on the number of items in (7.12) become active and remain so during the rest of the test. We are therefore able to administer the best items from the set until no free item in it is left. Then, an item for a new stimulus is selected and the process is repeated (Exercise 9.2).

Two of the three sections of the LSAT in the example in Section 9.1.6 were set-based. The model for the shadow tests for these two sections in the example was exactly as described here.

## 9.4 Controlling Item Exposure

It is necessary to control the exposure rates of the items in adaptive testing. The objective of maximum information in (9.11) involves a preference for the small subset of items in the pool with information functions that dominate the IIFs of all other items in the pool over an interval of $\theta$. Without control, these items would be frequently seen by test takers and easily passed on to a person who takes the test later.

The preference for the subset of most informative items is mitigated somewhat by the content constraints imposed on the test. To satisfy the constraints, it becomes necessary to select items with less than optimal information at the ability estimate. But then if the item pool has relatively few items with the combinations of attributes required by the constraints, these items easily become overexposed.

In this section, we discuss three different methods of item-exposure control for the STA. Two of these methods are based on the idea of adding one or more constraints to the model for the shadow test with a direct impact on the item-exposure rates; they differ only in the type of constraint that is imposed. The other method is an implementation of the well-known Sympson-Hetter method for the STA. Furthermore, in Chapter 11 we will discuss a few approaches to item-pool design for adaptive testing that prevent tendencies to overexpose items by building special constraints into the design of the item pool.

### *9.4.1 Alpha Stratification*

Although the selection of items with maximum information at the test taker's ability estimate $\widehat{\theta}$ makes intuitive sense, it is not necessarily a strategy that is always good. In the beginning of the test, when the errors in $\widehat{\theta}$ are relatively large, the item with the highest peak for its information

function at $\widehat{\theta}$ is not necessarily the one with the highest peak at the *true* $\theta$ value of the test taker. For example, it is easy to find two points on the ability scale in Figure 1.4, one for a test taker's true ability and the other for an estimate of it, where the best item at the latter is not the best at the former. Furthermore, if we select the items with the highest peaks for their information functions early in the test, we cannot use them later when the estimates $\widehat{\theta}$ are close to the true $\theta$.

In alpha-stratified adaptive testing, we select less informative items first and postpone the selection of the more informative items until the end of the test. This strategy is implemented by stratifying the pool on the values of item-discrimination parameter $a$ and restricting the selection of the items to the consecutive strata in the pool. That it is effective to stratify the pool on the values of $a$ for the items instead of their more complicated information functions follows from the analysis of the impact of $a$ on the IIF at the end of Section 1.2.4.

The option of alpha stratification is discussed here not because of its potentially beneficial impact on ability estimation but because of its tendency toward more uniform item usage. It prevents the adaptive-testing algorithm from capitalizing on a small subset of items with high values for $a$ and enforces a more uniform distribution of the items in the test on this parameter.

Alpha stratification can be implemented simply by introducing a new constraint into the model for the shadow tests. Suppose the item pool has been stratified into strata $Q_p$, with $p = 1, ..., P$. The question of how to stratify a pool is postponed until Section 11.5.1. We assume that fixed numbers of items $n_p$ are selected from the strata. The constraint to be imposed on the shadow tests is

$$\sum_{i \in Q_p} x_i = n_p. \tag{9.30}$$

The set $Q_p$ in this constraint is updated after $n_1, ..., n_P$ items in the test; (9.30) is thus another example of a dynamic constraint on an adaptive test. Because fixed numbers of items are selected from each stratum, the constraint on the total length of the adaptive test in (9.12) is redundant and can be removed from the model.

If the strata in the pool are chosen to be narrow, the only remaining parameter with an impact on the information functions of the items is the difficulty parameter $b_i$. (The guessing parameter has hardly any impact on item selection in adaptive testing.) It has therefore been suggested to simplify item selection for alpha-stratified adaptive testing and select the items from the strata with their values for $b$ closest to $\widehat{\theta}^{(g-1)}$. This suggestion implies an objective with a goal value for the shadow test (see Section 3.3.3).

Minimax implementation of this objective leads to the replacement of (9.11) by

$$\text{minimize } y \tag{9.31}$$

subject to

$$\left(b_i - \widehat{\theta}^{(g-1)}\right) x_i \leq y, \quad \text{for } i \in Q_p, \tag{9.32}$$

$$\left(b_i - \widehat{\theta}^{(g-1)}\right) x_i \geq -y, \quad \text{for } i \in Q_p. \tag{9.33}$$

## Empirical Example

The 50-item adaptive version of the LSAT in Section 9.1.6 was used to evaluate the impact of the alpha-stratification constraint in (9.30) on the exposure rates of the items. The item pool was divided into $P = 5$ strata each with 20% of the items. The maximum values of $a$ for the successive strata were: .559, .700, .813, .959, and 1.686. From each stratum, $n_p = 10$ items were selected for the test.

The model for the shadow test had all earlier content constraints for the LSAT, but the item-set structure for two of its sections was ignored. We simulated test administrations both with and without alpha stratification. For the case with alpha stratification, we also used the objective function in (9.31)–(9.33). The two cases were repeated with all content constraints removed from the model, which enabled us to evaluate the impact of alpha stratification on the exposure rates above and beyond the impact of the content constraints.

The distributions of the exposure rates of the 753 items in the pool for the four cases are displayed in Figure 9.4. For adaptive testing both with and without content constraints, the exposures rates with the alpha-stratification constraint were much closer to uniformity than without the constraint. The presence of the content constraints in the model did not have any discernible impact on these distributions. Although alpha stratification did have a favorable impact on the exposure rates, the results still show high rates for a few items in the pool.

We conclude that the introduction of the alpha-stratification constraint in the model appears to be an effective method for obtaining a more uniform distribution of the item-exposure rates. To reduce the rates for the last few items, stricter control is necessary, however. Stricter control is obtained if we increase the number of strata, $P$. But the problem would then likely become overconstrained, and the statistical quality of the ability estimates might deteriorate seriously.

More importantly, in practice, we always want to control the *conditional* exposure rates of the items at a series of well-chosen $\theta$ values because control of the marginal rates only does not preclude the possibility of high rates

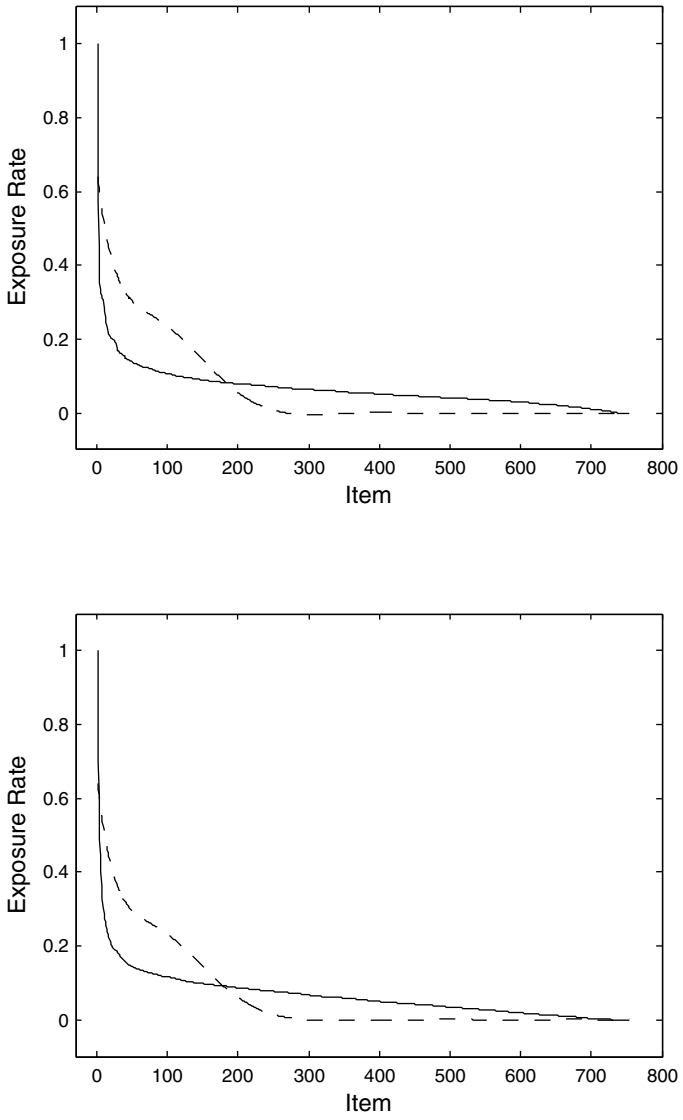


FIGURE 9.4. Item-exposure rates for adaptive testing without (dashed line) and with (solid line) alpha stratification for the cases without (top) and with (bottom) content constraints on the test. The items are ordered by exposure rate.

for test takers with ability levels close to each other. Without conditional control, it would still be easy for such test takers to share items.

Alpha stratification by itself does not control the conditional exposure rates of all items. We are therefore more in favor of combining a smaller number of strata with additional probabilistic control of the conditional item-exposure rates. The following sections describe two such probabilistic methods that can be used with the STA.

### *9.4.2 Sympson-Hetter Method*

The Sympson-Hetter method reduces the exposure rates of the more popular items by introducing a probability experiment in the adaptive-testing algorithm that is conducted after each item has been selected. The experiment is to decide if the item is actually administered or if it is rejected in favor of the item with the next-highest information at $\widehat{\theta}$.

For the case of conditional item-exposure control, the Sympson-Hetter experiment can be modeled as follows. Let $S_i$ be the event of item $i$ being selected and $A_i$ the event of item $i$ being administered. The conditional probabilities of these events given $\theta$ are denoted as $P(S_i \mid \theta)$ and $P(A_i \mid \theta)$, respectively. The probabilities $P(A_i \mid \theta)$ are the expected conditional exposure rates of the items. An item cannot be administered without being selected. Thus, $P(S_i \mid A_i, \theta) = 1$. It therefore holds that

$$P(A_i \mid \theta) = P(A_i \mid S_i, \theta)P(S_i \mid \theta). \tag{9.34}$$

For a given item pool and set of constraints on the tests, the probabilities $P(S_i \mid \theta)$ are fixed for all items. The conditional probabilities $P(A_i \mid S_i, \theta)$ serve as control parameters for the experiment. By manipulating these probabilities for items with $P(S_i \mid \theta) > 0$, their expected exposure rates can be increased or decreased. To find the optimal values for these parameters at a set of $\theta$ values, we iteratively adjust them in a series of simulations of adaptive test administrations prior to the operational use of the test until they are below an upper limit, $r^{\max}$. The limit is set by the program administrator (Exercise 9.4).

If the test is operational, the Sympson-Hetter experiment is usually implemented as follows: Instead of selecting the most informative item at $\widehat{\theta}$, an ordered list with a number of the most informative items is identified. The control probabilities $P(A_i \mid S_i, \widehat{\theta})$ for the items are those at the $\theta$ values closest to $\widehat{\theta}$. The probabilities are renormed (i.e., their sum for the list is set equal to one), and one item is sampled from this list that conducts the multinomial experiment defined by the control probabilities. All items on the list more informative than the item sampled are removed from the pool, and all items less informative items than the sample are returned to the pool.

A longer list of items for the Sympson-Hetter experiment implies lower exposure rates for the more popular items in the pool. But if the list be-

comes too long, we lose too many good items during the test, and pay a price in the form of less accurate ability estimation. In practice, we compromise between these two tendencies. For the STA, a natural candidate for the list is the set of free items in the shadow test; they are the most informative items given all content constraints in the model. However, for a shorter adaptive test, this set may become too small toward the end of the test.

