Chapter 16
Detecting Person Misfit in Adaptive Testing

Rob R. Meijer and Edith M.L.A. van Krimpen-Stoop

16.1 Introduction

An examinee’s test score does not reveal the operation of undesirable influences of test-taking behavior such as faking on biodata questionnaires and personality tests, guessing, or knowledge of the correct answers due to test preview on achievement tests. These and other influences may result in inappropriate test scores, which may have serious consequences for practical test use, for example, in job and educational selection, where classification errors may result. In the context of item response theory (IRT) modeling, several methods have been proposed to detect item score patterns that are not in agreement with the expected item score pattern based on a particular test model. These item score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level (θ). Research with respect to methods that provide information about the fit of an individual item score pattern to a test model is usually referred to as appropriateness measurement or person fit measurement. Most studies in this area are, however, in the context of paper-and-pencil (p&p) tests. As will be argued below, the application of person fit theory presented in the context of p&p tests cannot simply be generalized to a computerized adaptive test (CAT). In this chapter we introduce and review the existing literature on person fit in the context of a CAT.

Before we introduce person fit research, it is important to realize that not all types of aberrant behavior affect individual test scores. For example, a person may guess the correct answers of some of the items but also guess wrongly on some of the other items and, as the result of the stochastic nature of guessing, this process may not result in substantially different test scores under most IRT models to be discussed below. Whether aberrant behavior leads to nonfitting item score patterns depends on numerous factors such as the type and amount of aberrant behavior.

Furthermore, it may be noted that all methods discussed in this chapter can be used to detect nonfitting item score patterns, but several of these methods do not
allow the recovery of the mechanism that created the deviant item score patterns. Other methods explicitly test against specific violations of a test model assumption or against particular types of deviant item score patterns. The latter group of methods therefore may facilitate the interpretation of nonfitting item score patterns.

16.2 Practical and Theoretical Relevance of Person Fit Analysis in a CAT

There are a number of causes why an examinee may not respond according to the IRT model in a p&p test and in a CAT. Misfitting item score patterns may be the result of

- low-ability persons who copy answers from more able neighbors (see e.g., Levine and Rubin, 1979),
- exceptionally creative persons who discover novel interpretations for some items (e.g., Levine and Drasgow, 1988),
- persons who are randomly guessing the correct answers to all the items in the test because they only take the test to become familiar with the items that are administered, or
- persons who have preknowledge of (some of the) items in the test and as a result give correct answers to relatively difficult items.

In the context of a CAT, some of the above-mentioned causes are less likely. For example, copying answers from more able neighbors is very unlikely because in a CAT each examinee is administered different items in a different order. Also, alignment errors are unlikely in adaptive testing, because the answers are given via the keyboard and the next item is only shown after an answer to the current item is given.

When the exam is administered by means of a computer, misfit may be the result of

- persons who are unfamiliar with a computer or may have trouble settling in or warming-up. For these persons the earliest items are more likely to be answered incorrectly than later items;
- unmotivated persons in pretesting situations. For these persons the answers may be unrelated to the item difficulty and a pattern may be generated that seems the result of random response behavior;
- persons who have preknowledge of (some of the) items in the test, and as a result give correct answers to subareas of the test. For example, a group of test takers (sources) are administered a CAT and memorize the items presented. Then the sources make a list of their items and provide other test takers (memorizers) the opportunity to memorize these items. If these memorizers are successful, they may use their item preknowledge when administered a test. An alternative is that examinees may have extra training in various subareas of the exam, which may result in multidimensionality in a CAT that is assumed to measure a single ability.