Detection of Invalid Test Scores

Tendeiro, Jorge N.; Meijer, Rob R.

Published in:
Journal of Educational Measurement

DOI:
10.1111/jedm.12046

Publication date:
2014

JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education
Journal of Educational Measurement
Fall 2014, Vol. 51, No. 3, pp. 239–259

Detection of Invalid Test Scores: The Usefulness of Simple Nonparametric Statistics

Jorge N. Tendeiro and Rob R. Meijer
University of Groningen

In recent guidelines for fair educational testing it is advised to check the validity of individual test scores through the use of person-fit statistics. For practitioners it is unclear on the basis of the existing literature which statistic to use. An overview of relatively simple existing nonparametric approaches to identify atypical response patterns is provided. A simulation study was conducted to compare the different approaches and on the basis of the literature review and the simulation study guidelines for the use of person-fit approaches are given.

When a person is taking an examination or test, total (transformed) test scores are reported to provide information about an examinee's proficiency level. Total scores, may, however, give a false impression of the test taker's proficiency level. In general, test takers may produce invalid test scores due to item preknowledge or item score copying (Belov & Armstrong, 2010; Meijer & Sijtsma, 2001), but they may also produce invalid test scores due to misinterpretation of test questions, or guessing most answers to the test. Large differences between total test scores from repeat test takers or groups of test takers may also point at cheating behavior. Because test results have often far reaching consequences for individuals, test scores should routinely be checked on their validity. There is indeed a trend that large testing companies are starting to *monitor* scale scores via different types of quality control tools, time series techniques, and person-fit scores (Tendeiro & Meijer, 2012). In recent guidelines for the reporting of test scores (e.g., International Testing Committee, 2014) it is recommended that test results should be monitored routinely through statistical techniques for detecting invalid test scores. Also, Olson and Fremer (2013) published a report for the Council of Chief State School Officers in which they advocated using, besides other methods, person-fit statistics to detect irregularities in test behavior. However, these reports contain no specific guidelines advising which statistic or method to use.

The methodological contribution of this study consists of a thorough comparison of the power of different person-fit indices to detect invalid test scores under several testing conditions. On the basis of this simulation study indices that can be used best in practice will be selected. Some recently proposed indices based on detecting strings of item scores will be incorporated, as well as more traditional methods that are sensitive to reversals to the perfect Guttman pattern that were not used in earlier studies. From the existing literature it cannot be deduced how these different nonparametric statistics perform in realistically simulated test data. We hope to make a significant contribution to the literature concerned with monitoring the quality of test scores.

In this study the focus is on *simple* statistical techniques that can be used to investigate the fit of an item score pattern to the majority of item score patterns in the sample. These statistics are often referred to as group-based or nonparametric person-fit indices. The advantage of these statistics is that they are based on observed item scores and total scores, and do not require estimation of parameters as in parametric IRT models. Although there have been several person-fit review studies (e.g., Karabatsos, 2003; Meijer & Sijtsma, 2001), these studies do not incorporate recent developments in this area and from these studies it is unclear for practitioners which indices can best be used to detect test irregularities or invalid test scores. This study may serve as a guideline for practical data forensics using person-fit statistics. The purpose of this study is to provide the reader an overview of some traditional and recently proposed nonparametric fit indices based on different types of residual analyses between observed and expected item scores and to further refine nonparametric person-fit methodology. A comparison with a popular parametric index (l_z^*; Snijders, 2001) is also performed to better contextualize the performance of the nonparametric fit indices. This study has two main goals: (1) to provide an up-to-date overview of nonparametric person-fit research and (2) to provide practitioners with practical guidelines that may help choosing the best analytical approach possible, within the limits of the results obtained from our simulation study. The second goal is especially relevant for empirical applications because literature that discusses nonparametric person-fit indices taking into account various relevant research factors (e.g., test length, item discrimination, type of aberrant behavior, proportion of items and/or respondents providing atypical answers) is surprisingly scarce, as discussed in the next section.

It should be emphasized that the nonparametric person-fit indices discussed in this study are general indices that are not specifically aimed at detecting cheating. Instead, they are useful to detect unexpected score patterns due to one of a possibly wide range of aberrant answering behaviors. Thus, the most important application of individual person-fit indices is to check the *interpretability* of an examinee's proficiency level. If an examinee has an atypical person-fit score, the item score pattern cannot be described through the chosen statistical model and, consequently, it is very difficult to compare examinee's test scores with other test scores in the sample. Moreover, person-fit scores may help interpreting the *type* of aberrant behavior that originated the atypical item score pattern. Such analyses should always be complemented with other sources of information (e.g., seating charts, video surveillance, or follow-up interviews) because it is possible that different types of aberrant behavior lead to similar manifestations of unexpected score patterns.

This article is organized as follows. First, an overview of existing nonparametric person-fit indices is given. Second, the design of our simulation study is discussed in relation to previous findings in the literature. Third, the details of our simulation study are explained, and the major findings from the simulation study are presented. Fourth, the relative effectiveness of nonparametric indices is discussed, as well as a comparison with a popular parametric index.

Existing Nonparametric Person-Fit Approaches

Guttman Model Indices

We did not select all indices proposed in the literature, but selected indices on the basis of earlier review studies (Meijer & Sijtsma, 2001). The focus was given to indices that previous studies have shown to perform relatively well (e.g., Armstrong & Shi, 2009b; Karabatsos, 2003; Meijer, 1994; Meijer, Muijtjens, & van der Vleuten, 1996; Rudner, 1983), but for which a more thorough performance analysis and relative comparison is still lacking in the literature.

Let X_i denote the random variable consisting of the score on a *dichotomous* item i ($i = 1, \ldots, I$). The observed score of person n ($n = 1, \ldots, N$) on item i, that is, a realization of random variable X_i, will be denoted by x_{ni}. The item's proportion-correct score, also known as the item's *difficulty* or p-value, is the proportion of persons who answered the item correctly and is denoted by P_i ($i = 1, \ldots, I$). The p-value of item i is defined by $P_i = \int_\theta P_i(\theta) f(\theta) d\theta$, where $f(\theta)$ is the density of ability θ in the population. P_i can be estimated by the sample's proportion-correct, which is denoted by p_i. Without loss of generality, and unless stated otherwise, it is assumed that the items are ordered in increasing order of difficulty, that is, $p_1 \geq p_2 \geq \cdots \geq p_I$. This simplifies the presentation of the computational formulas for most person-fit indices that will be discussed. Respondent n's response vector and total score will be denoted by $\boldsymbol{x}_n = (x_{n1}, x_{n2}, \ldots, x_{nI})$ and s_n, respectively. The probability of answering item i correctly, conditional on total score, that is, $\text{Prob}(X_i = 1 | S = s)$ is denoted by $p_i(s)$. Furthermore, let $\boldsymbol{p} = (p_1, p_2, \ldots, p_I)$ denote the vector of proportions-correct in the sample.

