When data are collected through questionnaires or tests, respondents may produce invalid test scores as a result of, for example, lack of motivation, not understanding particular items, or faking. These types of behavior will often result in aberrant, inconsistent, or unexpected item response patterns. Therefore, in several guidelines for the quality of tests, such as in those of the International Test Commission (2013, p. 23), it is advised to “check aberrant or unexpected response patterns, (e.g., when difficult items are answered correctly and easy items incorrectly).” Also, Olson and Fremer (2013) discussed several data forensic methods to check for invalid test scores and advised to check for inconsistent response behavior.

Reynolds (2010) in his editorial for Psychological Assessment emphasized the importance of checking the validity of individual test scores. He stated that “reliability refers to test scores, not tests, and validity refers to the accuracy and appropriateness of test score interpretations—again, not to tests” (p. 3, italics added) and “each interpretation attributed to a test score must be validated: Tests are neither valid nor invalid; only the interpretations of performance on the test are the proper subject of validation research” (p. 3).

Although in personality and clinical assessment the test score validity is sometimes checked for each individual test score through validity scales, in practice, checking the validity of individual test scores, be it in typical or maximum performance testing, is still not popular.

In this tutorial, we discuss the principles of person-fit statistics that are used to check the validity of test scores through the alignment between a test model and individual observed data. The test model may be a parametric item response theory (IRT) model, such as the two- or three-parameter logistic model (Embretson & Reise, 2000), or a nonparametric IRT model (Sijtsma & Molenaar, 2002), but it may also be a factor model or a set of items that are selected using classical test theory. We will not discuss the details of different statistics that are available in the literature. Instead, we give references to studies that the reader can consult for more details of individual statistics and for a more detailed overview we refer to Meijer and Sijtsma (2001). In general, the statistics we discuss in this article can be applied as long as the data are fairly unidimensional.

There are many, mostly technical, papers written about these statistics by specialists (for an overview, see Meijer & Sijtsma, 2001). What is lacking is a tutorial for researchers and clinicians, that is, for those who are mainly interested in applying these statistics to analyze their data. In this article we would like to fill this gap. We first provide an empirical example that illustrates the central idea of the use of these statistics; second, we discuss the principles behind these statistics on the basis of simple group-based statistics and through the alignment between a test model and individual observed data. The test model may be a parametric item response theory (IRT) model, such as the two- or three-parameter logistic model (Embretson & Reise, 2000), or a nonparametric IRT model (Sijtsma & Molenaar, 2002), but it may also be a factor model or a set of items that are selected using classical test theory. We will not discuss the details of different statistics that are available in the literature. Instead, we give references to studies that the reader can consult for more details of individual statistics and for a more detailed overview we refer to Meijer and Sijtsma (2001). In general, the statistics we discuss in this article can be applied as long as the data are fairly unidimensional.

A Practical Guide to Check the Consistency of Item Response Patterns in Clinical Research Through Person-Fit Statistics: Examples and a Computer Program

Rob R. Meijer¹, A. Susan M. Niessen¹, and Jorge N. Tendeiro¹

Abstract
Although there are many studies devoted to person-fit statistics to detect inconsistent item score patterns, most studies are difficult to understand for nonspecialists. The aim of this tutorial is to explain the principles of these statistics for researchers and clinicians who are interested in applying these statistics. In particular, we first explain how invalid test scores can be detected using person-fit statistics; second, we provide the reader practical examples of existing studies that used person-fit statistics to detect and to interpret inconsistent item score patterns; and third, we discuss a new R-package that can be used to identify and interpret inconsistent score patterns.

Keywords
validity of test scores, person fit, item response theory, data forensics

¹University of Groningen, Groningen, Netherlands

Corresponding Author:
Rob R. Meijer, Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, Netherlands.
Email: r.r.meijer@rug.nl
we present existing research that illustrates how person-fit scores are related to psychological background variables, including the clinical implications in applied assessment. Finally, we discuss a recently developed R-package that can be used to calculate the statistics and that provide plots to diagnose the configurations of item score patterns. To be able to do this, a researcher should have access to the item scores of test takers, patients, or clients. With the popularity of computer-based assessment an increasing number of researchers will have access to such data.

We think that one of the reasons that these statistics are not often used in practice is that it is difficult for researchers to find software to calculate these statistics. Another reason is that many papers are too technical to understand for nonspecialists. Hopefully this tutorial will serve this researcher and other practitioners to apply these techniques more often.