### *9.4.3 Multiple-Shadow-Test Approach*

An effective way to get a larger set of free items is to use a *multiple-shadow-test approach* (MSTA). In this approach, at each step, a set of parallel shadow tests is selected, and the experiment is conducted over the list of the most informative items assembled from the free items in these tests. At first sight, it may seem cumbersome to implement an MSTA, but the only thing required to do so is to adjust the model for the single shadow test in (9.11)–(9.21) to that for simultaneous selection of multiple tests.

As an example, we formulate the core of a model for an adaptive test with set-based items and omit possible content constraints:

$$\text{maximize } y \quad \text{(objective)} \tag{9.35}$$

subject to

$$\sum_{i=1}^{S}\sum_{i=1}^{I_s} I_{i_s}(\widehat{\theta}^{(g-1)})x_{i_s t} \geq y, \quad \text{for all } t, \quad \text{(test information)} \tag{9.36}$$

$$\sum_{t=1}^{T} x_{i_s t} \leq 1, \quad \text{for all } i_s \in R_g, \quad \text{(item overlap)} \tag{9.37}$$

$$\sum_{t=1}^{T} z_{st} \leq 1, \quad \text{for all } s \in R_l, \quad \text{(stimulus overlap)} \tag{9.38}$$

$$\sum_{i_s \in \overline{R}_g} x_{i_s t} = g - 1, \quad \text{for all } t, \quad \text{(previous items)} \tag{9.39}$$

$$\sum_{s \in \overline{R}_l} z_{st} = l - 1, \quad \text{for all } t, \quad \text{(previous stimuli)} \tag{9.40}$$

$$\sum_{i_s=1}^{I_s} x_{i_s t} \gtreqless n_s^{\text{set}} z_s, \quad \text{for all } s \text{ and } t, \quad \text{(number of items per set)} \tag{9.41}$$

$$x_{i_s t} \in \{0,1\}, \quad \text{for all } i, s \text{ and } t, \quad \text{(range of variables)} \tag{9.42}$$

$$z_{st} \in \{0,1\}, \quad \text{for all } s \text{ and } t. \quad \text{(range of variables)} \tag{9.43}$$

Observe that the decision variables $z_{st}$ and $x_{i_s t}$ are now for the assignment of stimulus $s$ and the items in its set $i_s = 1_s, ..., I_s$ to shadow test $t$,

respectively. Of course, for the MSTA to be effective, we do not want the free items and stimuli in the shadow tests to overlap; hence the constraints in (9.37) and (9.38). The constraints in (9.39) and (9.40) guarantee the presence of the previous items and stimuli in each of the shadow tests. The constraints in (9.41) do not only control the size of the item sets but also coordinate the values of the variables $z_{st}$ and $x_{i_s t}$. If the set sizes are not constrained, alternatively the constraint in (7.13) or (7.14) should be used to play this role (see the discussion in Section 7.1).

## Empirical Example

We introduced Sympson-Hetter exposure control in the 50-item adaptive version of the LSAT in Section 9.1.6 using the MSTA. The number of parallel shadow tests was equal to $T = 2$. At each step, the Sympson-Hetter experiment was conducted over all free items in the two shadow tests.

To set the values of the control parameters in (9.34), we conducted an iterative series of simulated administrations of the LSAT for test takers with abilities equal to $\theta = -2.0, -1.5, ..., 2.0$. For each $\theta$ value, the limit for the exposure rates was $r^{\max} = .25$. To get stable estimates of the exposure rates, 1,000 administrations were simulated for each $\theta$ value. The values for the control parameters were adjusted according to the standard procedure for the Sympson-Hetter method. (See the literature at the end of this chapter.) Even though we ran a series of 21 simulations for each $\theta$ value, we were unable to produce a set of values for the control parameters with all exposure rates below the limit. The best results were obtained in the 15th iteration; the values found in this iteration were used in the main study.

In the main study, CAT administrations both without and with Sympson-Hetter exposure control were simulated. For the case without control, we used the STA with a single shadow test, whereas for the case with control, we used the MSTA.

Figure 9.5 shows the distributions of the conditional exposure rates for the item pool at $\theta = -2.0, -1.5, \ldots, 2.0$ for the two cases. Without exposure control, the conditional exposure rates were much too high for a considerable number of items. The introduction of control reduced the rates to below .25 for nearly all of the items. The only exceptions were sets of 5–10 items for each of the $\theta$ values with exposure rate slightly above .25. The maximum rate of .29 among these items occurred only once. These exceptions were due to the fact that no entirely satisfactory set of values for the control parameters could be found, a result not uncommon for the Sympson-Hetter method.

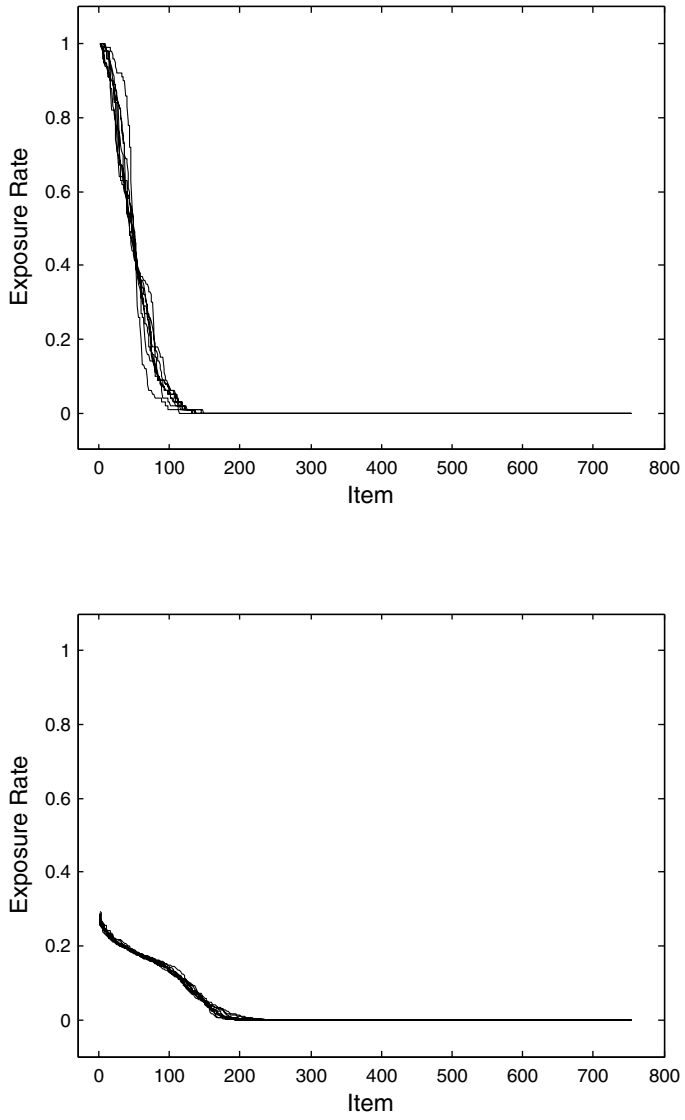


FIGURE 9.5. Conditional item-exposure rates at $\theta = -2.0, -1.5, \ldots, 2.0$ for adaptive testing without (top) and with (bottom) Sympson-Hetter exposure control. The items are ordered by exposure rate.

### *9.4.4 Method with Ineligibility Constraints*

Finding appropriate values for the control parameters in the Sympson-Hetter method is quite time-consuming. In the preceding example, 189 computer simulations were required to identify the set of values with acceptable exposure rates (21 simulations at nine $\theta$ values each). Also, every time the item pool is updated and/or the adaptive-testing algorithm changed, the control parameters have to be adjusted anew.

An alternative approach is based on the idea of eligibility decisions for the items in the pool. These decisions are based on probability experiments, too. But unlike the Sympson-Hetter method, the experiments are conducted only once for each test taker—before the test begins. If item $i$ is decided to be ineligible for the test taker, the following *ineligibility constraint* is added to the model for the shadow tests:

$$x_i = 0. \tag{9.44}$$

If an item is decided to be eligible, no constraint is added.

Obviously, the lower the probability of eligibility for an item, the lower its exposure rate. But if an item is eligible, it is not necessarily selected for administration. It is therefore not immediately clear what values the eligibility probabilities should have to constrain the exposure rates to a given range. Appropriate goal values for these probabilities can, however, be derived in the following way.

In addition to the event $A_i$ introduced for (9.34), we define event $E_i$ as item $i$ being eligible for a test taker. For an unfavorable combination of item pool and content constraints, adding a large number of ineligibility constraints could make the shadow-test model infeasible. We therefore define $F$ as the event of the shadow test remaining feasible after all eligibility constraints have been added to the model. The event of the model becoming infeasible is denoted as $\overline{F}$. If infeasibility happens, the model for the shadow test is solved with all ineligibility constraints removed. We discuss the likelihood of $\overline{F}$ and its consequences for the exposure rates of the items below.

An item can be administered only if it is eligible or the ineligibility constraints are removed from the model because it is infeasible. Thus, $P(E_i \cup \overline{F} \mid A_i, \theta) = 1$, and, analogous to (9.34),

$$P(A_i \mid \theta) = P(A_i \mid E_i \cup \overline{F}, \theta) P(E_i \cup \overline{F} \mid \theta). \tag{9.45}$$

It follows that the requirement that all expected exposure rates $P(A_i \mid \theta)$ be below a limit $r^{\max}$ is met if

$$P(E_i \cup \overline{F} \mid \theta) P(A_i \mid E_i \cup \overline{F}, \theta) \leq r^{\max} \tag{9.46}$$

for all $i$. This inequality does not impose any direct constraint on the probabilities of item eligibility, $P(E_i \mid \theta)$. But using probability calculus,

(9.46) can be shown to lead to the bound on these probabilities

$$P(E_i \mid \theta) \leq 1 - \frac{1}{P(F \mid \theta)} + \frac{r^{\max} P(E_i \cup \overline{F} \mid \theta)}{P(A_i \mid \theta) P(F \mid \theta)}, \tag{9.47}$$

with $P(A_i \mid \theta) > 0$ and $P(F \mid \theta) > 0$. (For a derivation, see Exercise 9.5.)

Suppose $e$ persons have taken the tests, and we want the items with a tendency to overexpose to meet the upper bound $r^{\max}$. The ineligibility constraints for test taker $e+1$ are then drawn with probabilities

$$P^{(e+1)}(E_i \mid \theta) = 1 - \frac{1}{P^{(e)}(F \mid \theta)} + \frac{r^{\max} P^{(e)}(E_i \cup \overline{F} \mid \theta)}{P^{(e)}(A_i \mid \theta) P^{(e)}(F \mid \theta)}, \tag{9.48}$$

with $P^{(e)}(F \mid \theta) > 0$ and $P^{(e)}(A_i \mid \theta) > 0$. The superscripts in these probabilities denote their status with respect to the succession of the test takers. Observe that all probabilities on the right-hand-side bound are for the preceding test taker $e$, while the eligibility probability on the left-hand side is for the new test taker $e+1$. Also, the inequality in (9.47) has been replaced by an equality.

The right-hand side of (9.48) can be easily estimated for an operational adaptive testing program using counts of the events $A_i$, $F$, and $E_i$. We recommend recording the counts conditional on the final ability estimates for the test takers, $\widehat{\theta}_n$. Since conditional exposure rates are robust with respect to small changes in $\theta$, the impact of the remaining estimation error in $\widehat{\theta}_n$ can be disregarded.