Sato (1975) proposed the caution index C given by

$$C_n = 1 - \frac{\text{Cov}(\mathbf{x}_n, \mathbf{p})}{\text{Cov}(\mathbf{x}_n^*, \mathbf{p})}, \tag{1}$$

where x_n^* is the so-called Guttman vector containing correct answers for the s_n easiest items (i.e., with the largest p-values) only. C is zero for Guttman vectors and its value tends to increase for response vectors that depart from the group's answering pattern, hence warning the researcher to be *cautious* about interpreting such item scores. Harnisch and Linn (1981) proposed a modified version of the caution index denoted C^* which bounds the caution index between 0 and 1:

$$C_n^* = \frac{Cov\left(\mathbf{x}_n^*, p\right) - Cov(\mathbf{x}_n, p)}{Cov\left(\mathbf{x}_n^*, p\right) - Cov(\mathbf{x}_n', p)},$$

where $\mathbf{x}_n'$ is the reversed Guttman vector containing correct answers for the s_n hardest items (i.e., with the smallest p-values) only. C^* is sensitive to the so-called 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).

Van der Flier (1980; see also Tatsuoka & Tatsuoka, 1982; Meijer, 1994) proposed the (normed) number of Guttman errors as a person-fit index, denoted $U1$. The normalization is done against the maximum number of Guttman errors given the respondent's total score s_n. The formula is given by

$$U1_n = \frac{\sum_{i<j}^{I}(1 - x_{ni})x_{nj}}{(I - s_n)s_n}.$$

Van der Flier (1980, 1982; see also Emons, Meijer, & Sijtsma, 2002) yet proposed an alternative index denoted $U3$ given by

$$U3_n = \frac{f\left(\mathbf{x}_n^*\right) - f(\mathbf{x}_n)}{f\left(\mathbf{x}_n^*\right) - f(\mathbf{x}_n^{'})},$$

where $f(\mathbf{x}_n)$ denotes the summation $\sum_{i=1}^{I} x_{ni} \log(\frac{p_i}{1-p_i})$. Expressions for the expected value and variance of $U3$ were also given by van der Flier (1980, 1982). These formulas are expected to hold under somewhat imprecise conditions (1982, pp. 295–296). However, Emons et al. (2002) showed that the standardized $U3$ index does seem to be problematic because its empirical distribution often deviates from the theoretical one. Similarly to the C index, both the $U1$ and the $U3$ indices are sensitive to Guttman errors (ranging between 0 for perfect Guttman vectors and 1 for reversed Guttman vectors).

A different type of index was introduced by Sijtsma (1986; see also Sijtsma & Meijer, 1992). Sijtsma observed that Mokken (1971) had already introduced an index H_i that allowed assessing the scalability of an item to the Guttman (1944, 1950) model. Sijtsma (1986) used the same index applied to the *transposed* data in order to come up with an index that could detect respondents that would not comply with the Guttman model. Assume, without loss of generality, that the rows of the data matrix are ordered to increasing order of total score s_n ($n = 1, \ldots, N$). The index formula is

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

with $t_n = s_n/I$, $t_m = s_m/I$, and t_{nm} is the proportion of items answered correctly by both respondents n and m (Sijtsma & Molenaar, 2002, p. 57). This index is equivalent to the ratio $\text{Cov}(\mathbf{x}_n, \mathbf{r}_{(n)})/\text{Cov}_{\max}(\mathbf{x}_n, \mathbf{r}_{(n)})$, where $\mathbf{r}_{(n)}$ is the vector of total item scores computed *excluding* respondent n, and the denominator is the maximum covariance given the marginal. Hence, H_n^T is actually similar to Sato's C equation (1). H_n^T is maximum 1 when $t_{nm} = t_n$ ($n < m$) and $t_{nm} = t_m$ ($n > m$); this means that 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_n^T equals zero when the average covariance of the response pattern of respondent n with the other response patterns equals zero. Index H^T was shown to perform relatively well in several simulation studies (Karabatsos, 2003; Sijtsma, 1986; Sijtsma & Meijer, 1992).

Van der Flier (1980, 1982) presented an index referred to as the *probability of exceedance* (PE); see Tendeiro and Meijer (2013) for recent developments of this index. The PE of the observed response vector x_n is determined as the sum of the

probabilities of all response vectors which are, at most, as likely as x_n, conditional on the total score

$$\text{PE}(\mathbf{x}_n) = \sum_{y} \text{Prob}\left(\mathbf{X} = \mathbf{y} | s_n\right), \tag{2}$$

where the probability that random vector $\boldsymbol{X}$ equals the observed response vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_I)$ is defined by

$$\text{Prob}(\mathbf{X} = \mathbf{y}) = \prod_{i=1}^{I} p_i^{y_i} (1 - p_i)^{1 - y_i},$$

and the summation in (2) extends to all response vectors $\boldsymbol{y}$ with total score s_n verifying $\text{Prob}(\boldsymbol{y}) \leq \text{Prob}(\boldsymbol{x}_n)$. Response vector $\mathbf{x}_n$ is considered nonfitting when its PE is smaller than a specified level, either predetermined by the researcher or estimated using data calibration or resampling procedures. The PE index is sensitive to deviances to the performance of the group of respondents as indicated by the estimated p-values. In other words, $\mathbf{x}_n$ is considered aberrant when it does not closely match the expected score pattern that is suggested by the population's p-values. The PE index is especially suited to tests of short or moderated length. In fact, the exact computation of the PE for tests with more than 20 items is unfeasible in practice, because its computation requires a complete enumeration of all response patterns with the same length and total-correct score as the response pattern under inspection. The number of such responses patterns increases quickly with I (it is equal to $\binom{I}{s_n}$). This was one of the motivations that led van der Flier to develop the $U3$ index as an alternative (Meijer & Sijtsma, 1995; van der Flier, 1980, 1982). One alternative to avoid this problem consists of using bootstrapping to estimate suitable sampling distributions. Tendeiro and Meijer (2013) also discuss the possibility of using some asymptotic distribution for this purpose, but the results found up to now were not encouraging.

Emons, Sijtsma, and Meijer (2005) proposed a comprehensive methodology for person-fit analysis in the context of nonparametric item response theory. The methodology (a) included van der Flier's (1982) global person-fit index $U3$ to make the binary decision about fit or misfit of a person's item-score vector, (b) used kernel smoothing to estimate the person-response function for the misfitting item-score vectors, and (c) evaluated unexpected trends in the person-response function using a new local person-fit index.