**An Illustrative Example**

We start to illustrate the principles of person-fit statistics using the Physical Functioning scale of the SF-36 (Ware & Sherbourne, 1992). We use this scale because it is a well-known and often-used scale and we use it for didactic reasons: The psychometric properties of this scale are ideal to explain person fit (high discriminating items and large spread between the item difficulties). However, to further illustrate the use of person fit for psychological constructs, we will also provide results for an “Inadequacy” scale (Luteijn, van Dijk, & Barelds, 2005) at the end of this article.

In Table 1 we depicted the item content, mean item score, and examples of three item score patterns on the 10 items of the well-known Physical Functioning scale. The items are scored from 0 through 2 (0 = limited a lot; 1 = limited a little; 2 = no, not limited at all) and our results are based on a sample of 714 persons (see Meijer, Tendeiro, & Wanders, 2015, for a more detailed description). Meijer et al. (2015) showed that the data predominantly measured one factor and that the total score scale can be used to give an overall impression of Physical Functioning. In Table 1 the items are ordered according to decreasing mean item score in this sample, that is, to increasing amount of physical effort that is needed to fulfill the task. For each person who fills out this questionnaire, we expect that, when an item that requires a more difficult task is endorsed, such as “walking more than a mile,” an “easier” item like “walking one block” is also endorsed. In practice, this will not always be the case because responses to items are probabilistic; for most persons there will be some reversals or errors (known as Guttman errors, i.e., answering a more “difficult” or less popular item correctly and an “easier” or more popular item incorrectly), but many errors may result in a total score that is difficult to interpret. For example, consider in Table 1 the item score patterns (items in order of increasing difficulty) of Person 493 [1 1 1 1 1 1 0 0 0] resulting in a total score of 7 and that of Person 560 [0 0 0 0 1 1 0 1 2 2], also with a total score of 7. Although both persons have the same total scores, when we consider the configuration of the item scores, the score pattern of Person 493 seems to be in line with the popularity of the items (more 0 scores on the items that require more strenuous tasks). However, it is clear that the score pattern of Person 560 is more difficult to interpret (i.e., aberrant from what is expected), because relatively easy physical tasks are difficult to conduct for this person (0 scores), whereas more difficult tasks are easier to conduct (1 and 2 scores). As a result, the total score for this person is difficult to interpret.

There may be different underlying mechanisms that cause aberrant response behavior, and below we provide examples of studies that investigated the reason of aberrant response behavior. However, irrespective of the underlying mechanisms, it is clear that the test scores of persons with unexpected answering behavior are difficult to interpret. With 714 score patterns in this sample, it is difficult to inspect every item score pattern and decide by eye-ball inspection only, which patterns are unexpected and need closer inspection. Therefore, several statistics have been proposed that provide information about how unexpected an item score pattern is. Sometimes also sampling distributions are known for these statistics under a test model, which can help a researcher to decide when to classify an item score pattern as normal or aberrant.
Almost all person-fit statistics are sensitive to the (weighted) number of Guttman errors. Consider a test consisting of 10 dichotomously scored items ordered from easy to difficult according to their item proportion-correct score. If a test taker has a total score of 6, then we expect that he or she answers the six easiest items correctly and the most difficult items incorrectly. When there is a reversal of item scores, that is, when the most difficult item is answered correctly and the easier item incorrectly, this is counted as an error. Many person-fit statistics are sensitive to the number of these errors. In general, the more errors, the more aberrant the pattern is. When we, again, assume that items are ordered from easy to difficult then a pattern \([1, 1, 1, 1, 1, 0, 0, 0, 0, 0]\) contains zero errors, whereas the pattern \([1, 1, 0, 1, 0, 1, 1, 0, 0, 1]\) contains 9 Guttman errors, because there are 9 \((0,1)\) item pairs.