The bounds on the probabilities in (9.48) are *self-adaptive*. As soon as $P^{(e)}(A_i \mid \theta)$ becomes larger than $r^{\max}$ for a test taker in the program, probability $P^{(e+1)}(E_i \mid \theta)$ goes down, whereas if $P^{(e)}(A_i \mid \theta)$ becomes smaller than $r^{\max}$, it goes up again. (For a derivation, see Exercise 9.6.) This feature of self-adaptation permits us to apply this type of exposure control in an operational testing program without any previous adjustment of control parameters. The exposure rates of the items automatically adapt to optimal levels. The same holds if we have to change the item pool during operational testing; for instance, to remove some items with security breaches.

The only precaution that has to be taken is setting $P^{(e+1)}(E_i \mid \theta) = 1$ until a shadow test is found and the item has been administered once at the ability level. This measure is necessary to satisfy the conditions of $P^{(e)}(F \mid \theta) > 0$ and $P^{(e)}(A_i \mid \theta) > 0$ for (9.48).

For a well-designed item pool, the ineligibility constraints will hardly ever lead to infeasibility. If it does, the adaptive nature of (9.48) automatically corrects for the extra exposure of the items in the test due to the removal of the constraints.

### Empirical Example

The empirical example with the adaptive version of the LSAT in the preceding section was repeated with the Sympson-Hetter exposure-control method replaced by the method with random ineligibility constraints discussed in this section. All other aspects of the earlier study remained the same, except that we were a bit more ambitious and set the limit on the conditional exposure rates at $r^{\max} = .2$. We simulated 3,000 administrations for test takers at $\theta = -2.0, -1.5, ..., 2.0$ each.

The distributions of the conditional exposure rates for the cases without and with exposure control are shown in Figure 9.6. For the case with control, the exposure rates were effectively reduced to below .20. Due to the remaining estimation error in the right-hand-side probabilities of (9.48), a few items exceeded the limit by .01–.02, however.

We also recorded the number of times the model for the shadow test was feasible in this study. No cases of infeasibility were observed.

## 9.5 Controlling the Speededness of the Test

Test items differ as to the amount of time they require. An obvious factor with an impact on the time an item requires is its difficulty, but other factors, such as the amount of reading or encoding of symbolic information that is required, also play a role.

If the test is fixed, each test taker gets the same set of items. Consequently, the test is equally speeded for each of them. But if the test is adaptive, each test taker gets a different set, and the test easily becomes *differentially speeded*. Some of the test takers may then be able to complete their test early, whereas others run out of time. If this happens, the test can lose much of its validity.

One solution to the problem of differential speededness would be to choose the time limit for the test takers as a function of the time intensity of the items he or she gets. This strategy is technically possible as soon as we have good estimates of the time required by the items but is bound to run into practical problems, such as the impossibility of planning adaptive testing sessions efficiently.

Another solution would be to keep a common time limit for all test takers but to stop the test as a function of the time intensity of the items administered. This strategy leads to a different test length for different test takers and therefore has the disadvantage of less accurate ability estimates for some of them. It also becomes difficult for the test to satisfy a common set of content specifications for all test takers.

A better solution therefore is to *constrain* the test to become equally speeded for all test takers. This approach requires good estimates of the time intensity of the items. But if they have been pretested by seeding

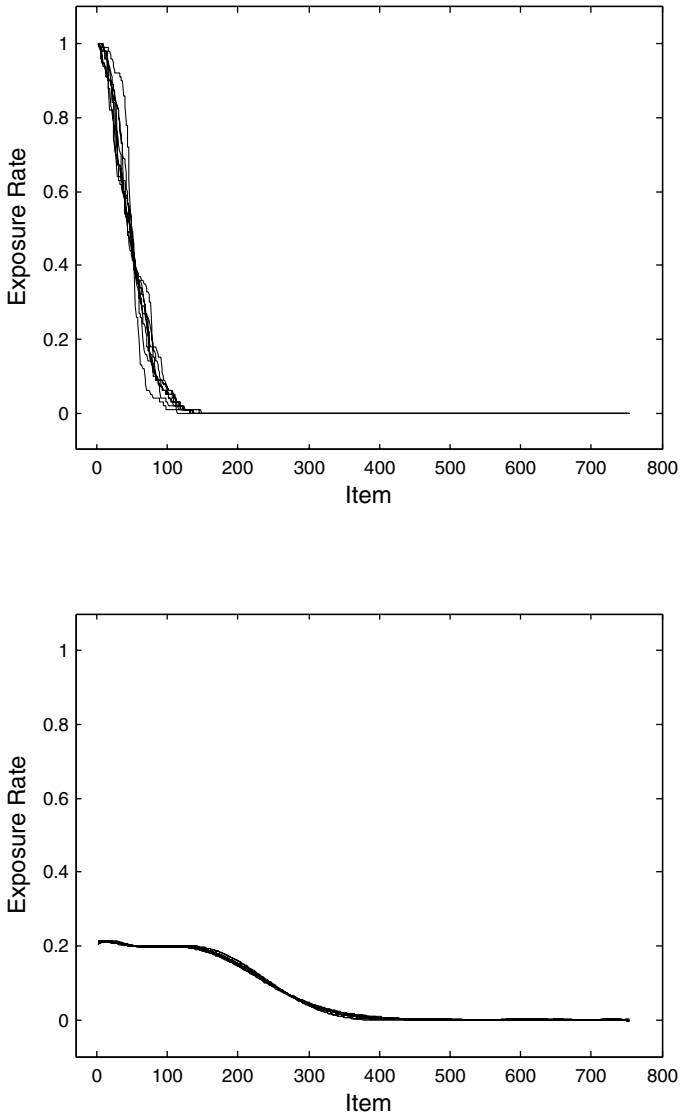


FIGURE 9.6. Conditional item-exposure rates at $\theta = -2.0, -1.5, \ldots, 2.0$ for adaptive testing without (top) and with (bottom) probabilistic ineligibility constraints. The items are ordered by exposure rate.

them into regular adaptive-test administrations, their response times have been automatically recorded, and estimates of the time parameters of the items can easily be obtained as part of regular item calibration. We discuss a lognormal model for the response times on test items that had excellent fit to a data set used in the empirical study later in this section and then show what constraints can be used to make a test equally speeded.

### *9.5.1 Response-Time Model*

The response-time model is for the distribution of random variable $T_i$ for the response time of a person operating at speed $\tau$ on item $i$. In addition to the person parameter, the model has two item parameters, one for the time intensity of item $i$, $\beta_i$, and the other for its discriminating power, $\alpha_i$.

The model has the density function

$$f(t_i;\tau,\alpha_i,\beta_i) = \frac{\alpha_i}{t_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_i - (\beta_i - \tau))\right]^2\right\}, \tag{9.49}$$

which is known as the lognormal density because it is identical to the density of a normal distribution for the logarithm of the response time. This model is not yet identified; we therefore propose the following constraint on the speed parameters for the persons $j = 1, ..., N$ in the data set:

$$\sum_{j=1}^{N} \tau_j = 0. \tag{9.50}$$

The speed parameter $\tau$ and time-intensity parameter $\beta_i$ are both on a scale from $-\infty$ to $\infty$, while the discrimination parameter takes values from 0 to $\infty$. Observe that the distribution of $\ln T_i$ has expected value $\beta_i - \tau$. Thus, the more time the item requires or the slower the person operates, the larger the expected logtime for the test taker on the item. The difference between $t_i$ and the expected value $\beta_i - \tau$ is modified by the discrimination parameter $\alpha_i$. The larger this parameter, the smaller the variance of the distribution, and the better the item discriminates between the response-time distributions of persons operating at a speed just above and below $\beta_i$.

Just as for the parameters in the IRT model, the estimates of the values of item parameters $\beta_i$ and $\alpha_i$ are treated as their true values during the operational use of the item pool. For details on this estimation procedure, see the literature at the end of this chapter.

We discuss two different types of control, one based on the features of the items only and another that also corrects for the individual speed of the test takers. Which type of control is appropriate depends on what the test was designed to measure. Following up on our discussion of the different cases of multidimensional testing in Section 8.1, the following two different cases are distinguished:

1. Both speed and ability are intentional factors measured by the test. In this case, it would be better to score the test with separate estimates of $\theta$ and $\tau$, but this is not yet current practice. Instead, typically a single score is supposed to reflect a combination of ability and speed. In this hybrid type of testing, the test has a time limit that is supposed to put the test takers under certain pressure. The limit has to be equally effective for all test takers. But since the test is adaptive, each test taker gets a different selection of items, and we have to control the selection of the items for their time intensity.

2. Ability is intentional but speed is a nuisance factor. In this case, the test is in fact a power test, and the scores can be interpreted as an estimate of $\theta$ only. The test may have a time limit for practical reasons, but we do not want the limit to be effective, let alone put a different amount of pressure on different test takers. In this case, item selection has to be controlled both for the test taker's speed and the time intensity of the items to prevent the test taker from running out of time.

### *9.5.2 Ability and Speed as Intentional Factors*

If the calibration sample is large and can be viewed as a random sample from the population of test takers, the identifiability constraint in (9.50) sets the average speed of the population of test takers equal to zero. As a consequence, the time-intensity parameter $\beta_i$ becomes the expected logtime on item $i$ for an average test taker in the population.

If the total time available for the test, $t_{\text{tot}}$, is well-chosen for the majority of the test takers, item selection can be controlled for differences in time intensity between the items by including the following constraint in the model for the shadow test:

$$\sum_{i=1}^{I} \exp(\beta_i) x_i \leq t_{\text{tot}}, \tag{9.51}$$

where the exponential transformation is needed because $\beta_i$ is on the logtime scale.

We assume that (9.15) is present in the model for the shadow test so that the left-hand sum of this constraint automatically contains the values of $\beta_i$ for the items already administered.

The constraint in (9.51) makes item selection equally speeded for all test takers at the level of the total test. If stricter control is required, in the sense of more homogeneous speededness throughout the test, we can also impose the constraint at the level of subsequent blocks of items in the test.

The constraint in (9.51) ignores the differences in the dispersion of the total time on the test between different test takers. For an adaptive test of

regular length, this is not much of a problem. Because the total time is the *sum* of the response times on the items, the standard deviation of the total time quickly decreases with increasing length of the test (Exercise 9.8).

### *9.5.3 Speed as a Nuisance Factor*

If speed is a nuisance factor according to the definition of the test, we should estimate the test takers' speed during the test and use these estimates as well to create unspeeded tests for all test takers.

Suppose the response times on the first $g-1$ items have already been recorded, and we are to select the $g$th item in the test. The maximum-likelihood estimate of the test taker's speed parameter $\tau$ after $g-1$ items can be shown to be equal to

$$\widehat{\tau}^{(g-1)} = \sum_{i \in \overline{R}_g} (\beta_i - \ln t_i)/(g-1). \tag{9.52}$$

The estimate is used in the following constraint on the test:

$$\sum_{i=1}^{I} \exp(\beta_i - \widehat{\tau}^{(g-1)}) x_i \leq t_{\text{tot}}. \tag{9.53}$$

This constraint thus controls the expected total time of the test takers on the test, no matter how fast they work.