CUSUM-Based Indices

A family of person-fit indices of a completely different nature is based on cumulative sum (CUSUM) procedures; see Page (1954), van Krimpen-Stoop and Meijer (2000, 2001), Armstrong and Shi (2009a,b), Tendeiro and Meijer (2012), and Tendeiro, Meijer, Schakel, and Maij-de Meij (2013). CUSUM procedures originally arose from the statistical process control field, which covers a range of statistical procedures which allow to control and monitor different types of production processes. A CUSUM (Page, 1954) is a chart which allows following a production process in real time. The process accumulates information observed in prior measurements and

has the ability of detecting a shift in the production process (i.e., an *anomaly*) at early stages. A CUSUM is characterized by *control limits*; these can be lower and/or upper limits, according to the nature of the CUSUM (e.g., one- or two-sided). Once a shift in measurements is big enough and the chart line crosses a control limit an alarm signal is given. At this point production stops, the source of the problem is identified, the problem is eliminated, and afterwards productions is resumed with a reset CUSUM chart.

Some researchers conceived applying the CUSUM technique on the detection of aberrant behavior in the context of educational and psychological testing using item response theory modeling (Bradlow, Weiss, & Cho, 1998; van Krimpen-Stoop & Meijer, 2000, 2001). Researchers anticipated that CUSUMs might be especially sensitive to *local sequences* of aberrant item scores. This type of aberrant behavior is typically not the main concern of the person-fit indices available in the literature, which therefore gives CUSUMs a special role in the person-fit field. The importance of detecting local aberrant behavior is important in many practical situations. For example, respondents with warm-up problems tend to fail more items at the beginning of the test than on other sections. In this case it would be useful to detect a possibly strange score pattern among the first items. Likewise, respondents who invest too much time in order to excel on each and every item might run out of time, and as a consequence might be forced to guess the answer to the last items of the test. In this case, the aberrant behavior is restrained only to the end of the test. CUSUMs are ideal to detect these, and other similar, types of aberrant behavior, as shown in some simulation studies (see Armstrong & Shi, 2009a,b; Tendeiro & Meijer, 2012).

Most research in education and psychology that uses CUSUM procedures is based on parametric item response theory (e.g., Armstrong & Shi, 2009a; Bradlow et al., 1998; Meijer & van Krimpen-Stoop, 2010; Tendeiro & Meijer, 2012; van Krimpen-Stoop & Meijer, 2000, 2001). The only article, to our knowledge, that discusses CUSUM methods nonparametrically is Armstrong and Shi (2009b). In this article the van Krimpen-Stoop and Meijer (2001) approach was adapted to the nonparametric field as follows. Let C^L and C^U denote the lower and upper CUSUM indices, respectively. Lower CUSUMs are typically sensitive to aberrant behavior pertaining to an underperformance of some kind; we shall refer to such type of aberrant behavior as *spuriously low* responding (Rupp, 2013). Upper CUSUMs, on the other hand, are typically sensitive to aberrant behavior that reveals an overperformance of some kind; we shall refer to this type of aberrant behavior as *spuriously high* responding (Rupp, 2013). Start by initializing the CUSUM statistics: $C_0^L = C_0^U = 0$. After administration of item i ($i = 1, \ldots, I$) the CUSUM statistics are iteratively updated as follows:

$$C_i^L = \min\left\{0, C_{i-1}^L + T_i\right\},$$
$$C_i^U = \max\left\{0, C_{i-1}^U + T_i\right\},$$

where C^L and C^U are then given by C_I^L and C_I^U, respectively. The increment T_i is equal to $(X_i - \text{Prob}(X_i = 1|S = s))$, hence it is conditional on the total correct score. T_i is a measure of the difference between the observed and expected score on

item i, conditional on the respondent's total score. T_i is negative (at least nonpositive) whenever an item is answered incorrectly and is positive (at least nonnegative) whenever an item is answered correctly.

A succession of items answered incorrectly will lead to a succession of negative increments T_i, which are accumulated by C^L. In case the lower CUSUM decreases below some *control limit* (to be estimated; Hawkins & Olwell, 1998) the respondent is flagged as having responded spuriously low. Observe that the upper CUSUM C^U is unable to detect this aberrant behavior because it is bounded below by zero. A succession of items answered correctly, on the other hand, will lead to a succession of positive CUSUM increments, which are accounted for by C^U. In case the upper CUSUM increases above some control limit (to be estimated) the respondent is flagged as having responded spuriously high. In this case the lower CUSUM C^L is unable to detect such a respondent because it is bounded above by zero. Summarizing, C^L is tailored to detecting local spuriously low responding whereas C^U is tailored to detecting local spuriously high responding.

The control limits are estimated such that false positives (i.e., falsely detecting inconsistent behavior) are limited by a preselected level α. A common approach is to estimate the control limits using calibration data sets (e.g., Tendeiro & Meijer, 2012). Calibration data sets may be computed from data simulated using the estimated IRT parameters from the real data set. This should only be attempted in case it is believed that the sample parameter estimates were not overly affected by the presence of atypical scores in the data. Alternatively, scores from previous test administrations may be used instead.

It is possible to devise a two-sided CUSUM that is not compromised uniquely to spuriously low or spuriously high responding (Armstrong & Shi, 2009a,b). This index is particularly useful when the researcher is not interested on any type of aberrant behavior in particular. The formula is

$$\mathrm{C}^U_{\max} = \max\left\{C^U_i\right\} \text{ and } \mathrm{C}^L_{\min} = \min\left\{C^L_i\right\}, \qquad i = 1, \ldots, \mathrm{I},$$
$$C^{LU} = \mathrm{C}^U_{\max} - \mathrm{C}^L_{\min}.$$

In this case aberrant response behavior (both spuriously low or high responding) is reflected in an increase of C^{LU}, hence an upper control limit must be estimated in order to come up with a decision rule.

Armstrong and Shi (2009b) proposed an alternative increment statistic T_i to both the lower and the upper CUSUMs ($i = 1, \ldots, I$). Let $p^L_i(s)$ and $p^U_i(s)$ denote alternative conditional probabilities that need to be specified by the researcher and that try to estimate the real probability of answering item i correctly under some "aberrant" type of response behavior. The increments statistics T_i are given by

$$T_i = \ln \frac{p_i\,(s)^{x_i}\,[1 - p_i\,(s)]^{1-x_i}}{p^L_i(s)^{x_i}\left[1 - p^L_i(s)\right]^{1-x_i}} \text{ (lower CUSUM)},$$
$$T_i = \ln \frac{p^U_i(s)^{x_i}\left[1 - p^U_i(s)\right]^{1-x_i}}{p_i\,(s)^{x_i}\,[1 - p_i\,(s)]^{1-x_i}} \text{ (upper CUSUM)}.$$

Increments T_i are defined as log-likelihood ratios. The numerator (resp. denominator) of T_i for the lower (resp. upper) CUSUM is the likelihood of item i's score based on the subsample of respondents with the same total score as the respondent under investigation. This likelihood gives a measure of "normal" behavior. The denominator (resp. numerator) of T_i for the lower (resp. upper) CUSUM, on the other hand, gives a measure of the likelihood of item i's score for respondents that display some kind of aberrant responding behavior Armstrong and Shi (2009b, pp. 415–417) defined $p_i^L(s)$ and $p_i^U(s)$ as quadratic functions of $p_i(s)$. Tendeiro and Meijer (2012) noted that this approach has some drawbacks in the parametric IRT setting, and suggested some improvements that may be extended to the current nonparametric framework.

