Chapter 2
Constrained Adaptive Testing with Shadow Tests

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

1. Introduction

The intuitive principle underlying the first attempts at adaptive testing was that a test has better measurement properties if the difficulties of its items match the ability of the examinee. Items that are too easy or difficult have predictable responses and cannot provide much information about the ability of the examinee. The first to provide a formalization of this principle was Birnbaum (1968). The information measure he used was Fisher’s well-known information in the sample. For the common dichotomous item response theory (IRT) models, the measure is defined as

\[ I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} \frac{(P_i'(\theta))^2}{P(\theta)[1 - P(\theta)]}, \]  

(1)

where \( P_i(\theta) \) is the probability that a person with ability \( \theta \) gives a correct response to item \( i = 1 \ldots n \).

For the one-parameter logistic (1PL) model, the information measure takes a maximum value if for each item in the test the value of the difficulty parameter \( b_i = \theta \). The same relation holds for the 2PL model, though the maximum is now monotonically increasing in the value of the discrimination parameter of the items, \( a_i \). The empirical applications discussed later in this chapter are all based on response data fitting the 3PL model:

\[ P_i(\theta_j) \equiv c_i + (1 - c_i) \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}. \]  

(2)

For this model, the optimal value of the item-difficulty parameter is larger than the ability of the examinee due to the possibility of guessing on the items. The difference between the optimal value and the ability

* Part of the research in this chapter received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this chapter are those of the author and do not necessarily reflect the position or policy of LSAC.
of the examinee is known to be a monotonically increasing function of
the guessing parameter, \( c_t \).

Both test theoreticians and practitioners immediately adopted the
information measure in (1) as their favorite criterion of test assembly.
The fact that item information additively contributes to the test preci-
sion has greatly enhanced its popularity. Though other criteria of item
selection have been introduced later (for a review see van der Linden
and Pashley, this volume), the most frequently used criterion in com-
puterized adaptive testing (CAT) has also been the one based on the
information measure in (1).

Though adaptive testing research was initially motivated by the in-
tention to make testing statistically more informative, the first real-life
testing programs to make the transition to CAT quickly discovered that
adaptive testing operating only on this principle would lead to unre-
realistic results. For example, if items are selected only to maximize the
information in the ability estimator, test content may easily become un-
balanced for some ability levels. If examinees happened to learn about
this feature, they may change their test preparations and the item cali-
bration results would no longer be valid. Likewise, without any further
provisions, adaptive tests with maximum information can also be un-
balanced with respect to such attributes as their possible orientation
towards gender or minority groups and become unfair for certain groups
of examinees. Furthermore, even a simple attribute such as the serial
position of the correct answer for the items could become a problem
if the adaptive test administrations produced highly disproportionate
use of particular answer keys. Lower ability examinees might benefit
from patterned guessing and some of the more able examinees might
become anxious and begin second-guessing their answers to previous
items. Examinees may get alerted by this fact and start bothering if
their answers to the previous questions were correct.

More examples of nonstatistical specifications for adaptive tests are
easy to provide. In fact, what most testing programs want if they make
the transition from linear to adaptive testing, is test administrations
that have exactly the same “look and feel” as their old linear forms but
that are much shorter because of a better adaptation to the ability levels
of the individual examinees. The point is that adaptive testing will only
be accepted if the statistical principle of adapting the item selections to
the ability estimates for examinees is implemented in conjunction with
serious consideration of many other nonstatistical test specifications.

Formally, each specification that an adaptive test has to meet im-
poses a constraint on the selection of the items from the pool. As a con-
sequence, a CAT algorithm that combines maximization of statistical
information with the realization of several nonstatistical specifications