A more sophisticated version of (9.53) is possible if we update our estimates of $\tau$ in a Bayesian fashion (i.e., as the posterior distribution of $\tau$ given the response times on the previous items, $t_{i_1}, ..., t_{i_{g-1}}$). The posterior distribution of $\tau$ can be used to predict the time distribution for each of the remaining items in the pool for the test taker, which are known to also be lognormal for the model in (9.49). (For a derivation, see the literature section at the end of this chapter.) Let $f(t_i \mid t_{i_1}, ..., t_{i_{g-1}})$ denote the density of the predicted response-time distribution for item $i$ in the remaining set of items in the pool, $R_g$. It is easy to calculate the $\pi$th percentiles of these distributions, which we denote as $t_i^{\pi_g}$. The following constraint on the shadow tests should be used:

$$\sum_{i \in \overline{R}_g} t_i + \sum_{i \in R_g} t_i^{\pi_g} x_i \leq t_{\text{tot}}. \tag{9.54}$$

This constraint restricts the sum of the actual response times on the first $g-1$ items and $\pi_g$th percentiles of the predicted response-time distributions on the $n-g$ free items in the shadow test by the total amount of time available for the test. It makes sense to specify a sequence of percentiles for the successive items that increases toward the end of the test. Such a sequence makes the predictions more conservative when less time is left.

Just as discussed for the constraint in (9.51) on the item parameters only, (9.53) and (9.54) can also be imposed at the level of subsequent blocks of items to realize a more homogeneous level of unspeededness throughout the test.

In principle, for this second type of control, if test takers understand the procedure, they may try manipulating the choice of items; for example, by working more slowly initially than they need to and speeding up later in the test, when the algorithm selects items that require less time. If the item parameters in the IRT model do not correlate with the parameters in the response-time model, such strategies have no impact on the ability estimates. If these parameters do correlate, they can only have an impact on the accuracy of the estimates of $\theta$ because of less than optimal item selection. But if $\theta$ is estimated statistically correctly, such strategies cannot introduce any bias in the estimates that would be advantageous to the test taker.

## Empirical Example

A simulation study was conducted in which the Bayesian constraint in (9.54) was used to control the speededness of the test. We used the adaptive version of the Arithmetic Reasoning Test from the Armed Services Vocational Aptitude Battery (ASVAB). The item pool was a previous pool of 186 items for this test calibrated using the 3PL model in (1.16). The test has a length of 15 items and is administered without any content constraints. An initial analysis of the actual response times in the data set showed that the time limit of 39 minutes (2,340 seconds) for the test was rather generous and the test thus could be viewed as a power test. This impression justifies the use of (9.54) to avoid differences in speededness between test takers due to the adaptive nature of the test.

The response-time model we used was the lognormal model with a different parameterization, which involved the constraint $\alpha_i = \alpha$ on the discrimination parameter for all items in the pool. The model showed a satisfactory fit to the response times of a sample of $N = 38{,}357$ test takers.

The items were selected to have maximum information at the interim estimates of $\dot{\theta}$. The first item was selected to be optimal at $\widehat{\theta}^{(0)} = 0$. Interim estimates during the test were obtained using expected a posteriori (EAP) estimation with a uniform prior distribution for the first estimate. The percentile chosen for the posterior predicted response-time distributions in (9.54) was the 50th percentile for the selection of the first item, but we moved linearly to the 95th percentile for the selection of the last two items in the test. The prior distribution of $\tau$ was a normal distribution with mean and variance equated to the mean and variance of the distribution of $\tau$ estimated from the sample of persons in the data set. After each item, the test taker's response time was used to update this prior distribution in a Bayesian fashion.

We simulated test takers with abilities $\theta = -2.0, 1.5, ..., 2.0$ and speed $\tau = -.6, -.3, 0, .3$, and .6. This range of values for the speed parameter covered the estimated values for the sample of test takers for the version of the response-time model used in this study.

We first simulated test administrations without the response-time constraint. The average times used by the test takers at the different ability and speed levels are reported in the first panel of Figure 9.7. The major impression from the results is that the majority of the test takers have ample time to complete the test, which therefore can be viewed as a power test. Of course, the slower test takers used more time, but this finding does not necessarily point to differential speededness of the test but only to individual differences between test takers. However, the fact that a minority of the test takers ran out of time does reveal that the test was differentially speeded.

We then repeated the study with the response-time constraint in (9.54) added to the shadow-test model. The results are given in the second panel of Figure 9.7. This time, all test takers were able to meet the limit of 2,340 seconds. Also, the amount of time they used was distributed much more uniformly over $\tau$. These results were obtained in spite of the fact that the test takers operated at exactly the same levels of speed as in the first study!

To stretch the limit, we also repeated the study with a time limit equal to $t_{\text{tot}} = 34$ minutes (2,040 seconds) and then $t_{\text{tot}} = 29$ minutes (1,740 seconds). The third and fourth panels in Figure 9.7 show the time used by the test takers for these limits. Again, the response-time constraint appeared to be effective. In spite of the decrease in time available, the test takers had no difficulty completing the test in time, and the distributions of the times they needed to complete the test were more uniform still.

It may be surprising to note that the test takers who ran out of time in the unconstrained version of the test were those with a low value for $\tau$ but a *high* value for $\theta$. (See the upper left panel of Figure 9.7.) This result can be explained by the correlation of .65 between the item-difficulty parameter $b_i$ and the time-intensity parameter $\beta_i$ we found for the data set. As a result of this correlation, adaptive selection of the items for the more able test takers led not only to items that were more difficult toward the end of the test but also to items that were more time-intensive. Consequently, the slower test takers among them ran out of time.

## 9.6 Reporting Scores on a Reference Test

To enhance the interpretation of test scores, testing agencies usually give the test taker a sample of the test items. In a program with a group-based fixed test administered only once, the test itself can be released for this purpose, but this practice is not possible for adaptive tests. The

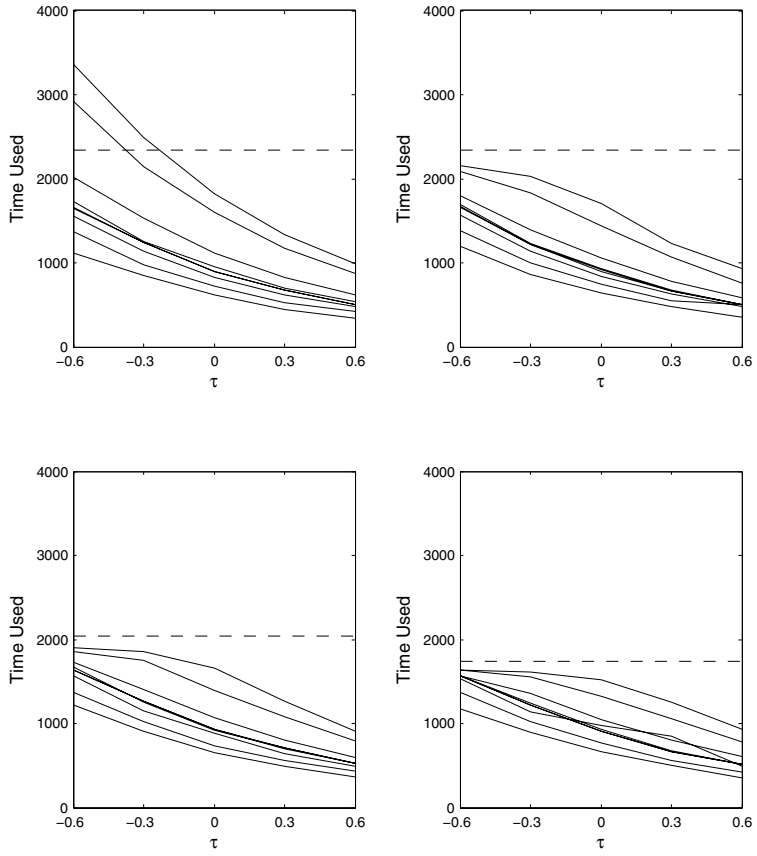


FIGURE 9.7. Average time used by examinees of different levels of ability $\theta$ and speed $\tau$ for the cases without the time constraint (top left) and with the time constraint and a time limit of 2,340 seconds (top right), 2,040 seconds (bottom left), and 1,740 seconds (bottom right). All curves are ordered by $\theta$, with $\theta = 2.0$ for the highest curves and $\theta = -2.0$ for the lowest curves. The dashed lines indicate the time limits.

testing agency would then have to replenish the item pool too frequently—an activity that not only would entail prohibitively large costs but could also be expected to lead to a quick deterioration of the quality of the pool.

A standard solution for an adaptive-testing program is to give the test takers a paper-and-pencil test with items representative of the pool as a reference test. If the test takers' ability estimates, $\widehat{\theta}$, have been equated to

the number-correct scores on the reference test, $X$, the reference test can be used to evaluate their performances on the adaptive test.

Typical equating methods used to transform $\widehat{\theta}$ to a number-correct score on the reference test are traditional equipercentile equating or true-score equating using the test characteristic function $\tau(\theta)$ in (1.22) for the reference test. If equipercentile equating is used, both the adaptive test and the reference test have to be administered to random samples from the population of test takers, and estimates of the population distributions of $\widehat{\theta}$ and $X$ need to be used to find the score transformation. This equating study is not necessary if the test characteristic function of the reference test, $\tau(\theta)$, is used. This function follows directly from the response functions of the items in the reference test, and substitution of the final estimate $\widehat{\theta}$ on the adaptive test into $\tau(\theta)$ gives an estimate of the test taker's true number-correct score on the reference test.

Both score transformations can be shown to be seriously biased in the presence of measurement error on the reference test and adaptive test. The reason for this bias is the use of a common transformation for all test takers, which has to compromise between their different error distributions along the scales of the two tests. An alternative is to use a local transformation at the position of the test taker on this scale. The discussion of such techniques is beyond the scope of this book. But an alternative that does fit the subject of this book can be derived from the constraints for matching observed-score distributions in Section 5.3.

When applied to adaptive testing, we impose the condition on the sums of powers of the response probabilities in (5.47) on the shadow tests. Let $j = 1, ..., n$ denote the items in the reference test. Suppose we are to assemble the shadow test for the selection of the $g$th item in the adaptive test. Analogous to (5.48), the target values are now the known constants

$$\mathcal{T}_{r(g-1)} = \sum_{j=1}^{n} p_j^r(\widehat{\theta}^{(g-1)}), \quad \text{for } r \leq R, \tag{9.55}$$

where $R$ is the condition in (5.47) with the highest order used. These target values have to be met by the shadow test; that is, by

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)})x_i, \quad \text{for } r \leq R. \tag{9.56}$$

These target values are enforced by imposing the following constraints on the shadow tests:

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)})x_i \leq \mathcal{T}_{r(g-1)} + \delta, \quad \text{for all } r \leq R, \tag{9.57}$$

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)})x_i \geq \mathcal{T}_{r(g-1)} - \delta, \quad \text{for all } r \leq R, \tag{9.58}$$

where $\delta$ is a small tolerance needed to avoid infeasibility due to an equality constraint in the model; see the discussion of (3.50) in Section 3.2.4.

A critical difference with the constraints for the case of a fixed test in Section 5.3 is that (9.57) and (9.58) control the conditions in (5.47) only at the current estimate $\widehat{\theta}_{g-1}$, while in (5.51) and (5.52) they are controlled at a fixed series of values $\theta_k$, $k = 1, ..., K$. In the example for a fixed test in Section 5.3.4, we saw that the best results were obtained for the cases with control at two or three $\theta$ values. We therefore recommend adding constraints at the values $\widehat{\theta}_{g-1} + \varepsilon$ and $\widehat{\theta}_{g-1} - \varepsilon$, where $\varepsilon$ is a constant chosen by the test administrator. The target values for the reference test at these points, $\mathcal{T}^{+\varepsilon}_{r(g-1)}$ and $\mathcal{T}^{-\varepsilon}_{r(g-1)}$, are calculated analogously to (9.55).

