LSAC RESEARCH REPORT SERIES

- Detection of Invalid Test Scores on Admission Tests: A Simulation Study Using Person-Fit Statistics

  Jorge N. Tendeiro
  Rob R. Meijer
  Casper J. Albers
  University of Groningen, Groningen, the Netherlands

- Law School Admission Council
  Research Report 15-03
  March 2015

A Publication of the Law School Admission Council
The Law School Admission Council (LSAC) is a nonprofit corporation that provides unique, state-of-the-art products and services to ease the admission process for law schools and their applicants worldwide. Currently, 222 law schools in the United States, Canada, and Australia are members of the Council and benefit from LSAC’s services. All law schools approved by the American Bar Association are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also members. Accredited law schools outside of the United States and Canada are eligible for membership at the discretion of the LSAC Board of Trustees; Melbourne Law School, the University of Melbourne is the first LSAC-member law school outside of North America. Many nonmember schools also take advantage of LSAC’s services. For all users, LSAC strives to provide the highest quality of products, services, and customer service.

Founded in 1947, the Council is best known for administering the Law School Admission Test (LSAT®), with about 100,000 tests administered annually at testing centers worldwide. LSAC also processes academic credentials for an average of 60,000 law school applicants annually, provides essential software and information for admission offices and applicants, conducts educational conferences for law school professionals and prelaw advisors, sponsors and publishes research, funds diversity and other outreach grant programs, and publishes LSAT preparation books and law school guides, among many other services. LSAC electronic applications account for 98 percent of all applications to ABA-approved law schools.

© 2015 by Law School Admission Council, Inc.

All rights reserved. No part of this work, including information, data, or other portions of the work published in electronic form, may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Publishing, and Creative Services, Law School Admission Council, 662 Penn Street, PO Box 40, Newtown, PA, 18940-0040.

This study is published and distributed by LSAC. The opinions and conclusions contained in this report are those of the author(s) and do not necessarily reflect the position or policy of LSAC.
# Table of Contents

Executive Summary ............................................................................................................. 1

Introduction ......................................................................................................................... 1

Person Fit for Multiple Subtests .......................................................................................... 2
  Existing Statistics ................................................................................................................. 2

Objectives ............................................................................................................................. 4

Simulation Study .................................................................................................................. 5

Results .................................................................................................................................. 7
  Empirical Type I Errors ....................................................................................................... 7
  Detection Rates ................................................................................................................... 7

Discussion ............................................................................................................................. 12

References ............................................................................................................................. 13
Executive Summary

While an admission test may strongly predict success in university or law school programs for most test takers, there may be some test takers who are mismeasured. To address this issue, a class of statistics called person-fit statistics is used to check the validity of individual test scores. However, most person-fit statistics are designed for a single test, and not much is known about the performance of these statistics for admission tests consisting of multiple highly correlated subtests. In this study, the performance of a number of person-fit statistics was evaluated based on data that simulated aberrant responding on highly correlated subtests. The results indicated that two of the statistics outperformed the others.

Introduction

Predictive validity studies make it possible to identify applicants who are most likely to succeed in college, graduate school, or professional-degree programs such as medical school and law school. These studies depend on information provided by grade point average (GPA) and scores on tests such as the SAT Reasoning Test or the Law School Admission Test (LSAT) and, often, on the combination of both GPAs and test scores. For a majority of applicants, predictive validity is high. For example, Anthony, Dalessandro, and Reese (2013) found a multiple correlation between first-year average in law school and a combination of LSAT score and undergraduate GPA (UGPA) of 0.47. However, correlations like these are based on groups of test takers; for some individual test takers, this relation may be much lower. One reason may be that they are mismeasured by the admission test; that is, the admission total score may not be a good indicator of the test taker’s true proficiency level. Because at an individual level the stakes are high, it is important to check the validity of individual test scores on admission tests. Indeed, some universities are using a compensatory model, in which low scores on admission tests may be compensated by high scores on other predictors, such as UGPAs (Anthony et al., 2013).