Which Person-Fit Index Is to Be Preferred in Practice?

Several person-fit indices have been presented up to now. In practice, a researcher has to decide which index to use. The answer is not straightforward. Each index has specific features that can make it more suitable in some circumstances than others. For example, Harnisch and Linn's C^* and van der Flier's $U1$ are sensitive to detecting Guttman errors, that is, seemingly strange response vectors where some easy items were answered incorrectly but some harder ones were answered correctly. CUSUMs, on the other hand, are tailored to detecting unusual (local) sequences of item scores. C^* or $U1$ may be less sensitive to local sequences of unexpected answer behavior than CUSUM-based statistics. There are many factors that can affect the performance of a person-fit index, such as the target population, the length of the test, the difficulty of the items (and the spread of the difficulty across all items), or the proportion of respondents that display aberrant responding behavior, just to mention a few. Perhaps the best way for a practitioner to make an educated choice is twofold:

(1) Take into account the type of aberrant response behavior that is expected (e.g., random response versus strings of unexpected scores) and the characteristics of the scale to be used, like mean item discrimination and test length. An index might perform exceptionally well for a particular scale in a particular population, but may be outperformed by others when used in a different setting.
(2) Consider results from simulation studies that compare performances of several indices. Which indices were shown to perform better under testing conditions similar to the ones in the practitioner's setting? Simulation studies typically assess the performance of person-fit indices on (simulated) real life scenarios. The ability to control the effects of relevant factors on detection and false positive rates helps comparing indices with each other.

Some simulation studies that compared the performance of several nonparametric person-fit indices include Karabatsos (2003), Meijer (1994), Meijer et al. (1996), and Rudner (1983). Rudner (1983) compared the performance of both parametric and nonparametric indices, which included the person biserial correlation (Donlon & Fischer, 1968) and C^*. The manipulated factors included test length ($I = 45, 80$), type of aberrant behavior (spuriously low and high, where both types of aberrant behavior were simulated in each data set), and proportion of items with atypical scores. C^* was reported performing consistently better than other nonparametric indices,

although it was outperformed by some parametric indices. Meijer (1994) compared the performance of $U3$, $U1$, and the nonnormed version of $U1$. It was concluded that factors such as item discrimination, type of aberrant behavior (guessing, cheating), and test length had an effect on the observed detection rates. Moreover, no big differences between the three indices were detected across all experiment cells. Meijer et al. (1996) compared the performances of C, C^*, $U3$, and its standardized version in a simulation study. One peculiarity of this study is that data sets in which all response vectors display aberrant behavior were considered. It was concluded that C^* performed similarly to the standardized $U3$ and better than C, and that detection rates improved with the increase of test length ($I = 17, 33$). Karabatsos (2003) conducted an extensive simulation study involving 36 person-fit indices, 11 of which were nonparametric. Karabatsos considered five types of aberrant behavior (cheaters, creative respondents, guessing, careless, and random respondents), four proportions of respondents that provided aberrant item scores (5%, 10%, 25%, 50%), and three test lengths ($I = 17, 33, 65$). Factors such as item discrimination and the proportion of items with scores displaying aberrant behavior were kept fixed; also, no replications were considered in each cell of the design. It was concluded that H^T was the best index across factors, considering both the nonparametric and parametric indices used in the study. Other nonparametric indices that performed well were C, C^*, and $U3$. It is interesting to observe that four of the five best performing indices were nonparametric. It was also found that detection rates tend to deteriorate when the proportion of aberrant respondents increases in the sample and that detection rates tend to improve as the test length increases.

A simulation study that further extends our current understanding of the performance of nonparametric person-fit indices was conducted. Most of the person-fit indices discussed above were used. The combination of nonparametric indices considered in this study is new. In the simulation study the following topics were investigated:

(1) The effect of several factors on the performance of each index, namely: test length, number of items displaying aberrant behavior, number of respondents displaying aberrant behavior, type of aberrant behavior, and presence of local sequences of unusual item scores. The rates of both false and true positives were taken into account.
(2) The correlations with the total test score. Low correlations (in absolute value) are a good indication that the index measures something different than the total score. This is positive because the total score, which ignores the individual characteristics of each item, only provides an incomplete picture of the answering behavior of an examinee. Furthermore, guidelines that help the practitioner to choose a suitable index in a specific setting are suggested based on the results.

Method

Rupp (2013) conducted a systematic literature review in order to clarify how simulation studies are usually set up in person-fit research. He discussed the many decisions that one needs to make when designing the study. We closely followed suggestions made by Rupp (2013), namely concerning sample size, test length, distributions

to sample parameters from, and proportion of respondents/items with imputed aberrant behavior.

Scores of $N = 1{,}000$ respondents were simulated using the 3PLM (Birnbaum, 1968). Ability parameters were randomly drawn from the standard normal distribution. Item difficulties were randomly drawn from the standard normal distribution constrained to the interval $(-2.5, 2.5)$. Guessing parameters were randomly drawn from uniform distributions in the interval $(0, .2)$.

Item Discrimination

Three intervals of real numbers were independently considered to randomly draw the item discrimination parameters using the uniform distribution: Interval (.5, 1.5) reflecting low item discrimination, interval (1.5, 2.5) reflecting high item discrimination, and interval (.5, 2.5) reflecting a mixture of items with low and high discrimination. The goal was to check whether item discrimination had an effect on the indices' performance; this was expected to be the case based on previous research (e.g., Karabatsos, 2003; Meijer, 1994). Item scores were randomly drawn from the Bernoulli distribution in which the 3PLM was used to compute the probability of answering each item correctly, conditional on the respondent's ability. Perfect response vectors (all 0s or all 1s) were discarded and item scores were generated once more in such a case. The reason was that most indices cannot be computed for perfect score vectors.

Test Length, Proportion of Respondents, and Items With Aberrant Scores

Three test lengths ($I = 15$, 25, and 40) were considered in order to find a possible effect of test length on the performance of the indices. Prior research has shown that test length has typically a large effect on detection rates (Karabatsos, 2003; Meijer, 1994; Meijer et al., 1996; Rudner, 1983). Test with lengths $I = 15, 25, 40$ are referred to as "short," "moderate," and "long" tests respectively, in spite of the seemingly lack of consensus in the literature on this feature (Rupp, 2013). The proportions of respondents for which aberrant scores were simulated (denoted "AbN") had three levels: .05, .10, and .25. The proportions of items for which aberrant scores were simulated (denoted "AbI") also had three levels: .20, .40, and .50. Detection rates are expected to increase as AbI increases (e.g., Rudner, 1983). However, the effect of AbN on detection rates reported in the literature is unclear. Meijer (1994) reported an increase in detection rates when AbI increased from 5.5% to 11%, whereas Karabatsos (2003) reported comparable detection rates for low to moderate AbI proportions (5%, 10%, 25%) but worse detection rates for large AbI (50%). Therefore, it was not clear what to expect in the present study.