These target values are enforced by the following additional constraints:

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)} + \varepsilon)x_i \leq \mathcal{T}^{+\varepsilon}_{r(g-1)} + \delta, \quad \text{for all } r \leq R, \tag{9.59}$$

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)} + \varepsilon)x_i \geq \mathcal{T}^{+\varepsilon}_{r(g-1)} - \delta, \quad \text{for all } r \leq R, \tag{9.60}$$

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)} - \varepsilon)x_i \leq \mathcal{T}^{-\varepsilon}_{r(g-1)} + \delta, \quad \text{for all } r \leq R, \tag{9.61}$$

$$\sum_{i=1}^{I} p_i^r(\widehat{\theta}^{(g-1)} - \varepsilon)x_i \geq \mathcal{T}^{-\varepsilon}_{r(g-1)} - \delta, \quad \text{for all } r \leq R. \tag{9.62}$$

Unlike the fixed set of constraints in (5.51) and (5.52), the current set is dynamic. It imposes the conditions in (5.47) at a window about $\widehat{\theta}^{(g-1)}$ of size $2\varepsilon$, which stabilizes if $\widehat{\theta}^{(g-1)}$ does.

We do not yet expect this application to automatically work as well as the one for the fixed test in Section 5.3. One reason is that the best combination of choices of values for $\delta$ and $\varepsilon$ still has to be determined. Another is that an unfortunate initialization of $\widehat{\theta}^{(0)}$ may lead to an initial series of wild estimates with response functions of the selected items that are hard to compensate for later in the test, when $\widehat{\theta}$ is close to the test taker's true $\theta$. In fact, it may well be that control at a fixed series of well-chosen $\theta$ values works equally well or even better than (9.57)–(9.62) for an adaptive test. Both points require further research.

This research is worth the effort because automatic equating of an adaptive test to a reference test entails two major advantages. First, the equating is local; that is, conditional on the final ability estimate of the test taker. As indicated in Section 5.3.3, this type of equating is strong and satisfies the important requirement of equitability: It yields equated scores with the same error distribution as the test taker would have had on the reference test. Second, by imposing the conditions in (5.47) on an adaptive test, the test automatically produces the same *observed number-correct scores* as the

reference test. No additional score transformation is needed. The test taker can interpret his or her number-correct score on the adaptive test directly as the number of items he or she would have had correct on the reference test. More amazingly, because the adaptive test is equated to the *same* reference test for all test takers, their scores are also equated mutually: Different test takers can directly compare their numbers of items correct even though they received a different selection of items from the pool.

## Empirical Example

The set of constraints in (9.57)–(9.62) was used in a simulation study with adaptive-test administrations from the same item pool for the LSAT used in the examples in the preceding sections. We first assembled a fixed test from the pool and then randomly took a set of nested reference tests from it, with test lengths of $n = 10$, 20, 30, and 40 items; that is, the reference test of 10 items was nested in the test of 20 items, and so on.

No constraints were imposed on the adaptive test. The tolerance parameter $\delta$ was set at .5. This choice is equivalent to a half score point on the number-correct scale. During a tryout of this study, we discovered a few cases of infeasibility later in the adaptive test due to first estimates $\widehat{\theta}$ that were far off. If this happened, the algorithm increased the value of $\delta$ by .2. The parameter $\varepsilon$ was always kept at .5. The target values in the constraints in (9.59)–(9.62) were thus calculated at $\widehat{\theta}^{(g-1)} - .5$, $\widehat{\theta}^{(g-1)}$, and $\widehat{\theta}^{(g-1)} + .5$. Three different replications of the study were conducted, in which the constraints in (9.57)–(9.62) were imposed for (i) $r = 1$, (ii) $r = 1$ and 2, and (iii) $r = 1$, 2, and 3.

We used a Bayesian initialization of $\widehat{\theta}$, where $\widehat{\theta}^{(0)}$ was calculated using regression on a background variable that was assumed to correlate .60 with $\theta$. The prior distribution was also provided by the correlation with the background variable. Subsequent estimates of $\theta$ were expected a posteriori (EAP) estimates. (For details on this procedure, see the literature section at the end of this chapter.)

Adaptive test administrations were simulated for 30,000 test takers randomly sampled from a standard normal distribution for $\theta$. For each simulated test taker, we recorded the number-correct score after $n = 10$, 20, 30, and 40 items. The distributions of these scores for the 30,000 test takers are shown in Figures 9.8–9.10. The observed-score distributions on the reference tests were calculated using the Lord-Wingersky algorithm with a standard normal distribution for $\theta$. (For this algorithm, see the literature section at the end of this chapter.)

The results for the set of constraints with $r = 1$, 2 and $r = 1$, 2, 3 were quite close. Also, they were generally better than those for $r = 1$ only. For the lengths of 30 and 40 items, the number-correct score distributions on the adaptive test matched those on the reference test reasonably well.

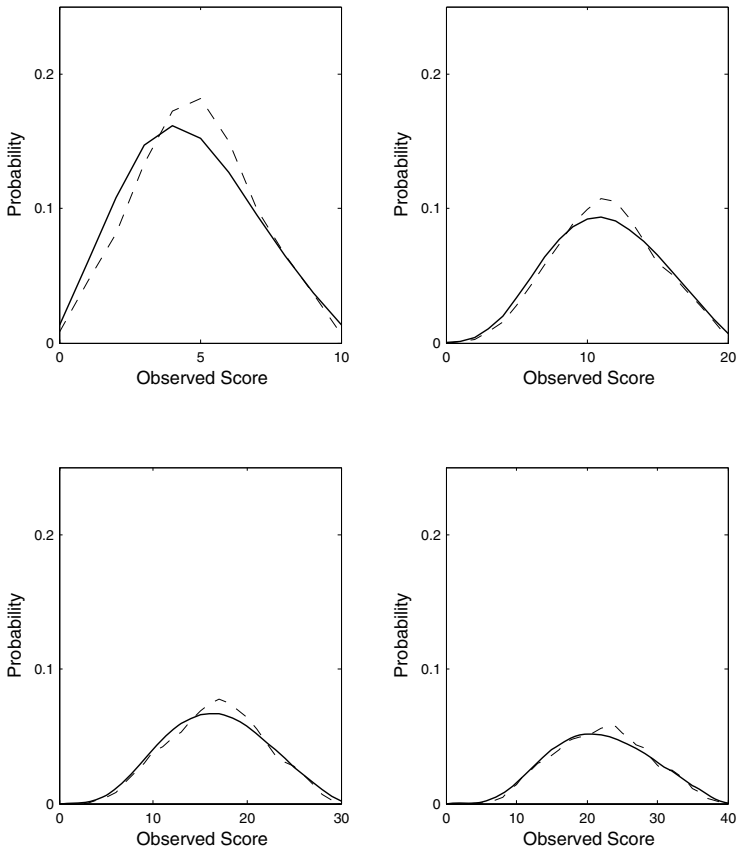


FIGURE 9.8. Observed-score distributions on an adaptive test (dashed lines) and a reference test (solid line) for $n = 10$ and 20 (top) and $n = 30$ and 40 (bottom) ($r = 1$).

For the shorter test lengths, there was still a mismatch. The question of whether the results for the test lengths of 30 and 40 items, which are typical lengths for real-world adaptive tests, are good enough for score-reporting purposes is subject to further research. This research should involve other choices of $\delta$ and $\varepsilon$ as well as the alternative of permanent control at a fixed series of values $\theta_k$, $k = 1, ..., K$. Also, the bias and accuracy of this equating method should be evaluated against the alternatives based on the test characteristic function and traditional equipercentile equating.

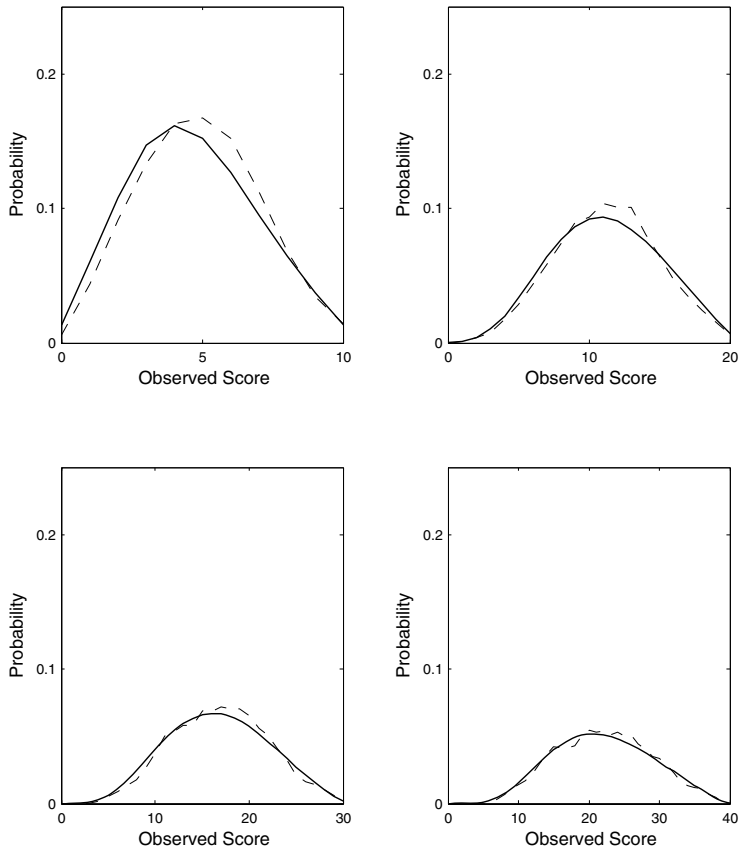


FIGURE 9.9. Observed-score distributions on an adaptive test (dashed lines) and a reference test (solid line) for $n = 10$ and 20 (top) and $n = 30$ and 40 (bottom) $(r = 1, 2)$.

In this study, we also checked the impact of the constraints on the statistical quality of the final ability estimator for the test. There was no discernible loss of accuracy over the main portion of the $\theta$ scale. At the extremes of the scale, there was some loss, equivalent to only a few test items. This loss could thus easily be compensated for by making the test slightly longer.

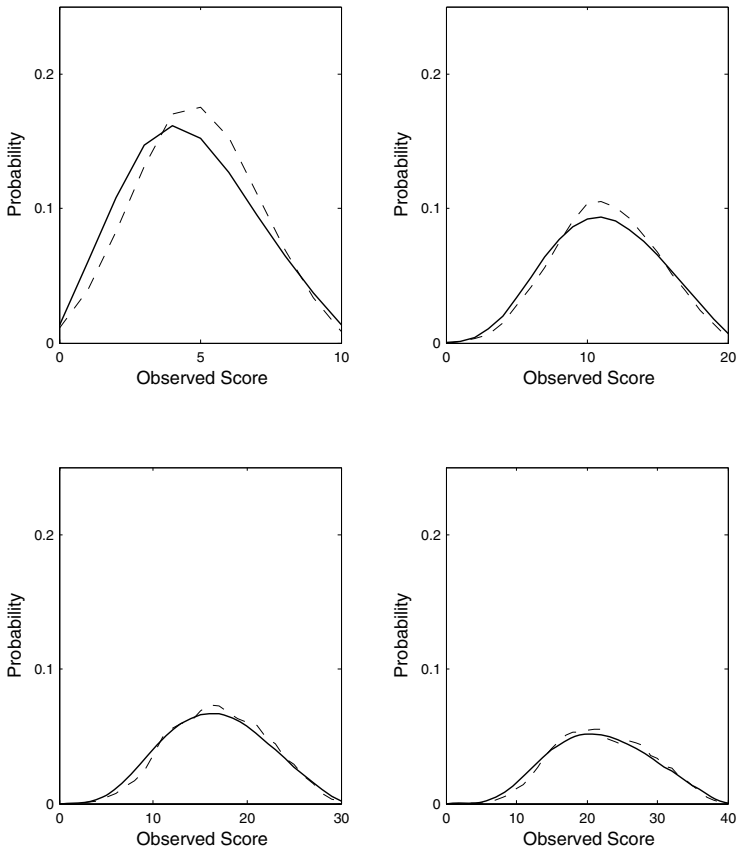


FIGURE 9.10. Observed-score distributions on an adaptive test (dashed lines) and a reference test (solid line) for $n = 10$ and 20 (top) and $n = 30$ and 40 (bottom) ($r = 1, 2, 3$).