One way to check the validity of test scores is to study the configuration of item scores. Very unlikely responses to individual items may result in test scores that do not provide an adequate picture of a test taker’s true proficiency level. To detect these inconsistent, or misfitting, item score patterns, several person-fit statistics have been proposed. However, almost all of these statistics have been applied to single (sub)tests (Meijer & Sijtsma, 2001). In practice, however, many admission tests consist of different subtests (below denoted as multiple subtests) and the sum of the total score across subtests is used to obtain an impression of a test taker’s proficiency level. For example, on the LSAT, the sum of the total number of questions answered correctly across the four scored sections (i.e., the raw score on Reading Comprehension, Logical Reasoning (two parts), and Analytical Reasoning) is converted to a score on a 120–180 scale (i.e., the scaled score). For these types of tests, it is unclear from the existing literature which person-fit statistic to use. The aim of this report is to compare the performance of a number of parametric and nonparametric statistics using simulated data that mimic the psychometric structure of educational tests consisting of multiple subtests.
The report is organized as follows. First, an overview of the current status of person-fit analyses in multiple-subtest settings is given, and the goals of our research proposal are presented. Second, the details of our simulation study are explained. Third, the most important findings from the simulation study are presented. Fourth, our main conclusions are summarized and possible future research directions discussed.

**Person Fit for Multiple Subtests**

Although person-fit statistics can be calculated for each subtest, there are a number of advantages to analyzing person fit across subtests. First, the power of person-fit statistics to detect aberrant response behavior increases with increased test length (e.g., Meijer & Sijtsma, 2001). As several studies have shown, the power of person-fit statistics is low for relatively short tests (say, less than 15 items). Second, from a psychological point of view it is interesting to study the response behavior across all items of a test to investigate whether unexpected response behavior occurs on specific subtests or across different subtests. Although person-fit scores can be calculated for each subtest and compared across subtests, the information about aberrant response behavior across all subtests is lost. Third, because (a transformed) total score across all subtests is often used for decision-making, it makes sense to study consistency of response behavior across all items of the test.

**Existing Statistics**

A number of different statistics based on a likelihood approach have been proposed in the literature (Karabatsos, 2003; Meijer & Sijtsma, 2001). For unidimensional subtests, an often used person-fit statistic is the loglikelihood statistic $z_l$ (Drasgow, Levine, & Williams, 1985). Let $X_i$ denote the random variable consisting of the score on a dichotomous item $i$ ($i = 1, \ldots, I$), and let $P_i(\theta)$ denote the item response theory (IRT) model describing the probability of a person with latent trait $\theta$ answering item $i$ correctly. The most common dichotomous IRT models are the one-, two-, and three-parameter logistic models; see, for example, Embretson and Reise (2000). The $l_z$ value for person $n$ ($n = 1, \ldots, N$) is given by

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}},$$

where $l_0 = \log L(\theta) = \sum_{i=1}^{I} \left\{ X_i \log P_i(\theta) + (1 - X_i) \log [1 - P_i(\theta)] \right\}$ is the logarithm of the likelihood function, and $E(l_0)$ and $Var(l_0)$ are the mean and variance of $l_0$, respectively. The $l_z$ statistic can be used to determine the fit of a person response pattern to the IRT model defined by $P_i(\theta)$. Drasgow, Levine, and McLaughlin (1991) proposed a statistic similar to Equation (1) for the multiple-subtest setting. They compared six practical person-fit statistics by means of a simulation study, and
concluded that the $l_z$-based statistic performed reasonably well across varying testing conditions. Recently, Conijn, Emons, and Sijtsma (2014) extended the Drasgow et al. (1991) study by comparing different alternative methods based on the $l_z$ statistic that combined information from different multi-unidimensional tests. They distinguished five different approaches:

1. In the first approach the $l_z$ statistic was calculated across all subtests as though the test were unidimensional. This statistic was denoted $l_{z(uni)}$. This can be referred to as the unidimensional approach.
2. In the second approach, the $l_z$ statistic was calculated for each subtest separately (denoted $l_{z(sub)}$). This can be referred to as the subscale analysis approach.
3. In the third approach the sum of the individual subtest statistics, $l_{zm}$, was calculated. This can be referred to as the multi-test extension approach.
4. In the fourth approach, the statistic $l_{z(com)}$ was based on the $l_{zm}$ values of each possible subset of subtests (including the single subtests). A pattern was classified as aberrant if at least one of the resulting statistics suggested misfit.
5. In the fifth approach, a pattern was classified as aberrant when $l_z$ was used for the subtests and $l_{zm}$ was used for the complete test. This approach was denoted $l_{z(sel)}$.

Throughout this report, the five approaches above will be referred to as the ‘uni’, ‘sub’, ‘m’, ‘com’, and ‘sel’ approaches, respectively, irrespective of the specific person-fit statistic used ($l_z$ in this case).

Conijn et al. (2014) based their study on polytomously scored items within the context of noncognitive assessment. Both *globally* (i.e., across the entire test) and *locally* (i.e., within one or two subtests), inconsistent response vectors were generated. Based on simulated datasets, using parametric bootstrap procedures and controlling for false discovery rates, Conijn et al. (2014) recommended using the $l_{z(sel)}$ to detect inconsistent response vectors in general, although the power of $l_{zm}$ for detecting global misfit was also substantial. Global misfit was better detected by $l_{z(uni)}$ and $l_{zm}$, whereas local misfit was better detected by $l_{z(sub)}$ and $l_{z(sel)}$. However, these results are difficult to generalize for educational tests. First, the Conijn et al. study was conducted in the noncognitive domain using polytomous data and using subtests consisting of 6 and 12 items. Conijn et al. (2014) used polytomous data simulated on the basis of IPIP-50 item parameters under the graded response model, with trait values drawn from a standard normal multivariate distribution with correlations between the trait values of .4, .6, and .8. Although these choices can very well be defended in a noncognitive domain, for educational tests such as the LSAT the number of items per subtest is typically larger (say, 25–30 items), and
observed correlations between total scores are often around .7–.8 between subtests. Drasgow et al. (1991) discussed correlations (after correction for attenuation) of \( r = .73 \) between SAT Verbal and Quantitative test; \( r = .80 \) between the enhanced ACT English and Mathematics test; and \( r = .80 \) for the general Aptitude test battery Verbal Aptitude and Numerical Aptitude. It is currently unknown how any of the multiple-subtest approaches described above performs under more realistic assumptions in the framework of cognitive assessment. The goal of the present report is to clarify this issue.

**Objectives**

In the present report, the methodology in Conijn et al. (2014) is extended in three ways. First, the power of the five multiple-subtest approaches discussed above is investigated when they are applied to dichotomously scored items. A simulation study mimicking high-stakes educational tests was set up for this purpose. The results to be expected are mostly unknown. As some authors have discussed (e.g., Reise & Revicki, 2014), the psychometric characteristics of maximum performance tests are different from the psychometric structure of typical performance tests as discussed in Conijn et al. (2014). Second, the improved statistic proposed by Snijders (2001), denoted \( l'_{\pi} \), is used instead of the original \( l_{\pi} \) statistic. The \( l'_{\pi} \) statistic provides a better approximation of the asymptotic distribution to the standard normal (see Snijders, 2001). Third, in addition to \( l'_{\pi} \), nonparametric person-fit statistics are also considered. Karabatsos (2003) found that nonparametric statistics are a good alternative and often perform better than the parametric alternatives. Tendeiro and Meijer (2014) also found that the \( C^* \) and \( H^T \) statistics are among the nonparametric statistics that performed best for unidimensional tests. It was therefore decided to use these statistics under the current multiple-subtest framework. Below, both the \( C^* \) and \( H^T \) statistics are defined.