**Polytomous Items**

For polytomous items, like the items in Table 1, Guttman errors are calculated based on the number of persons that take a so-called item step. An item step is the imaginary step that a person takes when deciding to answer in a subsequent answer category. For example, consider Figure 1 where we depicted the item steps for two items “walking one block” (Item PF9) and “walking more than a mile” (Item PF7). Remember that these items have three answer categories scored 0, 1, and 2. In this case, there are two item steps: The first step is the step from 0 to 1 (and thus answering 1), and the second step is the step from 1 to 2 (answering 2). In a sample we can first determine, for each item, the number of persons that answered in each category. For Item PF9 this is: 34 persons have a 0 score, 90 persons have a 1 score, and 590 have a 2 score (see Figure 1). On the basis of these numbers we can calculate the number of persons that took the first step. For Item PF9, this is the total number of persons, 714, minus 34 persons that chose 0 = 680 (for Item PF7: 714 – 115 = 599). The number of persons that took the second step for Item PF9 equals 714 – 34 = 590 (for Item PF7: 714 – 115 – 175 = 424). As can be seen in Figure 1, there are fewer persons that took the second step for the item “walking more than a mile” that reflects having no difficulty with a more difficult physical task than for the item “walking one block.” Now, if we determine the number of persons that take each item step for all items, we can order the item steps across all items from easy to difficult. For each person, we can then compare his/her item scores with the item step ordering. Large differences point at aberrant response behavior. To further illustrate this, consider the pattern of item steps taken by person 493 (in order of increasing difficulty of the item steps): \([1 1 1 1 0 1 1 0 0 0 0 0 0 0 0]\), with 0 representing an item step not taken and 1 representing item step taken. Again, this person produces a normal response pattern with only a few Guttman errors on the item step level: There are three \((0,1)\) item pairs. In contrast, the pattern of item steps taken by person 24 equals \([0 1 1 0 0 1 0 0 1 1 0 1 1 0 1 1]\). This is an unexpected pattern with 52 Guttman errors. It is important to realize that, although the items steps are used for the calculation of the person fit statistics, for the interpretation of a misfitting item score pattern, the different ordering of the item steps is of not much use. Interpretation should be done on the item level. Although there are differences in the exact way the different person-fit statistics are calculated, this is the basic principle behind most statistics for polytomous item scores.

Some statistics are normed against the perfect Guttman patterns, that is, the statistic equals 0 when it perfectly fits the Guttman model and it equals 1 when it perfectly fits the reversed Guttman pattern. Thus, the higher the person-fit score, the more aberrant the item score pattern. Sometimes a statistic is normed against the maximum number of possible Guttman errors given the total score. In Table 1 we give the numerical values for the \(\hat{G}_N^p\) statistic (Emons, 2008; the superscript “\(p\)” denotes “polytomous” and the subscript “\(N\)” denotes “normed”), which is normed against the maximum number of Guttman errors. Some statistics follow a standard normal distribution for a large number of

\[\text{Figure 1. Illustration of item steps for two items of the PF scale.}\]
items, which eases the interpretation of the numerical values of these statistics. For example, for the $I^p$ statistic (Drasgow, Levine, & Williams, 1985) values smaller than $-1.96$ are often interpreted as indicating unexpected response behavior at a Type I error rate of 5%. This specific asymptotic approximation does not always work as expected; an adjusted computational formula is available when the item scores are dichotomous (the so-called $I^p$ statistic; Snijders, 2001; see also Magis, Raîche, & Béland, 2012, for an accessible overview). The $I^p$ statistic (and its dichotomous corresponding statistic, $I^p$) evaluates the likelihood of a score pattern under an IRT model. Thus, to apply these statistics, it is assumed that the data can be described by an IRT model (suited to either dichotomous or polytomous items; see Embretson & Reise, 2000, for an overview of both types of models). In Table 1 we also provide values for the $I^p$ statistic. As can be seen Person 493 has plausible responses, resulting in a low value for $G^p$ and a positive value for $I^p$. Persons 560 and 24 provide very unlikely responses, and as a result there are extreme values for both $G^p$ and for $I^p$. For example, it is strange that Person 24 has trouble climbing one flight of stairs (Item PF5) but not several flights of stairs (PF4).

Thus, person-fit statistics can be used as descriptive statistics, where misfit can be directly defined in relation to the other patterns in the sample (such as for the $G^p$ statistic) or to an IRT model (such as the $I^p$ statistic). In the first type of application, the question is answered: How improbable is a score pattern compared with the other persons in the sample? In the second case, the question is answered: How likely is a score pattern given the assumed IRT model? Thus, to classify an item score pattern as aberrant researchers can take, for example, the 5% or 10% most aberrant response patterns in the sample, or they may use a theoretical sampling distribution and determine in advance a Type I error rate. For most person-fit statistics the sampling distribution is unknown. Then, resampling methods can be used to obtain such a distribution (e.g., see de la Torre & Deng, 2008, for an approach based on the $I^2$ statistic). In general, it is difficult to provide rules-of-thumb for choosing a Type I error rate. However, because falsely rejecting the null-hypothesis in most applications of person-fit research will not have large consequences (because almost always additional information will be obtained), choosing a 5% or 10% Type I error rate is a good choice.