Aberrant Response Behavior

Two types of aberrant behavior were independently generated: spuriously low and spuriously high responding. Item scores reflecting spuriously low responding were generated as follows. A proportion of respondents (.05, .10 or .25) with high ability and enough 1s (at least 20%, 40%, or 50% of the response vector) was randomly selected. High ability was defined by theta values above .5, although for some cells

in the design smaller values had to be used in order to have enough respondents available. Then, for each selected respondent, the adequate proportion of 1s was randomly chosen and replaced by scores drawn from a Bernoulli distribution with a .2 probability. Hence, 1s were changed to 0s with 80% probability. Item scores reflecting spuriously high responding were similarly generated, with the difference that respondents had originally low ability (below $-.5$ with occasional exceptions when not enough respondents had been selected) and that 0s were randomly chosen and replaced by scores drawn from a Bernoulli distribution with a .8 probability. The items were randomly chosen from the entire response vector, that is, local aberrant behavior was not taken into account at this point. In order to also consider local aberrant responding (the ideal setting to study the detection rate of the CUSUMs), new data were generated with the added constraint that the items whose scores needed to be generated to reflect aberrant behavior should be consecutive.

Three types of data sets were analyzed depending on the types of atypical respondents that were included. Data sets with only spuriously low aberrant respondents, with only spuriously high aberrant respondents, and with equal proportions of both low and high aberrant responders (spuriously mixed) were simulated. The first two types of data are useful to understand how indices perform when one specific type of aberrant behavior is dominant. The third type of data gives relevant information for cases in which various types of aberrant behavior are present in the data.

Control limits for each nonparametric person-fit index were estimated using a separate calibration data set. Scores of 10,000 "normal" respondents (i.e., with no scores changed to display aberrant behavior) were simulated in the same conditions as described above. For each index, appropriate quantiles of the empirical distribution were computed in each testing condition. A false positive rate of 5% was used in all cases.

Summarizing, the simulation study consisted of a 3 (item discrimination) $\times$ 3 (total number of items) $\times$ 3 (proportion of respondents displaying aberrant behavior) $\times$ 3 (proportion of items displaying aberrant behavior) $\times$ 3 (type of aberrant behavior: spuriously low, high, mixed) $\times$ 2 (type of data: general, CUSUM) completely crossed design. The following nonparametric person-fit indices were considered in our study: C^*, $U1$, $U3$, H^T, PE (only for $I = 15$), and the lower, upper, and two-sided CUSUMs proposed by van Krimpen-Stoop and Meijer (2001). The l_z^* parametric index (Snijders, 2001), which is a corrected version of the popular l_z (Drasgow, Levine, & Williams, 1985), was also computed and used for comparison purposes. Each data set was replicated 100 times. The adequacy of the chosen number of replications was verified by estimating the asymptotic Monte Carlo errors (MCEs; see for instance Koehler, Brown, & Haneuse, 2009) associated to the detection rate of each index, across all experiment factors. It was verified that the MCEs for each person-fit index were never larger than .02. This low error level was deemed adequate for the intended purposes of this study. Moreover, mean detection (i.e., true positive) rates, false positive rates, and correlations between person-fit scores and total scores were averaged across replications. All functions were programmed in R (R version 3.0.1; R Core Team, 2013) and are available from the contacting author upon request. Also, an R package which includes most of the indices used in this study is currently available (PerFit; Tendeiro, 2014).

Results

The General-Type Data

It was checked whether empirical Type I error rates were consistent with the nominal 5% error rate. This was indeed the case with the exception of l_z^* (about 2% across all experiment conditions) and H^T (about 6% across all experiment conditions). Moreover, to investigate the influence of item discrimination, test length, proportion of respondents displaying aberrant behavior, and proportion of items displaying aberrant behavior on the detection rates of each person-fit index, three (for spuriously low, high, and mixed responding) four-way ANOVAs that included all main and second-order effects were conducted. Omega-squared effect sizes were computed and the common thresholds were used (.01, .06, and .14 for small, medium, and large effect sizes, respectively). All interaction effects had effect sizes of no practical importance (omega-squared below .03 in all cases). Also, the proportion of respondents displaying aberrant behavior (factor AbN) had no relevant main effect in all cases, and the proportion of items displaying aberrant behavior (factor AbI) had no practical effect in the spuriously low responding situation, but did have a small to medium effect on the spuriously high and mixed cases, as discussed below. The detection rates associated to the l_z^* index also showed a large effect from both item discrimination and test length, and a medium to large effect of AbI and AbN. More specific details are provided next.

For all indices the item discrimination factor had a large effect on the detection rates. Effect sizes were similar for all the Guttman-type indices (omega-squared between .42 and .43 for spuriously low responding, between .39 and .42 for spuriously high responding, and between .44 and .46 for the spuriously mixed case). The effect of item discrimination on detection rates was less strong for the CUSUM indices and l_z^* (omega-squared between .32 and .39 for spuriously low responding, between .30 and .37 for spuriously high responding, and between .35 and .42 for the spuriously mixed case). Figure 1 further shows that the detection rates increased with the item discrimination for all indices analyzed (the detection rates are averaged over all the other factors). The CUSUM indices are associated with low detection rates (the two-sided CUSUM was the best of the three indices and it is displayed in the plots). Also the PE and the l_z^* indices performed worse than the remaining indices. The apparent poor performance of the PE index is possibly due to the fact that it was only assessed for short tests (i.e., when the total number of items equaled 15). The index that consistently outperformed the other indices is H^T.

The test length factor had a large positive effect on the detection rates of all the Guttman-type indices excluding the PE. The effect for C^*, $U1$, $U3$, H^T was larger in spuriously low and mixed responding cases (omega-squared between .24 and .28) than in the case of spuriously high responding (omega-squared between .21 and .25). The effect for the l_z^* index was slightly lower in comparison (.20, .23, and .17 for spuriously low, mixed, and high responding, respectively). The effect consists of an increase of the detection rates with the test length, in a much similar way as the effect of item discrimination plotted in Figure 1.

