Chapter 5
Sequencing an Adaptive Test Battery

Wim J. van der Linden

5.1 Introduction

Switching a testing program from a linear to an adaptive format increases its efficiency considerably. The gain in efficiency can be used to shorten the length of the test or increase the accuracy of the scores. The gain is especially relevant to testing programs in which a battery of tests has to be administered in a single session but the testing time has to remain feasible. Examples of such programs are diagnostic testing for instructional purposes (e.g., Boughton, Yao, & Lewis 2006; Yao & Boughton, 2007) and large-scale assessments of education. These programs generally involve the reporting of profiles of scores of students, schools, or districts. In order to use such profiles for decision making, each of their individual scores should have satisfactory accuracy. The more advantageous combination of testing time and score accuracy made possible by the use of a battery of adaptive instead of linear tests has been highlighted earlier, for instance, in Brown and Weiss (1977) and Giallucca and Weiss (1979).

In an adaptive test battery, each individual test is assembled from a different item pool. For the first test, an initial ability estimate is chosen and the first item is selected from the first pool to be optimal at this estimate. The response is then used to update the initial estimate, and the second item is selected to be optimal at the update. The process is repeated until a predetermined number of items or accuracy level is reached, whereupon a new test from a new pool is started. Using this format, a typical saving of the length of the individual tests by some 40–60% percent relative to a linear version of them is possible.

It is easy to confuse the case of a battery of unidimensional adaptive tests addressed in this research with a multidimensional adaptive test, particularly if the abilities correlate. Multidimensional adaptive testing has its own procedures of optimal item selection (Mulder & van der Linden, 2009; Segall, 1996, this volume,
For these procedures, it is not necessary (and, in fact, even disadvantageous) to constrain the selection of items to one and the same subpool until a predetermined number of items is administered, as is done in testing with a battery of unidimensional tests.

In the current research, the focus was not on the question of how to select the individual items. Readers with an interest in this question should refer to Thissen and Mislevy (2000) or van der Linden and Pashley (this volume, chap. 1). Instead, the interest was in the optimal sequencing of the test battery, the idea being that instead of always administering the tests in the battery in the same predetermined order for each test taker, its efficiency could be increased further by optimizing the order for the individual test takers. In fact, the best approach seems to repeat the principle of adaptation at the level of the selection of the tests. We would then pick the first test to be optimal over the initial estimates of each of the test taker’s abilities measured by the battery. The second test would be chosen to be optimal given the test taker’s responses to the first test. And so on.

A statistical notion relevant to such an approach is that of collateral information, i.e., the information about the test taker’s ability measured by one test available in the other tests. We expect the amount of collateral information in test batteries to be substantial because they are typically designed to measure a set of strongly related but distinct abilities, for instance, abilities in early mathematics or language acquisition in elementary education. It would be a waste to ignore such information when choosing a next test from the battery. The problem of how to improve subtest scores by borrowing information from other subtests has received considerable attention recently (e.g., Wainer et al., 2001). The solution to the problem of sequencing the tests in a battery addressed in this research can be viewed as an adaptive solution to this more general problem.

An appropriate framework for implementing the adaptive approach is multilevel item-response theory (IRT) in combination with an empirical Bayes procedure for the selection of the tests. The two-level model used in this research consists of distinct response models for the unidimensional item pools as first-level models and a specification of the joint distribution of their ability parameters for the population of test takers as a second-level model. The second-level model allows the borrowing of information from earlier response vectors about the abilities measured by later tests. An empirical Bayes approach represents this information in the form of prior and posterior distributions: each time a test is completed, the test taker’s response vector is used to update the posterior predictive distributions of the abilities measured by the remaining tests, whereupon the updates are used as prior distributions for the selection of the next test. The approach is empirical in that the second-level model for the distribution of the abilities is estimated from test data collected during pretesting of the items.

For the results to be realistic, we have to allow for content constraints on the item-selection process. These constraints are necessary to implement the test agency’s specifications for each individual test, such as its required content distribution, possible logical relations between the items in the pool (e.g., items that should not