We would like to stress that extreme person-fit scores indicate that an item score pattern is very unlikely and that, by definition, the researcher should be very careful when interpreting such total scores. The reason is that, in such cases, a person’s item score pattern cannot be scaled with respect to the score patterns of other persons. Having said this, it is of course interesting to study what type of underlying mechanisms may determine aberrant response behavior. Meijer (1996) mentioned different types of possible reasons for aberrant response behavior: Sleeping (trouble getting started, and only answering items according to his/her trait value after having answered a number of items), guessing, cheating, alignment errors, plodding (working very methodologically and as a result generate almost perfect Guttman item response patterns), extremely creative response behavior, and deficiency of subabilities. However, not every type of possible aberrant response behavior could be identified empirically.

### What Does an Extreme Person-Fit Score Mean?

Below, we discuss several studies that show that there is a relation between person-fit scores and different types of underlying test behavior.

#### Lack of Reading Skills or Interpretation Problems

Meijer, Egberink, Emons, and Sijsma (2008) analyzed data from Harter’s (1985) Self-Perception Profile for Children in a sample of children ranging from 8 through 12 years of age. They argued that, for some children, the scale scores should be interpreted with care and caution. Through combined information obtained from extreme person-fit scores, observational data, interviews, and theory about self-concept they showed that for some children item scores did not adequately reflect their trait level due to a less developed self-concept and/or problems understanding the meaning of the questions. Also, Meijer and Tendeiro (2014) found inconsistent response patterns for racial/ethnic groups in a high-stakes educational test. This was the result of insufficient knowledge of English.

#### Idiosyncratic Response Behavior or Traitedness

Reise and Waller (1993) and Ferrando (2012) discussed a general form of inconsistent response behavior when analyzing personality questionnaires where for some respondents the trait as measured by the items does not apply. This type of response behavior may lead to an idiosyncratic interpretation of the items, and test-takers may endorse less popular items and reject easier items.

#### Test Taking Motivation

Ferrando (2012) found a group of students filling out a Neuroticism and Extraversion scale (for research purposes) that were low in what he called “person reliability,” that is, the test-takers were largely insensitive to the ordering of the items. As a consequence, the resulting response patterns were almost random. This was probably due to unmotivated test behavior. In a recent study, Niessen, Meijer, and Tendeiro (2014) found that when collecting data for research
purposes some students’ response patterns were also more or less random as a result of unmotivated test behavior, resulting in extreme person-fit scores. Also Schmitt, Cortina, and Whitney (1993) showed using hierarchical regression analysis that person-fit scores added information to criterion scores in a validation sample. When they removed persons with extreme person-fit scores validities increased.

**Personality and Pathology**

Conijn (2013) investigated consistency of response behavior on the Outcome Questionnaire-45 and found that patients having more severe distress and patients with psychotic disorders, somatoform disorders, and substance-related disorders were particularly likely to show misfit. Conijn (2013) concluded that person-fit analysis has potential in outcome measurement for subgroups that are at risk of producing invalid results. Conrad et al. (2010) screened for atypical suicide risk for persons enduring alcohol and drug treatment. They found that extreme person-fit scores indicated patients that endorsed suicidal ideation items, whereas these persons did not endorse items that reflect less severe forms of depressed feelings.

Other researchers like Meijer and van Krimpen-Stoop (2001) and Woods, Oltmanns, and Turkheimer (2008) found that males produced more inconsistent responses than females. Schmitt, Chan, Sacco, McFarland, and Jennings (1999) and Ferrando (2009) found that respondents low on conscientiousness generated less consistent response patterns than test takers high in conscientiousness.

Zickar and Drasgow (1996) analyzed data in which respondents were instructed to answer honestly or to fake the good answer on a personality scale. They found that using the person-fit approach a larger number of faking respondents were detected at low false positive rates than using a social desirable scale.