Assume, without loss of generality, that items are ordered by increasing estimated item difficulty, where item difficulty is defined as the proportion of respondents who answered the item correctly. Also, denote the unit weight test score of person \( n (n = 1, \ldots, N) \) by \( s_n \). The \( C^* \) statistic, also referred to as the modified caution index (Harnisch & Linn, 1981), is given by

\[
C^* = \frac{\text{Cov}(x^*_n, p) - \text{Cov}(x_n, p)}{\text{Cov}(x^*_n, p) - \text{Cov}(x_n, p)^*}
\]

where \( x_n \) denotes the response vector of person \( n (n = 1, \ldots, N) \), \( x^*_n \) is the so-called Guttman vector containing correct answers for the \( s_n \) easiest items only, \( x^*_n \) is the reversed Guttman vector containing correct answers for the \( s_n \) hardest items only, and \( p \) denotes the vector of item proportion-correct in the sample. \( C^* \) is
sensitive to Guttman errors. A Guttman error is a pair of scores (0,1), where the 0-score pertains to the easiest item and the 1-score pertains to the hardest item. $C^*$ ranges between 0 (perfect Guttman vector) and 1 (reversed Guttman vector). Now, in order to define the $H^T$ statistic, rearrange the rows of the data matrix such that the order of total score $s_n (n = 1, \ldots, N)$ is increasing. The $H^T$ statistic for person $n$ is given by

$$H^T = \frac{\sum_{m>n}(t_{nm} - t_n t_m)}{\sum_{m>n}(t_m - t_n t_m) + \sum_{n<m}(t_n - t_m t_n)},$$

with $t_n = s_n / I, t_m = s_m / I$, and $t_{nm}$ being the proportion of items answered correctly by both respondents $n$ and $m$ (Sijtsma & Molenaar, 2002, p. 57). This statistic is equivalent to the ratio $\text{Cov}(x_n, r_{(n)}) / \text{Cov}_{mn}(x_n, r_{(n)})$, where $r_{(n)}$ is the vector of total item scores computed excluding respondent $n$, and the denominator is the maximum covariance given the marginal. $H^T$ is maximum 1 when no respondent with a total score smaller/larger than $t_n$ can answer an item correctly/incorrectly that respondent $n$ has answered incorrectly/correctly, respectively. $H^T$ equals zero when the average covariance of the response pattern of respondent $n$ with the other response patterns equals zero.

It is unknown how nonparametric person-fit statistics perform in a multiple-subtest setting. Hence, both the $C^*$ and $H^T$ statistics are used in the multiple-subtest approaches presented. Thus, a total of 15 alternative person-fit statistics especially designed to fit multiple-subtest designs are considered in this study.

### Simulation Study

Data were generated as follows. Item scores were generated for four subtests, each consisting of 25 items. Each dataset consisted of 10,000 item response vectors. The two-parameter logistic model (Embretson & Reise, 2000) was used, with discrimination parameters uniformly distributed between [0.5, 2.0] and difficulty parameters standard normally distributed (bounded between $-2.5$ and $+2.5$). Moreover, four person $\theta$ parameters were generated for each simulated test taker, one per subtest. These parameters were randomly drawn from a multivariate normal distribution with between-subtest correlations of .8. As observed in the Introduction, this value is in accordance with values reported in Drasgow et al. (1991) and is also in accordance with our own observation for often used high-stakes tests. These item and person parameters resulted in data that were very similar to the empirical data from a number of large-scale, high-stakes educational admission tests (see also Rupp, 2013).
Aberrant response patterns were generated following a procedure similar to the procedure devised by Conijn et al. (2014). Two proportions of aberrant response patterns in the dataset were considered: \( AbN = .10 \) (i.e., 1,000 response patterns) and \( AbN = .30 \) (3,000 response patterns). For each dataset, the response patterns assigned to reflect aberrant response behavior were randomly and equally distributed among eight conditions: 20%, 40%, 60%, or 80% of the item scores across all subtests were randomly selected and changed (global misfit), and 50% or 100% of the item scores of either 1 or 2 randomly chosen subtests were changed (local misfit). Thus, for a fixed proportion \( AbN \), both global and local misfitting response patterns were generated in each dataset. Two types of aberrant response behavior were simulated: spuriously low response behavior, where the probability of answering each selected item correctly was equal to .20; and spuriously high response behavior, where the probability of answering each selected item correctly was equal to .80.