Factor AbI had a strong effect on the detection rate of spuriously high responding for all Guttman-type indices (omega-squared between .14 and .18). In this situation,

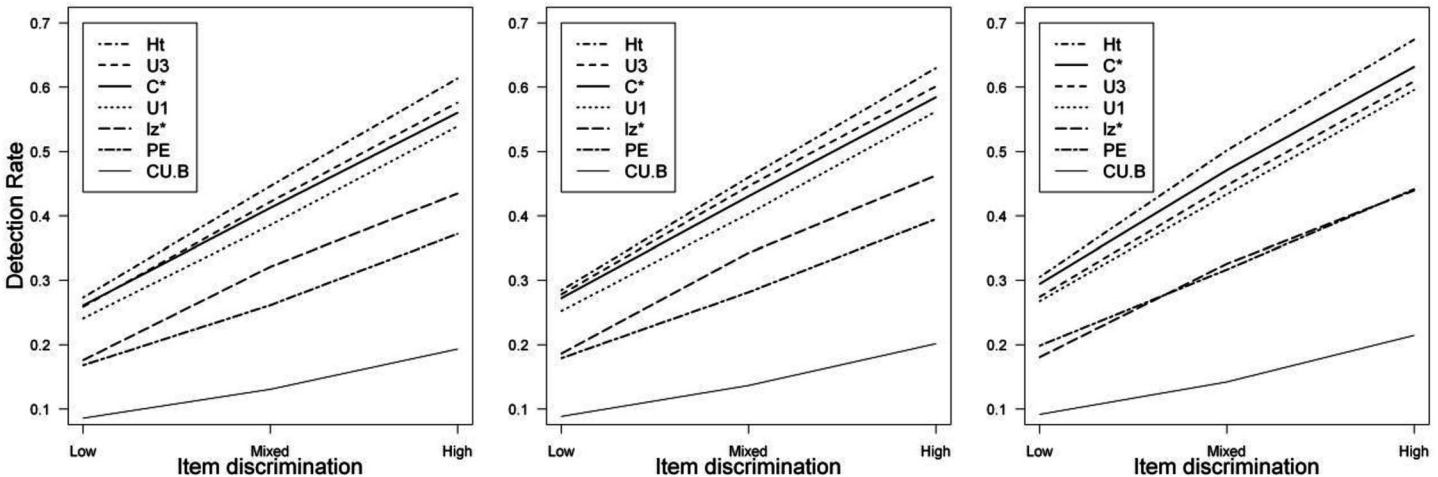

Figure 1. Effect of item discrimination on detection rates for spuriously low (left), spuriously high (right), and mixed (middle) situations for the general-type data.

the detection rates increased when AbI increased from 20% to 40%. Increasing AbI from 40% to 50% had no relevant impact on the detection rates. For the spuriously mixed case the effect of AbI on detection rates was moderate for all indices (omega-squared between .05 and .10).

It can therefore be concluded that the H^T index performed better than the remaining indices across all test lengths and discrimination levels. For 15 items the PE seems to be less powerful than the competing indices. The parametric l_z^* performed worse than several nonparametric indices (C^*, $U1$, $U3$, H^T). This result might also be partly explained by the fact that the empirical Type I error rate associated to l_z^* was lower than the nominal 5% rate, as previously observed. It may also be observed that AbN had a negative moderate effect on l_z^* (omega-squared between .10 and .13), thus the performance of this index seemed to be negatively affected by increasing number of aberrant respondents in the sample.

The CUSUM-Type Data

Similarly to the general-type data, both test length and item discrimination had a large effect on the detection rates. The item discrimination factor had a smaller effect on the detection rate of the CUSUM indices (omega-squared between .14 and .18) when compared to the other nonparametric indices (omega-squared between .30 and .37). Figure 2 shows that the detection rate increased with item discrimination. Although this increase is faster for the Guttman-type indices, the CUSUMs overperformed the former. This observation is especially true when detecting spuriously low aberrant behavior (left panel) and when the items' discrimination is low to moderate (in all cases). When items have high discrimination then some nonparametric indices (in particular H^T) perform similarly to the CUSUMs.

It is clear that the lower CUSUM was the most suited index to detect spuriously low responding whereas the upper CUSUM performed best in the case of spuriously high responding. The two-sided CUSUM was the best index in the spuriously mixed condition and it may therefore be used as a compromise when no specific type of aberrant response behavior is supposed to be predominant. Interesting is that, again, among the Guttman-type indices, H^T is the one that performs best. The fact that H^T performs quite well as item discrimination increases was new to us.

The effect of the test length factor, on the other hand, was particularly strong using CUSUM indices (omega-squared between .32 and .45) when compared to the other nonparametric indices (omega-squared between .20 and .27). This finding is in line with the cumulative feature of a CUSUM: long tests allow accumulating more evidence of aberrant response behavior when it is present in the data. As before, the longer the test the larger the detection rate (the plot displaying this effect resembles Figure 2).

Factor AbI also had a larger effect on the detection rate of the CUSUMs (omega-squared between .20 and .35) when compared to the other nonparametric indices (omega-squared between .06 and .20). It was verified that detection rates increased with the number of items displaying aberrant responding behavior, and that this increase was more pronounced for the CUSUMs.

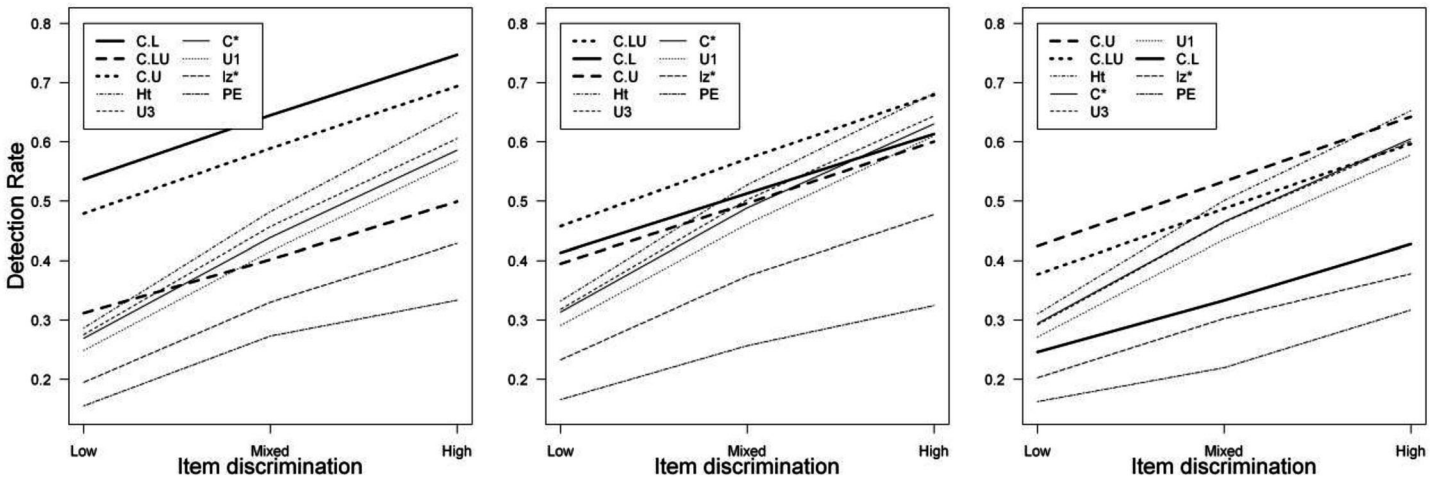

Figure 2. Effect of item discrimination on detection rates for spuriously low (left), spuriously high (right), and mixed (middle) situations for the CUSUM-type data.