Recently, Wardenaar, Wanders, Roest, Meijer, and de Jonge (2014) investigated the usefulness of person-fit statistics to answer the question whether observed association between depression and acute coronary syndromes are overestimated due patients’ tendency to endorse mainly somatic items on depression scales. Wardenaar et al. (2014) using the Beck Depression Inventory (BDI) data found that somatic items were most often endorsed by the majority of persons and that atypicals patients scored higher on the depressive mood/cognitions (e.g., “sense of failure”) and lower on somatic items. The clinical implication is that person-fit statistics do tell us that (a) relatively high scores on the BDI are due to high scores on somatic complaints and (b) BDI scores from atypical respondents should be interpreted differently than BDI scores from typical respondents. That is, for typical respondents the association between depression and acute coronary syndromes seems to be overestimated, whereas for atypical respondents it is not. Accounting for interpersonal differences in item responding can help to improve the validity of depression assessments in psychosomatic research.

<table>
<thead>
<tr>
<th>Table 2. Person-Fit Statistics Available in the R Package PerFit.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics</td>
</tr>
<tr>
<td>Nonparametric</td>
</tr>
<tr>
<td>r.pbis</td>
</tr>
<tr>
<td>C</td>
</tr>
<tr>
<td>G, G_n</td>
</tr>
<tr>
<td>A, D, E</td>
</tr>
<tr>
<td>U3, ZU3</td>
</tr>
<tr>
<td>C*</td>
</tr>
<tr>
<td>NC1</td>
</tr>
<tr>
<td>Hf</td>
</tr>
<tr>
<td>Gf</td>
</tr>
<tr>
<td>G_N</td>
</tr>
<tr>
<td>U3p</td>
</tr>
<tr>
<td>Parametric</td>
</tr>
<tr>
<td>l_z</td>
</tr>
<tr>
<td>l_z^p</td>
</tr>
<tr>
<td>l_z*</td>
</tr>
</tbody>
</table>

**Which Person-Fit Statistic to Use?**

There are numerous person-fit statistics. In Meijer and Sijtsma (2001), an overview is given of different statistics. There are also studies in which the power of different statistics is compared using simulated data. Karabatsos (2003) concluded that the $H$ coefficient had the highest power given a controlled false positive rate. An interesting conclusion from his study was that group-based statistics like $H$, $C$, and $U3$ (for references, see Table 2) outperformed parametric statistics like $l_z$. A researcher can calculate these statistics using the program PerFit (Tendeiro, 2015) described below. Also, Tendeiro and Meijer (2014) recently compared different group-based statistics for dichotomous item scores. They concluded that, given a fixed Type I error rate of .05 in general $H$, followed by $U3$, and $C$ had the highest power to detect misfitting response vectors.

Another criterion to select a person-fit statistic is whether there is a theoretical sampling distribution available. For
dichotomous data and parametric IRT analyses, the $l_z$ statistic is to be recommended because of the availability of clearly interpretable cutoff scores, that is, for $l_z$ its distribution has a better approximation to the assumed standard normal distribution than the $l$ statistic. Thus, for $l_z$ the Type I error rate is easier to control than for other statistics. For polytomous items and long tests also $l_z$ is most often used, probably because of the easy interpretation.

Drasgow, Levine, and McLaughlin (1991) proposed a multiple-subtest extension of $l_z$ denoted $l_{zm}$. Recently, Conijn (2013) compared the use of $l_z$ in combination with $l_{zm}$ and found that for noncognitive data a combination of both statistics resulted in the highest power for a fixed Type I error rate.

**Strategy to Detect and Interpret Aberrant Response Patterns**

Rupp (2013), following Meijer and Sijsma (2001), recently formulated some guidelines on how to perform a person-fit analysis. He suggested that each person-fit analysis should contain “(1) a statistical detection step,” where the scores on at least one person-fit statistic are computed; “(2) a numerical tabulation step,” where the item score patterns are shown; “(3) a graphical exploration step” to display the item response patterns; “(4) a quantitative exploration step” that takes into account possible covariates; and “(5) a qualitative explanation step,” such as think-aloud protocols or interviews that may help to explain aberrant response behavior. Using the R-package discussed below, Steps 1 through 3 can be performed easily. With respect to Steps 4 and 5, it depends on the type of application envisaged by the researcher whether these steps should and can be taken. In Step 4 scores on other tests can be incorporated in the analysis. As we discussed above, Ferrando (2009), for example, found that persons low on conscientiousness were more likely to produce inconsistent response patterns than persons scoring high on conscientiousness; other researchers found differences between males and females (e.g., Meijer & van Krimpen-Stoop, 2001). In Step 5, a clinician may be able to discuss extreme person-fit scores (aberrant response behavior) with a client or with other informants. A clinical interview may provide such a setting. However, Steps 4 and 5 are sometimes difficult to realize or may not always be of interest. For example, removing very unlikely item score patterns from unmotivated students that fill out a questionnaire for research purposes may be appropriate without giving an additional explanation. However, as Meijer and Tendeiro (2014) pointed out, in high-stakes educational testing it is not always easy to withhold test-takers a test score on the basis of person-fit scores (or any other “statistical evidence”). In these contexts, however, it may help understand particular types of scores. For example, as discussed above, Meijer and Tendeiro (2014) found that several test takers with language problems produced unexpected response patterns as a result of guessing and thus scored low on the test.