## 9.7 Multidimensional Adaptive Test Assembly

### *9.7.1 Minimizing Error Variances*

If the item pool is multidimensional, the model for the shadow tests can be derived from the standard model for the assembly of a fixed multidimensional test in (8.14)–(8.20). In Section 8.4.3, we showed how to specialize this model to different cases of multidimensional test assembly by imposing certain restrictions on the sets of weights in the key constraints in (8.16)

and (8.17). These cases included multidimensional testing in which all abilities measured by the items are intentional, cases where only some of them are intentional and the others are a nuisance, and cases of testing with an interest in a linear combination of the abilities.

The modification of the model in (8.14)–(8.20) that is required for a multidimensional model for shadow tests in adaptive testing is straightforward. As before, we formulate the model for the selection of the $g$th item in the adaptive test. The updates of the estimates of the two ability parameters after the first $g-1$ items are denoted as $\widehat{\theta}_1^{(g-1)}$ and $\widehat{\theta}_2^{(g-1)}$. For brevity, we denote the probabilities of correct and incorrect responses under the two-dimensional model in (1.17) at these estimates as

$$p_i^{(g-1)} = p_i(\widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}), \tag{9.63}$$

$$q_i^{(g-1)} = 1 - p_i(\widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}).$$

The core of the shadow-test version of the model in (8.14)–(8.20) is

$$\text{minimize } y \tag{9.64}$$

subject to

$$\sum_{i=1}^{I} a_{1i} a_{2i} p_i^{(g-1)} q_i^{(g-1)} x_i \leq y, \tag{9.65}$$

$$\sum_{i=1}^{I} a_{1i}^2 p_i^{(g-1)} q_i^{(g-1)} x_i \geq w_1 \kappa, \tag{9.66}$$

$$\sum_{i=1}^{I} a_{2i}^2 p_i^{(g-1)} q_i^{(g-1)} x_i \geq w_2 \kappa, \tag{9.67}$$

$$\sum_{i=1}^{n} x_i = n, \tag{9.68}$$

$$\sum_{i \in \overline{R}_g} x_i = g - 1, \tag{9.69}$$

$$x_i \in \{0, 1\}, \quad \text{for all } i, \tag{9.70}$$

$$y \geq 0. \tag{9.71}$$

If the item pool contains set-based items, the constraint in (9.29) has to be added to this model.

Observe that we no longer control two variance functions over an entire grid of points in the ability space, that is, at points $(\theta_{1kl}, \theta_{2kl})$, with $k = 1, ..., K$ and $l = 1, ..., L$, as we did for the assembly of a fixed test in Chapter

8, but only control the variances of the ability estimators at $(\theta_1, \theta_2) = (\widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)})$. For this reason, the original sets of weights in (8.16) and (8.17) specialize to the two relative weights $w_1$ and $w_2$ in (9.66) and (9.67).

If both $\theta_1$ and $\theta_2$ are intentional, the weights should be chosen to reflect the relative importance of minimization of the variance of $\widehat{\theta}_1$ relative to the variance of $\widehat{\theta}_2$. If both objectives are equally important, we should choose

$$w_1 = w_2 = 1, \tag{9.72}$$

which means that the weights can be removed from (9.66) and (9.67).

If $\theta_2$ is a nuisance ability, we should choose

$$w_1 = 1, \tag{9.73}$$

$$w_2 = 0. \tag{9.74}$$

This choice puts all weight on the constraint in (9.66) and makes (9.67) redundant.

The criterion for picking the best item from the free items in the shadow test follows directly from the definition of the variances of estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ in (8.5) and (8.6). To evaluate the candidacy of item $i$, we calculate projections for these variances after administration of item $i$ at the current estimate $(\theta_1, \theta_2) = (\widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)})$. We propose using the following projections derived from (8.5) and (8.6):

$$\mathrm{Var}(\widehat{\theta}_1^{(g-1+i)} \mid \widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}) = \frac{A_2}{A_1 A_2 - A_{12}}, \tag{9.75}$$

$$\mathrm{Var}(\widehat{\theta}_2^{(g-1+i)} \mid \widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}) = \frac{A_1}{A_1 A_2 - A_{12}}, \tag{9.76}$$

with

$$A_1 = \sum_{j \in \overline{R}_g \cup \{i\}} a_{1j}^2 p_j^{(g-1)} q_j^{(g-1)}, \tag{9.77}$$

$$A_2 = \sum_{j \in \overline{R}_g \cup \{i\}} a_{2j}^2 p_j^{(g-1)} q_j^{(g-1)}, \tag{9.78}$$

$$A_{12} = \sum_{j \in \overline{R}_g \cup \{i\}} a_{1j} a_{2j} p_j^{(g-1)} q_j^{(g-1)}, \tag{9.79}$$

where $\overline{R}_g$ still denotes the set of $g-1$ items already administered.

The item with the minimal value for

$$w_1 \mathrm{Var}(\widehat{\theta}_1^{(g-1+i)} \mid \widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}) + w_2 \mathrm{Var}(\widehat{\theta}_2^{(g-1+i)} \mid \widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}) \tag{9.80}$$

is the best item for administration. The values of the weights $w_1$ and $w_2$ in this sum are subject to the same restrictions as in (9.72) or (9.73) and (9.74).

The expression (9.80) represents a weighted version of the criterion of $A$-optimality discussed in Section 8.4.1 for the case of a fixed test. Alternatively, we could use the criterion of $D$-optimality and evaluate projections of the determinant of the covariance matrix in (8.3) for each of the free items in the shadow test (Exercise 9.10).

### *9.7.2 Computational Aspects*

Although the criterion for the second-stage selection of the item in (9.80) can easily be calculated in real time for the set of free items in the shadow test, the computation time required to calculate the shadow tests is an as yet unexplored aspect of multidimensional adaptive testing. As discussed in Section 8.4.1, the model in (9.63)–(9.71) should be solved for a sequence of increments of $\kappa$, and the best solution according to (9.80) should be picked. The empirical example with a fixed test in Section 8.6 illustrated this approach. With adaptive test assembly, the computations have to be performed in real time.

It is possible to optimize these computations along the same lines as in Section 9.1.5. If we start the constraints in (9.66) and (9.67) with a large value for $\kappa$ and relax by small decreases of $\kappa$, the preceding solution is always in the feasible space of the next problem and is an attractive initial solution for the next shadow test. However, except for the case of adaptive testing for a linear combination of abilities (see the literature at the end of this chapter), we do not yet have much experience with this type of multidimensional adaptive-test assembly, and our computational ideas still have to be tried out empirically.

### *9.7.3 Maximizing Kullback-Leibler Information*

An alternative approach to multidimensional adaptive testing is to choose an objective function for the shadow-test model based on the multivariate version of the posterior expected Kullback-Leibler information in (9.26). As already observed in Section 9.2.1, the generalization of the Kullback-Leibler measure to more than one dimension only involves the replacement of the unidimensional response model in (9.24) by the two-dimensional model in (1.17). Likewise, in (9.25) we have to integrate the multidimensional version of this measure over the joint posterior distribution of $(\theta_1, \theta_1)$ given the previous responses $(u_{i_1}, ..., u_{i_{g-1}})$.

To maximize Kullback-Leibler information in multidimensional adaptive testing, the objective function in the standard model for the shadow test in (9.11)–(9.21) has to be replaced by

$$\text{maximize} \sum_{i=1}^{I} K_i(\widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)}) x_i. \tag{9.81}$$

The only price we pay for using this objective function is that the different cases of multidimensional test assembly introduced in Section 8.4.3 can no longer be addressed with explicit weights for the ability dimensions, as we were able to do for (9.66), (9.67), and (9.80). But alternative ways to address these cases do exist.

We discuss the following three cases:

1. If $\theta_1$ and $\theta_2$ are both intentional abilities, the free item in the shadow test with the largest contribution to (9.81) should be selected. This criterion leads to the selection of items that are most informative in the sense of giving the best discrimination between the current estimates $(\widehat{\theta}_1^{(g-1)}, \widehat{\theta}_2^{(g-1)})$ and the points in the two-dimensional ability space at which the posterior distribution is concentrated.

2. If $\theta_1$ is intentional but $\theta_2$ is a nuisance ability, we recommend the following change for the objective function in (9.81): (i) Substitute the current estimate $\widehat{\theta}_2^{(g-1)}$ for $\theta_2$ in the denominator of the Kullback-Leibler information in (9.24) and (ii) perform the integration in (9.25) over the marginal posterior distribution of $\theta_1$ given $(u_{i_1}, ..., u_{i_{g-1}})$. As a result of this change, the items in the shadow test become most informative with respect to the variation in the two-dimensional ability space in the direction along the $\theta_1$ axis, while the variation along the $\theta_2$ axis is ignored.

3. If one is interested in a linear combination $\lambda\theta_1 + (1-\lambda)\theta_2$, with $0 < \lambda < 1$, following our earlier suggestion in Section 8.4.3, we recommend rotating the ability space such that this combination coincides with the first dimension in the new space. The case then becomes identical to the preceding case.

### *9.7.4 Empirical Examples*

We simulated administrations of a 50-item adaptive test for the three different cases of multidimensionality in the preceding section using the same pool of 176 mathematics items for the ACT Assessment Program as in the empirical example with a fixed multidimensional test in Section 8.6. For each case, the items were selected using the modification of the Kullback-Leibler information criterion in (9.81) above. Otherwise, the model for the shadow test was entirely identical to the model for the fixed test in the example in Section 8.6. The test was started with $(\widehat{\theta}_1^{(0)}, \widehat{\theta}_2^{(0)}) = (0, 0)$ for each simulated test taker. The ability estimates were updated using joint EAP estimation.

The case where we are interested in a linear combination of abilities had equal weights $\lambda = .5$. That is, we used $\xi_1 = .5\theta_1 + .5\theta_2$ as the intentional

ability and $\xi_2 = -.5\theta_1 + .5\theta_2$ as the nuisance ability. The case was implemented using the transformations $a_1^* = a_1 + a_2$ and $a_2^* = a_2 - a_1$ for the two discrimination parameters. (See Exercise 8.3 for details.)

The test administrations were simulated at the 441 different combinations of $\theta_1, \theta_2 = -2.0, -1.8, ..., 2.0$. At each combination, we calculated the average variances in (8.5) and (8.6) for the final estimates $\widehat{\theta}_1$ and $\widehat{\theta}_s$ across the replications. The number of replications at each different combination was equal to 25. The plots with average variances were smoothed using a standard procedure in *Matlab*.