Thus, the simulation study consisted of four experimental conditions resulting from a \( 2 ( AbN = .10, .30 ) \times 2 \) (type of aberrant behavior = spuriously low, high) completely crossed design. One hundred replications were simulated per condition. As explained in the Introduction, 15 alternative multiple-subtest person-fit methods were considered. These consisted of the ‘uni’, ‘sub’, ‘m’, ‘com’, and ‘sel’ approaches based on the \( I^*, C^* \), and \( H^T \) person-fit statistics.

A bootstrap procedure was used to decide which item response patterns should be flagged as aberrant by each of the 15 person-fit methods. For a given dataset, 12 model-fitting datasets were generated based on the same estimated item parameters and on a fixed \( \theta \) parameter (12 values: \( \theta = -2.50, -2.25, -2.00, -1.75, -1.50, -1.25, -1.00, 0.00, 0.25, 0.50, 0.75, 1.00 \)); 1,000 replications were used. The 15 person-fit methods were applied to each of the 12 model-fitting datasets. This provided bootstrap distributions for each person-fit method under the null model of consistent response behavior, for each of the 12 \( \theta \) values above. Then, for each item response pattern in the original dataset, \( p \)-values were computed for each person-fit method. The bootstrap distribution corresponded to the \( \theta \) value closest to the response vector’s estimated \( \theta \) from the original dataset. This procedure was sufficient to allow flagging of misfitting response patterns for the ‘uni’ and the ‘m’ approaches (a 5% significance level was used), for which only one \( p \)-value was available. The ‘sub’, ‘com’, and ‘sel’ approaches rendered multiple \( p \)-values for each person-fit method; therefore the inflation of Type I error rates had to be prevented. This was achieved by controlling the false discovery rate (FDR), which is the expected proportion of false positives. The FDR is equal to the family-wise error rate (probability of at least one false positive) when all null hypotheses are true, although it can be more powerful when at least one null hypothesis is false (Benjamini & Hochberg, 1995). The FDR method developed by Benjamini and Hochberg (1995), called the BH method, was applied. According to the BH method, a set of \( p \)-values is ordered, \( p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)} \), and the value \( k \) is found as the largest \( i \) for which \( p_{(i)} \leq (i/m)\alpha \). Finally, all null hypotheses
corresponding to the ordered $p$-values $p_(i),\ldots,p_(k)$ are rejected. For the ‘sub’, ‘com’, and ‘sel’ approaches, an item response vector was flagged as misfitting whenever at least one null hypothesis was rejected according to the BH method.

Note that both spuriously low and spuriously high scores were generated. Although in high-stakes testing, detecting test takers with spuriously high scores (i.e., test takers with higher scores than expected on the basis of their proficiency level) seems to be most important to flag, in practice there is also the need to have information about spuriously low scores. This is, for example, the case when schools or universities use a compensatory admission model. Anthony et al. (2013, p. 14) stated that “some students with low test scores or low UGPAs are admitted to law school, but usually they are not typical of the low-scoring applicants who are rejected. Instead, they are admitted because the school has some other evidence of their ability to do well in law school. (...) This model allows, for example, a high LSAT score to compensate for a low UGPA or, conversely, a high UGPA to compensate for a low LSAT score when schools are making admission decisions.” Thus, sometimes schools select persons with admission test scores that are unexpected, given other evidence. In these cases it is interesting to know whether the results on the admission test are perhaps due to mismeasurement of such persons. In fact, schools may use information from these kinds of statistics, if they are given by the test provider, as a clue to search for additional evidence.