**How Should We Manage Unexpected Score Patterns and Clinical Implications of Person-Fit Scores?**

An individual’s total score that is based on very unlikely item scores is difficult to interpret. Person-fit statistics thus function as a validity index for individual test scores. How to proceed with a person who produces an unlikely score pattern? This depends on the test setting. In an interesting article, Conrad et al. (2010) showed that through the use of person-fit statistics, persons could be identified with an atypical subtype at high risk for suicide that did not show typical depression and other internalizing disorders. Clinically, this is interesting information that adds additional information to the total score and that may help clinicians become more sensitive to the additional information that can be retrieved from a total score. In this case, additional information can be obtained in a clinical interview. Also, the study by Wardenaar et al. (2014) showed that BDI scores should be interpreted carefully and that total scores have different meanings for different persons. As another example, with the increasing use of computer-based administration of clinical questionnaires through small electronic devices (e.g., phones) checking the validity of test scores becomes even more important. Patients are asked to fill out a questionnaire at irregular intervals and motivation problems may arise. In these cases, it is important to check the validity of test scores and patients can be asked to rethink answers or to provide additional responses (retesting) when responses are very inconsistent.

**A Computer Program and Examples**

**Computer Program.** To help researchers calculate person-fit statistics and to interpret misfitting item score patterns, the R-package PerFit (Tendeiro, 2015; Tendeiro, Meijer, & Niessen, 2015) can be used. Although there are some programs that can be used to calculate person-fit statistics, such as WPERFIT (Ferrando & Lorenzo, 2000) and PERSON (Choi, 2010), one advantage of PerFit is that the program only needs a data file with item scores and that item and person parameters can be estimated through the program. However, the user can also use already estimated item parameters. Another advantage of PerFit is that the package calculates a number of group-based statistics that are not available in other packages such as, for example, the powerful $H$ statistic. In Table 2, we provide an overview of all the statistics that can currently be calculated using the program.
Assessment 23(1)

(more methods will be added in the near future). There are also references to articles that discuss the individual statistics. Examples of syntax for computing the statistics and for creating plots and cutoff scores are in the help files of the package. Furthermore, the program plots nonparametric person response functions (PRFs; Emons, Sijtsma, & Meijer, 2004; Sijtsma & Meijer, 2001) for dichotomous item score patterns. PRFs give a visual representation of the fit of an individual item score pattern. More specifically, a PRF gives the probability of answering an item correctly or endorsing an item as a function of the item difficulty. For example, in Figures 2 and 3 we depict two examples of PRFs for persons who completed a dichotomously scored Inadequacy scale (to be discussed in more detail below). The PRFs for Person 592 (normal response behavior; \( G^p_N = .04 \)) and Person 176 (aberrant response behavior; \( G^p_N = .60 \)) are shown. The \( y \)-axis represents the probability of endorsing an item and the \( x \)-axis displays the difficulty values (1 − proportion-correct score), which is an indicator of item popularity (higher values indicating less popular items). Person 592 (see Figure 2) displays an expected pattern: The less popular the items are, the smaller the probability that this person endorses an item. The PRF of Person 176 (see Figure 3), on the contrary, seems to show that for less popular items, the probability of endorsing becomes larger. This is, of course, strange. Such response behavior is worth further inspection for a full interpretation of the results.