The results for the case where $\theta_1$ and $\theta_2$ are both intentional are given in Figure 9.11. As expected, the two plots show surfaces at about the same height. The fact that the test was adaptive led to an improvement of the surfaces relative to those for the fixed test where $\theta_1$ and $\theta_2$ are intentional in Figure 8.2. The fact that the improvement was modest is due to the unfavorable size of the item pool in relation to the test length.

Figure 9.12 shows the estimates of the two variance functions for the case where $\theta_1$ is the only intentional ability. Since the item selection no longer had to compromise between the variances of the estimators of two intentional ability parameters, the surface for $\widehat{\theta}_1$ was more favorable than in Figure 9.11, while, as expected, the surface for $\widehat{\theta}_2$ was less favorable.

An excellent variance function was obtained for the linear combination of interest, $\xi_1$, in the third case. As Figure 9.13 shows, the surface was not only low but also nearly uniform over the entire space of combinations of values of $\xi_1$ and $\xi_2$ studied. Apparently, the items in this pool were written and selected with this kind of combination of abilities in mind. Of course, the large improvement in the variance of the estimator of $\xi_1$ in Figure 9.13 relative to that of $\theta_1$ in Figure 9.12 was obtained at the price of a much deteriorated variance for the combination $\xi$. But the latter did not interest us.

## 9.8 Final Comments

So far, we have only discussed adaptive test assembly with adaptation at the level of individual items. Alternative options are testlet-based adaptive testing, multistage adaptive testing, and adaptive linear on-the-fly testing. These alternatives have a decreasing level of adaptation: In testlet-based adaptive testing, the ability estimate is updated after testlets of 3–5 items, in multistage adaptive testing after longer subtests, and in linear on-the-fly testing the possibility of adaptation arises only if the individual tests given to test takers are assembled to be optimal at an *a priori* estimate of their ability derived from background information; for instance, an earlier test score.

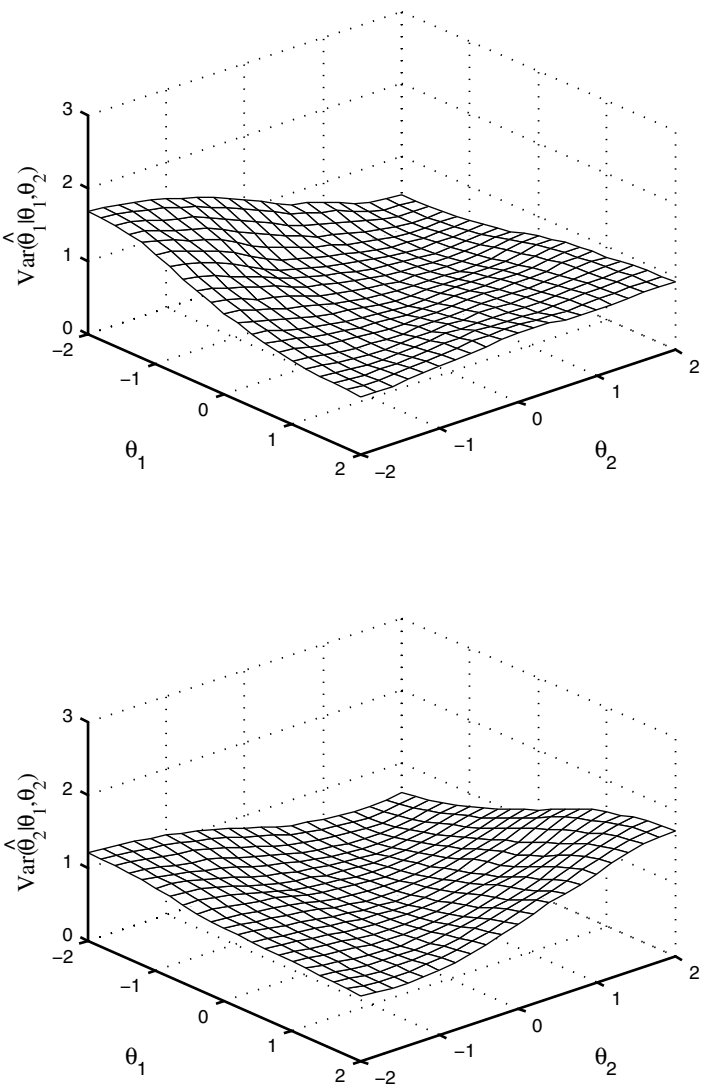


FIGURE 9.11. Estimated variance functions for $\widehat{\theta}_1$ (top) and $\widehat{\theta}_2$ (bottom) for a two-dimensional test with both $\theta_1$ and $\theta_2$ intentional abilities.

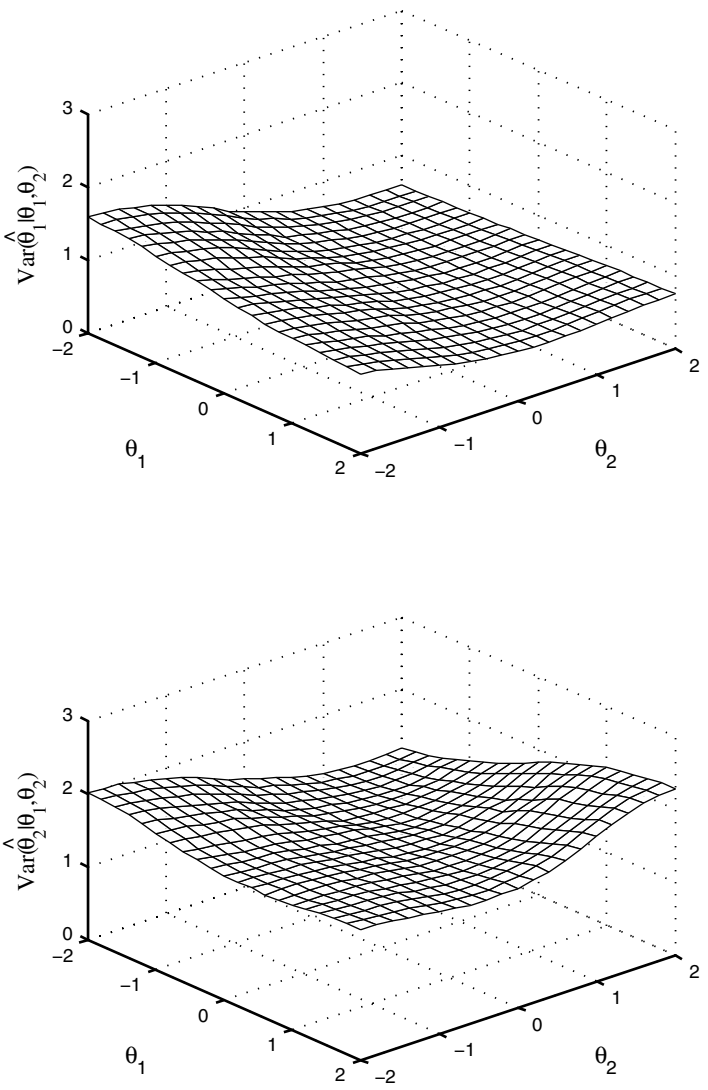


FIGURE 9.12. Estimated variance functions for $\widehat{\theta}_1$ (top) and $\widehat{\theta}_2$ (bottom) for a two-dimensional test with $\theta_1$ an intentional ability and $\theta_2$ a nuisance ability.

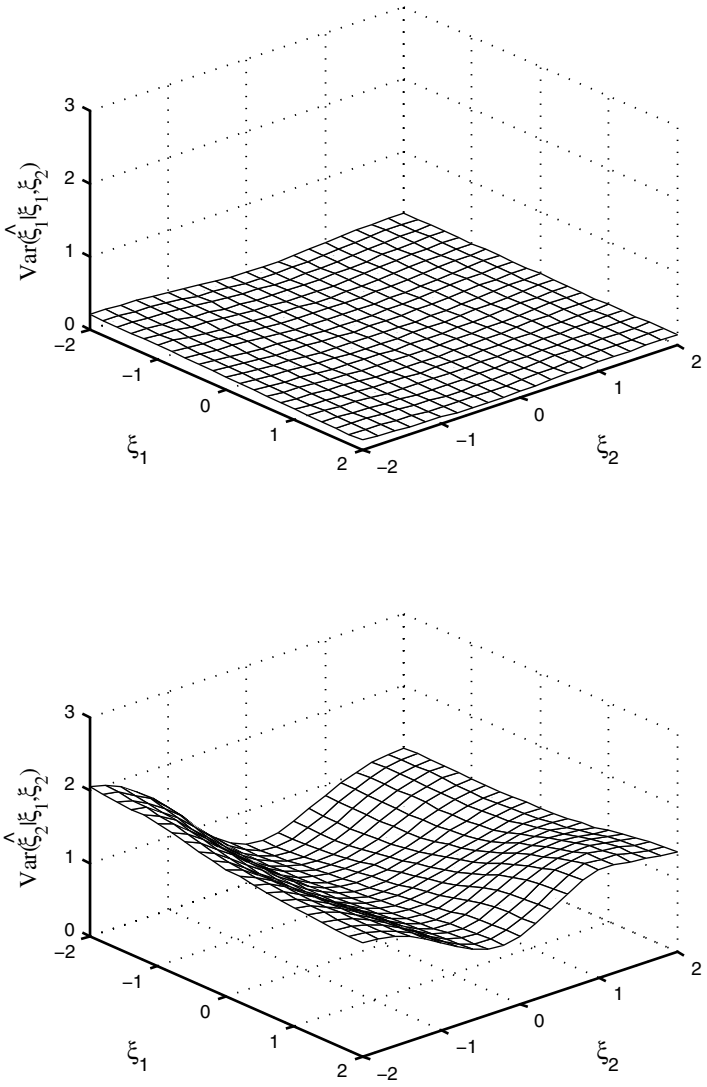


FIGURE 9.13. Estimated variance functions for $\widehat{\xi}_1$ (top) and $\widehat{\xi}_2$ (bottom) for a two-dimensional test with $\xi_1 = .5\theta_1 + .5\theta_2$ an intentional ability and $\xi_2 = -\theta_1 + .5\theta_2$ a nuisance ability.

The two main reasons for accepting a lower level of adaptation that have been put forward are: (i) the wish to have content specialists review intact units in the test before they become operational, and (ii) the opportunity for test takers to go back and review their response to an earlier item. The former is certainly a valid reason. If one wants the content specialist to review units in the test because the results of automated test assembly cannot be trusted (for example, because the items have not yet been coded completely), adaptation is only possible at the level of these intact units.

But the latter is not valid; in adaptive testing, it is always possible to have test takers review earlier items. An obvious strategy is to give them permission to go back to any of the last $m$ items they have taken and change their response if they find it necessary to do so. The only consequence is the necessity to recalculate the current ability estimate from the likelihood for the response vector with the corrected responses. The test is then continued with the new estimate.

In fact, this type of item review is much more efficient than the review offered in testlet-based or multistage adaptive testing, where updating the ability estimate is always postponed until a fixed number of items have been answered. Typically, test takers change their responses only for an occasional item. In this case, the revision of the ability estimate can be expected to lead to a minor change, particularly after the first ten items or so.

An instructive way to view testlet-based, multistage, and adaptive linear on-the-fly testing is as severely restricted versions of the STA. The graphical representation of the method in Figure 9.1 shows that the STA is the case of adaptive testing with the largest number of stages ($n$) as well as the largest choice of alternative test units (all items in the set of feasible tests in the pool) available at each stage. The other forms of adaptive testing are obtained by restricting the number of stages and/or the number of units available after an update. The restriction is maximal for adaptive linear on-the-fly testing, which is in fact equivalent to an STA with only one shadow test. Obviously, the price for these restrictions is lower precision of the test scores.