The parametric l_z^* performed worse than all other indices except PE. Moreover, the performance of l_z^* deteriorated with the increase of aberrant respondents (omega-squared between .18 and .28).

This study clearly indicates the usefulness of the nonparametric CUSUM technique to detecting local aberrant behavior. The other person-fit indices are less prone to detect this type of behavior. Among the Guttman-type family of indices, the H^T is the one that seems to perform the best.

Comparing Detection Rates With Previous Studies

The detection rates found in this study are comparable to the ones reported in Rudner (1983). Meijer (1994), Meijer, Molenaar, and Sijtsma (1994), and Meijer et al. (1996) found higher detection rates. It should be noted, however, that these simulation studies are based on fundamentally different methodological options (item difficulties were equally spaced, aberrant behavior was differently inputted, and critical values were differently estimated).

Correlations With Total Score

Correlations of each person-fit index with the total score on the final data sets, where aberrant scores have already been inputted, were also analyzed. As discussed before, low correlations show that the index measures something other than the total score by taking individual item characteristics into account. Averaging the correlations of each person-fit index with the total score over all research factors and replications showed that these correlations were typically low, more specifically, below .20 (in absolute value). Only in data sets with highly discriminating items these correlations increased to .25. The increase of test length was associated with larger correlations, whereas the increase of AbN was associated with smaller correlations. Factor AbI did not seem to have an effect on the values of the correlations.

In particular, both H^T and the CUSUM indices did not seem to correlate highly with the total score, hence reinforcing their utility as person-fit tools.

Attempt to Provide Some Guidelines to Practitioners

The large number of person-fit indices in the literature may result in problems of choosing an appropriate index. On the one hand, each index has its own specific features that can make it particularly attractive to detect specific types of aberrant behavior. Our simulation study indeed showed how indices may perform differently across conditions. This is positive because it allows the researcher to optimize the detection rates by carefully choosing one (or more) indices. The disadvantage is that there is some ambiguity concerning the choice of the *best* index to detect a specific type of aberrant behavior, precisely because there are several options available.

In our opinion there is no simple solution to this quandary. As far as we know, there are no easy-to-use guidelines in the person-fit literature. Here we will attempt to fill this gap by providing some useful guidelines. Our choices are based on the findings from our simulation study that we just reported.

Some criteria that do not overly help choosing the best person-fit index are the observed Type I empirical error rates and the low correlations with the total score. All indices showed similar properties according to each of these criteria. Another criterion that was not considered was the availability of sampling distributions. There are some asymptotic results for the *U*3 and the PE indices, but practice showed that these approximations cannot always be trusted (see Emons et al., 2002 for a thorough discussion). Using calibration samples and/or resampling techniques to estimate the sampling distributions (e.g., based on a pilot group) is a possible solution to this limitation (e.g., see Tendeiro et al., 2013, for one possible implementation in practice). The drawback of this approach is that results might be biased in case data are heavily affected by the presence of aberrant scores. We expect that this is only problematic when a large proportion of aberrant response vectors is present in the data. Deriving sampling distributions from simulated data based on the item characteristics of the test (using either parametric or nonparametric IRT models) is also a possibility. Alternatively, one might settle on selecting a prespecified proportion of extreme person-fit index scores (e.g., for 1% of the sample).

The H^T index seemed to perform best in detecting both spuriously low and high responding for general-type data, closely followed by *U3*, *C**, and *U1*. These findings are in line with known literature. Karabatsos (2003) also reported H^T as the best among a large family of both nonparametric and parametric indices. Rudner (1983) identified *C** when compared to two other nonparametric indices, and Meijer (1994) reported similar performances between *U1* and *U3*. Moreover, one of the most interesting results in this study is that H^T (followed by *U3*, *C**, and *U1*) performed well in detecting sequences of aberrant scores (CUSUM-type data) when item discriminations were high. H^T is almost never used in spite of its good performance in detecting different types of aberrant response behavior.

Results concerning CUSUM-type data identified the lower,- upper-, and two-sided CUSUMs as the best indices to detect spuriously low, high, and mixed types of responding, respectively, especially for scales with items displaying low to moderate discrimination. These findings are not theoretically unexpected but similar studies performing comparative analyses of the several CUSUMs under various experiment conditions are surprisingly absent from the literature.

In practice it is often difficult to know in advance which type(s) of aberrant behavior might be present in the data. The spuriously mixed condition was added to our simulation study in order to help addressing this problem. The results showed that, among the Guttman-type indices, H^T (followed by *U3*, *C**, and *U1*) performed best. The two-sided CUSUM was found to be the best CUSUM index to detect spuriously mixed responding.

How Can We Use Person-Fit Statistics in an Educational Context?

Finally, we address how the results of our findings can be used in an educational context. Rupp (2013) provided a framework on how to conduct a person-fit analysis. He distinguished "(1) a statistical detection step and numerical tabulation step," in which item score patterns are classified as normal or aberrant using at least one

powerful person-fit statistic; "(2) a graphical exploration step" in which item response patterns are (graphically) displayed; "(3) a quantitative exploration step" where possible covariates are used to explain aberrant response behavior; and "(4) a qualitative explanation step," where interviews and/or think-aloud procedures are used to explain aberrant response behavior.

Following these steps as much as possible and based on the selection of the statistics provided in the present study, Meijer and Tendeiro (2014) conducted a person-fit analysis on two sections of a high-stakes test. Based on 4,000 archival item score patterns they showed that test takers' score patterns that were classified as misfitting had relatively low scores, which may point at extensive guessing. Although they did not find different inconsistent test-taking behavior between male and female test takers, they did find significant differences in person-fit values for test takers whose first language was not English as compared to other groups of test takers distinguished by the testing company. This example showed that person-fit indices provide useful information that may be used to enhance the interpretation of test scores.