Results

Empirical Type I Errors

In each of the four experimental conditions across all replications, the empirical Type I error rates for the 15 person-fit methods were analyzed. In general, the conditions based on 10% of misfitting item response patterns were associated with slightly higher empirical Type I error rates. Moreover, 11 of the 15 person-fit methods were somewhat conservative (most empirical Type I error rates are between .01 and .04). For the $T_(uni)$, $T_(sub)$, $T_(com)$, and $T_(sel)$ methods, the empirical and nominal Type I error rates were similar in most conditions (ranging between .04 and .05). The largest Type I error rates (.07) were found for the $T_(uni)$ and the $T_(com)$ methods in the two experimental conditions associated with 10% of misfitting item response patterns.

Detection Rates

The detection rates associated with each person-fit method are now discussed. The analysis is broken down in terms of overall detection rates (i.e., taking into account both global and local misfit) and detection rates of globally and locally misfitting item response patterns.
The overall detection rates were similar for a given proportion of misfitting item response patterns in the sample (\( AbN = .10 \) or \( AbN = .30 \)). In other words, the imputed type of aberrant behavior (spuriously low, spuriously high) had no practical effect on the overall detection rates. Therefore, the results concerning the overall detection rates are presented in terms of factor \( AbN \) only. Figure 1 displays the detections rates discriminated by person-fit method (upper left), the multiple-subtest approach (upper right), and person-fit statistic (lower left). The first observation is that the detection rates decreased from the \( AbN = .10 \) condition to the \( AbN = .30 \) condition. This indicates that increasing the proportion of misfitting response vectors in the dataset actually deteriorated the performance of the person-fit methods, which is not an unusual result (e.g., St-Onge, Valois, Abdous, & Germain, 2011). Moreover, it can be seen that the \( \hat{t}_{(uni)} \) and \( H^T_{(com)} \) methods performed the best (the other methods were grayed out in the figure for simplicity). The \( \hat{t}_{(uni)} \) method performed better for a lower proportion of misfitting item response patterns (\( AbN = .10 \)), and the \( H^T_{(com)} \) method performed better for \( AbN = .30 \). The best multiple-subtest approach across all conditions was the ‘com’ approach, followed by the ‘uni’ approach. Also of great interest is that, on average, the approaches based on the nonparametric \( H^T \) statistic were the best ones (Figure 1, lower-left panel).
FIGURE 1. Overall detection rates (for both globally and locally misfitting item response patterns) discriminated by person-fit method (upper left), multiple-subtest approach (upper right), and person-fit statistic (lower left). The x-axis concerns proportions of misfitting item response patterns in the datasets.

Figure 2 shows that to detect globally misfitting item response patterns, the $l_{z(uni)}^*$ method performed best. As expected, for all person-fit methods, the detection rates increased as the severity of the misfit increased. Moreover, note that the $H_{(com)}^T$ method performed exceptionally well in comparison to $l_{z(uni)}^*$ when $AbN = .30$ (Conditions 2 and 4) and the proportion of item scores that were changed was low to moderate (20–40% of the entire item response vector).
FIGURE 2. Detection rates for globally misfitting item response patterns. Only the two best person-fit methods are identified for simplicity. The x-axis concerns proportions of item scores that were changed to reflect misfit. Condition 1: $AbN = .10$, spuriously low; Condition 2: $AbN = .30$, spuriously low; Condition 3: $AbN = .10$, spuriously high; Condition 4: $AbN = .30$, spuriously high.

