

Kernel-Smoothed DIF Detection Procedure for Computerized Adaptive Tests (CT 00-08)

Ratna Nandakumar
Julie Cwikla-Banks
Louis Roussos

Abstract

As part of the continuing efforts of the Law School Admission Council (LSAC) to maintain the highest quality of testing methodology for its Law School Admission Test (LSAT), a research program is investigating a wide range of factors in regard to the possibility of computerizing the LSAT.

One important component is the maintenance of the fairness of the test with regard to certain subgroups of interest in the testing population. In any testing situation, it is essential that items are fair to all subgroups of test takers. In other words, if on an item, equally able test takers from two subgroups are not performing equally well, then such an item will be cause for concern. This phenomenon is known as differential item functioning (DIF). A number of different methodologies have been studied for the detection of DIF on paper-and-pencil tests, but new techniques need to be developed for computerized tests.

In a previous LSAC report, Nandakumar and Roussos (2001) reported their development of CATSIB, a new DIF procedure for computerized tests, which is a modified version of SIBTEST (a DIF assessment methodology for paper-and-pencil tests). CATSIB can be used for DIF assessment at the pretest stage and, if necessary, can continue with DIF monitoring at the operational stage, using the combined pretest and operational data. By conducting a large-scale simulation study, Nandakumar and Roussos (2001) showed that CATSIB is a practical and reliable statistical procedure for detecting DIF in pretest items that are being calibrated in a CAT setting. CATSIB has exhibited good Type 1 error (false detection) rates and high power (true detection) rates for sample sizes of 250 to 500, which are sizes that are typically encountered in CAT settings. These small sample sizes were used because pretest sample sizes for computerized adaptive tests, for a variety of reasons, tend to be smaller than those obtained for paper-and-pencil pretest items. Thus, it is important to develop DIF detection procedures that have maximum detection power even for small samples.

The current paper extends this DIF research by considering whether a modification of the CATSIB procedure would increase its power to detect DIF without increasing its false detection rate (Type 1 error rate). The modification is known as "kernel smoothing," a

statistical technique that can be used to accurately estimate the expected score of a group of test takers on an item. If this estimation is done for two different groups, and group ability differences are adjusted for, then such estimation can be used to test for whether DIF is present in the item.

Thus, the current paper combines the kernel-smoothing estimation approach with CATSIB in the hope of obtaining a more efficient DIF detection procedure. It is hoped that by combining these two statistical methodologies, CATSIB and kernel-smoothing estimation, a more effective DIF detection procedure for CAT data can be obtained. The resulting procedure will be referred to as KS-CATSIB.

A simulation study was conducted to investigate the DIF estimation bias of KS-CATSIB in comparison to CATSIB with small samples. Sixteen studied items varying in difficulty and discrimination were considered for this purpose. A sample of 500 test takers was used in the reference group and a sample of 250 test takers was used in the focal group. The results showed that KS-CATSIB exhibited a large amount of bias in comparison to CATSIB. A number of different methods were employed to try to reduce the bias, but even the most successful method still resulted in a large amount of bias in comparison to CATSIB.

Therefore, it is currently recommended that kernel smoothing not be employed in the CATSIB DIF procedure until new methods for dealing with the statistical bias are developed and studied.