## 9.9 Literature

Pioneering work on adaptive test assembly was reported in Lord (1970; 1971; 1980, chapter 10) and in numerous research reports by David J. Weiss of the University of Minnesota (see Weiss, 1982). More recent treatments of adaptive testing have been given in van der Linden and Glas (2000) and Wainer (2000), while Parshall, Spray, Kalohn, and Davey (2002) should be consulted for the more practical aspects of adaptive testing.

The shadow-test approach (STA) was introduced in van der Linden and Reese (1998) and further extended in van der Linden (2000a). The weighted-deviations method (WDM) was published in Stocking and Swanson (1993). For an extensive comparison between the STA and the WDM, see van der Linden (2005d); a case study with these two approaches is reported in Robin, van der Linden, Eignor, Steffen and Stocking (2004).

The idea of using use Kullback-Leibler information as an objective in adaptive test assembly was formulated in Chang and Ying (1996). Alternative Bayesian criteria for item selection in adaptive testing were proposed in van der Linden (1998d) and van der Linden and Pashley (2000).

Chang and Ying (1999) were the first to suggest the use of alpha stratification in adaptive test assembly. The implementation of alpha stratification using a shadow test with (9.30)–(9.33) was investigated in van der Linden and Chang (2003). The Sympson-Hetter method of item-exposure control for adaptive testing was presented in Sympson and Hetter (1985); a case study with an application of this method to the adaptive version of the Armed Services Vocational Aptitude Battery (ASVAB) can be found in Hetter and Sympson (1997). Stocking and Lewis (1998) suggest using a conditional version of the Sympson-Hetter method, whereas formal properties of the Sympson-Hetter method and some alternative adjustment schemes for its control parameters are presented in van der Linden (2003). For details of the multiple-shadow-test approach (MSTA), the reader should consult Veldkamp and van der Linden (submitted). The method of item-exposure control based on random ineligibility constraints was introduced in van der Linden and Veldkamp (2004); a derivation of the bounds on the ineligibility probabilities in (9.47) as well as a proof of their property of self adaptation can be found in this reference. For the conditional version of this method, see van der Linden and Veldkamp (submitted). A review of all current methods of item-exposure control will be offered in Segall (in preparation).

The version of the lognormal model for response times was formulated in van der Linden (2005a); this reference should also be consulted for statistical methods for calibrating an item pool and testing the fit of this model against response-time data. The Bayesian constraint on the predicted response times for the test taker in (9.54) was introduced in van der Linden, Scrams, and Schnipke (1999). For other uses of response times to improve adaptive testing, see van der Linden (in preparation).

The idea of equating number-correct scores on an adaptive test to those on a reference test using an adaptive version of the constraints for matching observed-score distributions was developed in van der Linden (2001a). Details of the procedure for the Bayesian initialization of the ability estimator used in the empirical example in Section 9.6 are given in van der Linden (1999a).

Multidimensional adaptive testing with the criterion of minimum error variance for a linear combination of abilities in (9.79) was examined in

van der Linden (1999b), whereas a more systematic treatment of multidimensional adaptive testing with the posterior expected Kullback-Leibler criterion is given in Veldkamp and van der Linden (2002). For a case study of multidimensional adaptive testing with the STA, see Li (2002). A Bayesian version of the criterion of $D$-optimality for use in adaptive test assembly was studied in Luecht (1996) and Segall (1996, 2000).

## 9.10 Summary

1. In adaptive test assembly, items for the test are selected sequentially; each subsequent item is selected to be most informative at an update of the test taker's ability estimate calculated after the preceding response.

2. Adaptive tests have to be assembled to three different sets of constraints: (i) content constraints representing the test specifications, (ii) dynamic constraints necessary to implement the process of adaptation, and (iii) extra constraints to deal with such new problems in adaptive-testing programs as the necessity to constrain the exposure rates of the items, make the test equally speeded for each test taker, and equate the scores to a reference test released for score interpretation.

3. In a shadow-test approach (STA), the constraints are imposed on shadow tests assembled prior to the selection of a new item. Shadow tests are fixed tests that (i) have maximum information at the current ability estimate, (ii) meet all constraints, and (iii) contain all items already taken by the person. The item that is actually administered is the most informative free item in the shadow test. Because each shadow test meets all constraints, the adaptive test always does also.

4. The models used to assemble shadow tests are the same integer programming models as for fixed tests. The only modifications necessary are: (i) the addition of the second and third sets of constraints mentioned in the first point of this summary and (ii) the replacement of an objective function that controls the TIF at a series of fixed values $\theta_k$, $k = 1, ..., K$, by a function that maximizes to the value of the TIF only at the current estimate $\widehat{\theta}_{g-1}$.

5. The STA can be viewed as a projection method for adaptive item selection. At each next ability estimate, it predicts the optimal feasible remaining part of the test and picks the best item from it for administration.

6. The STA can also be viewed as a solution to the dilemma invoked by the combination of sequential item selection and the fact that the constraints can only be satisfied if all items are selected simultaneously. The STA resolves this dilemma by treating adaptive test assembly as a sequence of $n$ simultaneous optimization problems.

7. Adaptive test assembly can be conducted with objectives other than maximizing Fisher's information; for example, objectives based on Kullback-Leibler information and/or Bayesian objectives based on the posterior distribution of $\theta$ updated during the test.

8. If the pool contains set-based items, the model for the shadow tests can be derived from the standard model for the simultaneous selection of items and stimuli in Chapter 7. The model then has to be extended with a constraint that sets the variables of the stimuli already chosen equal to one.

9. An effective way to reduce the item-exposure rates of the more popular items in the pool is through alpha stratification; that is, restricting the selection of the items to subsequent strata in the pool with increasing values for the item-discrimination parameter. Alpha stratification can be implemented by adopting a special constraint in the shadow-test model.

10. An alternative way to reduce the exposure rates for the more popular items is through probabilistic control using the Sympson-Hetter method. The method can be implemented effectively using a multiple-shadow-test approach (MSTA) with the Sympson-Hetter experiment conducted over a list of items composed from the free items in parallel shadow tests. The model for the MSTA can be derived from the standard model for the simultaneous selection of multiple tests in Chapter 6.

11. A more efficient form of probabilistic exposure control is through random ineligibility constraints on the items in the pool. The constraints can be drawn with probabilities that are self-adaptive (that is, they automatically maintain the item-exposure rates at optimal levels and do not require any previous adjustment of control parameters).

12. Adaptive tests tend to be differentially speeded in the sense that some test takers get selections of items that require more time than others. If the items have been calibrated under a response-time model, the test can be made equally speeded by including a special constraint in the model for the shadow tests.

13. If the speed at which the test takers operate is an intentional part of the ability measured by the item pool, the constraint should be

based on the item parameters in the response-time model only. But if it is a nuisance factor, the constraint should be based both on the item and person parameters in the model.

14. Number-correct scores on adaptive tests can be automatically equated to scores on a reference test released for score interpretation by including an adaptive version of the constraints for matching the observed-score distribution in Section 5.3 in the model for the shadow tests. These constraints also allow for a direct comparison of the number-correct scores of different persons taking the adaptive test.

15. If the item pool is multidimensional, the model for the shadow tests can be derived from the standard model for the assembly of a fixed multidimensional test in Chapter 8. The item that is actually administered is the item with a minimal value for a weighted projection of the variances of the ability estimators. The weights offer direct control of the relative importances of these variances in the different cases of multidimensional testing discussed in Chapter 8.

16. Alternatively, the model for the shadow tests can be given an objective function based on a multidimensional generalization of (posterior) expected Kullback-Leibler information. This option allows us to deal with different cases of multidimensional testing, too, but without explicit weights for the individual ability estimators.

17. Testlet-based adaptive testing, multistage testing, and adaptive linear on-the-fly testing are restricted versions of the STA, with smaller numbers of stages at which the ability estimate is updated and smaller numbers of test units to choose from after the updates. The price paid for these restrictions is lower precision of the test score.

## 9.11 Exercises

9.1 Formulate a version of the standard model for a shadow test in (9.11)–(9.21) for the case of an adaptive test with random test length in Section 9.1.1.

9.2 In Section 9.3, the next stimulus in an adaptive test is identified by the most informative free item in the shadow test. An alternative criterion would be to select the free stimulus in the shadow test with the largest average information for its items. Which criterion is best?

9.3 Show that the sum of the exposure rates of the items in the pool in adaptive testing is always equal to the test length, $n$. What is the effect on the other items in the pool of lowering the exposure rate of an item with a tendency toward overexposure to less than $r^{\max}$?

For an adaptive test of 15 items, would it be possible to design an exposure-control method that forces the exposure rates of all items in a pool of 350 to be between .05 and .10?

9.4 Let $t = 1, 2, ...$ be the iteration steps in a sequence of adaptive-testing simulations conducted to adjust the control parameters $P(A_i \mid S_i, \theta)$ in the Sympson-Hetter method. After each step, we have new estimates of the probabilities of selecting the items, $P^{(t)}(S_i \mid \theta)$. The Sympson-Hetter method uses the following adjustment rule for the control parameters:

$$P^{(t+1)}(A_i \mid S_i, \theta) = \begin{cases} 1 & \text{if } P^{(t)}(S_i \mid \theta) \leq r^{\max}, \\ r^{\max}/P^{(t)}(S_i \mid \theta) & \text{if } P^{(t)}(S_i \mid \theta) > r^{\max}. \end{cases}$$

Motivate the adjustment of the control parameters for $P^{(t)}(S_i \mid \theta) > r^{\max}$. Does it make sense to set the control parameters equal to 1 if $P^{(t)}(S_i \mid \theta) \leq r^{\max}$? Formulate a few alternatives for this part of the adjustment rule.

9.5 Show that the upper bound for $P(E_i \mid \theta)$ in (9.47) follows from (9.46) along with the assumption that $P(E_i \cap F \mid \theta) = P(E_i \mid \theta)P(F \mid \theta)$. How reasonable is this independence assumption?

9.6 For a well-designed adaptive test, the probability of a feasible test for test taker $e$ can be expected to be equal to $P^{(e)}(F \mid \theta) = 1$. Use this condition to show the property of self-adaptation for the updates of the probabilities of item eligibility in (9.47).

9.7 Motivate the choice of the constraint on the time intensities of the items in the shadow test in (9.50) using the identifiability constraint in (9.49) and the fact that $t_{\text{tot}}$ has been chosen to be large enough for the population of test takers.

9.8 Suppose the response times of the test takers have a common standard deviation of 30 seconds for all items in the pool. What would be your estimate of the standard deviation of the total time for a test taker on a test of 25 items?

9.9 In the empirical example in Section 9.5.3, we began the test with the 50th percentile of the posterior predicted response-time distributions in the constraint in (9.53). Why is this choice reasonable? Why does it make sense to move to the 95th percentile for the selection of the last two items in the test.

9.10 Suppose both abilities measured by the items in a two-dimensional pool are intentional. In addition, suppose it holds that if the weights $w_1$ and $w_2$ for an adaptive test are set as in (9.71), the STA meets the criterion of $D$-optimality (i.e., it minimizes the determinant of

the covariance matrix in (8.4)). Formulate the criterion for item selection that has to replace (9.80) if the items from the shadow test are selected to be $D$-optimal?