The detection of item response patterns reflecting aberrant behavior at the subtest level showed four person-fit methods that performed relatively well: the $t'_{uni}$, $H^T_{(com)}$, $H^T_{(sub)}$, and $H^T_{(sel)}$ methods (Figure 3). As before, $H^T_{(com)}$ performed the best in comparison to the other statistics when $AbN = .30$ (Conditions 2 and 4), whereas $t'_{uni}$ performed the best when $AbN = .10$ (Conditions 1 and 3). The $H^T_{(sub)}$ and $H^T_{(sel)}$ methods performed especially well in detecting misfit within one subtest when all the item scores were changed to reflect aberrant behavior (see Conditions 2 and 4, “One subtest 100%” in the x-axis).
FIGURE 3. Detection rates for locally misfitting item response patterns. Only the four best person-fit methods are identified for simplicity. The x-axis identifies the four types of local misfit that were simulated (one subtest, 50% of the item scores changed; one subtest, 100% of the item scores changed; two subtests, 50% of the item scores changed; two subtests, 100% of the item scores changed). Condition 1: $AbN = .10$, spuriously low; Condition 2: $AbN = .30$, spuriously low; Condition 3: $AbN = .10$, spuriously high; Condition 4: $AbN = .30$, spuriously high.
Discussion

Admission to college or professional-degree programs is (partly) based on entrance and admission tests. Although many studies have shown that these tests have high predictive power for first year achievement, a test score for an individual applicant may be more or less informative. Test providers and admission committees recognize this. For example, the LSAC advises test takers on its website that “if you believe that your test score does not reflect your true ability—for example, if some circumstance such as illness prevented you from performing as well as you might have expected—you should consider taking the test again.” To be able to obtain accurate information about test score validity, sound psychometric methods are needed.

In the present report, we studied different methods that can be used to provide additional information about the total scores. Based on the results of these studies we conclude that a safe choice is to use $I_{z(uni)}$ and $H_T^{(com)}$ because in most simulated cases these statistics outperformed the other statistics. Second, we conclude that when it is expected that there may be a substantial proportion of aberrant response patterns in the data, $H_T^{(com)}$ is a better choice than $I_{z(uni)}$ because the number of aberrant response patterns biases the item parameters and reduces the power of $I_{z(uni)}$. This effect was also found in Karabatsos (2003) for unidimensional tests. A third conclusion is that when researchers expect that aberrant responses will only be shown on one of the subtests, the $H_T^{(sub)}$ and $H_T^{(sel)}$ performed best. Because, in general, researchers do not know what to expect and thus also do not know whether a particular response behavior will occur on any one of the subtests, $I_{z(uni)}$ and $H_T^{(com)}$ are the best choices.

There are two important caveats. First, in this report the between-subtest correlations of the $\theta$ parameters was high ($r = .8$). The choice of this correlation was based on empirical evidence that for a number of admission tests these numerical values are realistic. We checked to see how the results could change in case a lower correlation between the subtests would be more appropriate. We used the value $r = .6$. The results indicated that the Type I error rate associated with the $I_{z(uni)}$ statistic was extremely large (11–16%) in comparison to the 5% nominal error rate. This seems to indicate that the performance of the $I_{z(uni)}$ statistic is especially sensitive to violations of the unidimensionality assumption, unlike the remaining statistics under consideration. This is in line with findings in Albers, Meijer, and Tendeiro (2014). $H_T^{(com)}$ was the best person-fit statistic across all experimental conditions. Second, the current study was based on item discrimination values between 0.5 and 2.0. We then restricted the item discrimination values to between 0.5 and 1.5 and re-ran the analysis. Results showed that the detection rates decreased slightly, and that the $H_T^{(com)}$ statistic performed the best in detecting both
global and local aberrant response patterns. The results derived from both analyses indicate a larger flexibility of the $H^T_{(com)}$ statistic under varying testing conditions.

In this report we used a bootstrap procedure to determine whether a pattern could be classified as normal or aberrant. As our results showed, this procedure could not completely control the Type I errors. Snijders (2001) discussed a theoretical sampling distribution for the $I^r$ statistic that can be used for unidimensional tests. Recently, Albers et al. (2014) investigated the theoretical asymptotic sampling distribution for $I^r$ for tests with multiple subtests. They found that these distributions provide reasonable approximations, even for tests consisting of four subtests of only 10 items each. So in future research, these distributions may be a good alternative to using bootstrap methods.

References