Discussion

In several manuals and guidelines with respect to educational and psychological testing it is recommended to check the data quality at the person level (e.g., Olson & Fremer, 2013). As discussed in this study, there are different approaches that can be used to flag response patterns that are inconsistent with respect to the expected pattern. In this study a number of existing group-based or nonparametric person-fit indices that can be applied without fitting a parametric IRT model were presented. Indices that were sensitive to both the number of Guttman errors as well as to strings of correct and incorrect scores were discussed. The advantage of nonparametric indices is that they are not based on strict model assumptions concerning the data and are very easy to calculate. Because from earlier research it is unclear which indices are most powerful under varying testing conditions, a simulation study was conducted. Indices were compared with respect to the power to detect misfitting response vectors using simulated data. Some guidelines that may help practitioners to choose between the different indices were provided; these guidelines are based on both the current study as well as on prior research). In general the H^T index, followed by *U3*, *C**, and *U1*, seemed to lead to the highest power to detect misfitting response vectors for spuriously high, low, and mixed scoring persons. For strings of 0-scores and 1-scores CUSUM indices had higher power than Guttman-based indices. Test length, item discrimination, and proportion of aberrant scores in the response vector had moderate to large effects on detection power (depending on the experiment condition). The parametric l_z^* index was outperformed by most of the nonparametric indices in the simulation study.

The main limitation of this study relates to using simulated data. The methodological options followed in the study design were chosen as to better approximate real test settings. It is observed that it would be extremely difficult to conduct a similar study based on real data because of the lack of control over the various factors of

interest (such as the proportion of aberrant respondents or the type of misfit). Results reported in this article may be generalized to real test settings to the extent that the real testing conditions approximate (some of the) simulated conditions.

Other limitation concerns the decision on the course of action to take in case a respondent is flagged as (potentially) aberrant. Marking respondents by means of an extreme value of a person-fit index may not serve as proof that some kind of aberrant behavior did take place, nor it clarifies which type of aberrant behavior has been identified. Person-fit indices should be complemented with other sources of information (e.g., seating charts, video surveillance, or follow-up interviews). A nice example was given by Meijer et al. (2008).

As a general rule, practitioners are advised to carefully choosing the person-fit indices that best suit their analyses. It is important to acknowledge that different indices may have a different sensibility to detect aberrant behavior under various testing conditions. Thus, it is recommended that person-fit values always be considered with an eye toward the complete testing setting.

References

Armstrong, R. D., & Shi, M. (2009a). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement*, *33*, 391–410. doi:10.1177/0146621609331961

Armstrong, R. D., & Shi, M. (2009b). Model-free CUSUM methods for person fit. *Journal of Educational Measurement*, *46*, 408–428. doi:10.1111/j.1745-3984.2009.00090.x

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback-Leibler divergence and K-Index. *Applied Psychological Measurement*, *34*, 379–392. doi:10.1177/0146621610370453

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, *93*, 910–919. doi:10.1080/01621459.1998.10473747

Donlon, T. F., & Fischer, F. E. (1968). An index of an individual's agreement with group-determined item difficulties. *Educational And Psychological Measurement*, *28*, 105–113. doi:10.1177/001316446802800110

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67–86. doi:10.1111/j.2044-8317.1985.tb00817.x

Emons, W. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement*, *26*, 88–108. doi:10.1177/0146621602026001006

Emons, W. M., Sijtsma, K., & Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychological Methods*, *10*, 101–119. doi:10.1037/1082-989X.10.1.101

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150. doi:10.2307/2086306

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Claussen (Eds.), *Measurement and precision* (pp. 60–90). Princeton NJ: Princeton University Press.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement*, *18*, 133–146. doi:10.1111/j.1745-3984.1981.tb00848.x

Hawkins, D. M., & Olwell, D. H. (1998). *Cumulative sum charts and charting for quality improvement*. New York, NY: Springer.

International Test Commission. (2014). ITC guidelines for quality control in scoring, test analysis, and reporting of test scores. Retrieved February 25, 2014, from http://intestcom.org.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement In Education*, *16*, 277–298. doi:10.1207/S15324818AME1604_2

Koehler, E., Brown, E., & Haneuse, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulated-based statistical analyses. *The American Statistician*, *63*, 155–162. doi:10.1198/tast.2009.0030

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, *18*, 311–314. doi:10.1177/014662169401800402

Meijer, R. R., Egberink, I. L., Emons, W. M., & Sijtsma, K.(2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's Self-Perception Profile for Children. *Journal of Personality Assessment*, *90*, 227–238. doi:10.1080/00223890701884921

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*, 111–120. doi: 10.1177/014662169401800202

Meijer, R. R., Muijtjens, A. M., & van der Vleuten, C. M. (1996). Nonparametric person-fit research: Some theoretical issues and an empirical example. *Applied Measurement in Education*, *9*, 77–89. doi:10.1207/s15324818ame0901_7

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement In Education*, *8*, 261–272. doi:10.1207/s15324818ame0803˙5

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135. doi:10.1177/01466210122031957

Meijer, R.R., & Tendeiro, J. N. (2014). The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us. Law School Admission Council, Research report 14-03.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2010). Detecting person misfit in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 315–329). New York, NY: Springer.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Olson, J., & Fremer, J. (2013). *TILSA test security guidebook: Preventing, detecting, and investigating test security irregularities*. Washington, DC: Council of Chief State School Officers.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, *41*, 100–115. doi:10.1093/biomet/41.1-2.100

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement*, *20*, 207–219. doi:10.1111/j.1745-3984.1983.tb00200.x

Rupp, A.A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3–38.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo, Japan: Meishi Tosho.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitative Methoden*, *7*, 131–145.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, *16*, 149–157. doi:10.1177/014662169201600204

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331–342. doi:10.1007/BF02294437

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics*, *7*, 215–231. doi:10.2307/1164646

Tendeiro, J. N. (2014). PerFit: Person Fit (R package version 1.1/r4). Available at http://R-Forge.R-project.org/projects/perfit/

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, *36*, 420–442. doi:10.1177/0146621612446305

Tendeiro, J. N., & Meijer, R. R. (2013). The probability of exceedance as a nonparametric person-fit statistic for tests of moderate length. *Applied Psychological Measurement*, *37*, 653–665. doi:10.1177/0146621613499066

Tendeiro, J. N., Meijer, R. R., Schakel, L., & Maij-de Meij, A. M. (2013). Using cumulative sum statistics to detect inconsistencies in unproctored Internet testing. *Educational and Psychological Measurement*, *73*, 143–161. doi:10.1177/0013164412444787

van der Flier, H. (1980). *Vergelijkbaarheid van individuele testprestaties [Comparability of individual test performance]*. Lisse, The Netherlands: Swets & Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, *13*, 267–298. doi:10.1177/0022002182013003001

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2000). Detection of person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 201–219). Boston, MA: Kluwer-Nijhoff.

van Krimpen-Stoop, E. A., & Meijer, R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*, 199–218. doi:10.3102/10769986026002199

Authors

JORGE N. TENDEIRO is Assistant Professor, Department of Psychometrics and Statistics, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands; j.n.tendeiro@rug.nl. His primary research interests include item response theory in general, with a focus on person-fit analysis.

ROB R. MEIJER is Professor, Department of Psychometrics and Statistics, Faculty of Behavioural and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands; r.r.meijer@rug.nl. His primary research interests include the theoretical development of item response theory and the application of item response theory in psychological and educational measurement.