**Examples.** To further illustrate the use of PerFit, we discuss two empirical examples. The first example is concerned with the detection of unmotivated students in the context of a web survey. Data were collected from 294 students for research purposes in a population of students at a university. The survey consisted of several short questionnaires, including the Big Five Inventory (John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008), a medium length personality questionnaire. The students had to participate in a number of studies in order to receive course credit. Because the students had no direct gain in filling out the questionnaire seriously, there were concerns about the motivation of the students to complete the questionnaires and thus with respect to general data quality. We used person-fit analysis to detect careless responding students. One of the statistics used was the normed number of Guttman errors, which was computed for each subscale using PerFit. Respondents with aberrant responses could be identified and removed from the data set after inspection. An example is a female student 113, who had normed Guttman errors in the top 5% on two subscales, Extraversion (\( G^p_N = .37 \)) and Neuroticism (\( G^p_N = .41 \)). Inspecting the score pattern showed that she strongly endorsed the statement that she can be tense, but also agreed that she is relaxed and handles stress well. To further validate this extreme person-fit score, we inspected some other indicators of aberrant response behavior (see Meade & Craig, 2012). Further inspection showed that this student also responded suspiciously to an instructed response bogus item and completed the questionnaire in 14 minutes, whereas for most students the estimated completion time was about 20 minutes. Based on this information it can be defended to remove this item score pattern from the data.

As another example, we analyzed data from the Inadequacy scale from the Dutch Personality Questionnaire Junior (Dutch: Nederlandse Persoonlijkheidsvragenlijst-Junior: NPV-J; Luteijn et al., 2005). The NPV-J is a Dutch personality scale for adolescents and consists of five subscales. It is based on the California Personality Inventory. The Inadequacy scale was selected because it has good psychometric properties (Weekers & Meijer, 2008) and a relatively large number of items. Data were collected from 866 persons who attended primary and secondary education. Each item was either answered with “agree” or “disagree.” Figure 4 shows the distribution of \( G^p_N \). Note that, in this

**Figure 2.** Person response function for a normal behaving person on the Inadequacy scale.

**Figure 3.** Person response function for an aberrant respondent on the Inadequacy scale.
Table 3. Item score patterns and person-fit statistics for the Inadequacy scale.

<table>
<thead>
<tr>
<th>Item</th>
<th>Item content</th>
<th>Mean</th>
<th>Person 592</th>
<th>Person 176</th>
<th>Person 1,066</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN98</td>
<td>When I fail I feel sad</td>
<td>.56</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IN100</td>
<td>Nobody can do anything right</td>
<td>.50</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN52</td>
<td>When I am working on something I often think of other things</td>
<td>.40</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN34</td>
<td>Tired often</td>
<td>.33</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN70</td>
<td>Others talk about me</td>
<td>.33</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN48</td>
<td>Dream about things that I would not tell others</td>
<td>.32</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IN50</td>
<td>Ponder often</td>
<td>.30</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN57</td>
<td>Nervous often</td>
<td>.28</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN32</td>
<td>Worry about what I look like</td>
<td>.26</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN4</td>
<td>Often get headache when I worry</td>
<td>.24</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN96</td>
<td>Feel like something bad is going to happen</td>
<td>.24</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN102</td>
<td>Others are bothered when I am around</td>
<td>.24</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN6</td>
<td>Often bad mood</td>
<td>.20</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IN36</td>
<td>Few people understand me</td>
<td>.19</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN72</td>
<td>Often wake up at night</td>
<td>.19</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN93</td>
<td>Often fail</td>
<td>.19</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN19</td>
<td>When excited, my voice sounds strange</td>
<td>.17</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN59</td>
<td>Always get blamed when things go wrong</td>
<td>.17</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN13</td>
<td>Scared in the dark</td>
<td>.15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN75</td>
<td>Often bad dreams</td>
<td>.15</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>IN14</td>
<td>Others are happier than me</td>
<td>.14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN1</td>
<td>Cry for no reason</td>
<td>.13</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>IN8</td>
<td>Feel worthless</td>
<td>.11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IN91</td>
<td>Sweat when thinking of difficult homework</td>
<td>.11</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Limitations

Person-fit statistics are sensitive to unexpected response patterns conditional on the total score or trait value. This implies that if a respondent decides to fill out a questionnaire very consistently, for example, only choosing the first option resulting in, for example, all 1 scores, this pattern is not necessarily unexpected given the total score or trait value and such a pattern will not be detected as misfitting. Thus, person-fit statistics are especially sensitive to inconsistent response patterns. When researchers are interested in detection of, for example, long strings of similar scores that may be the result of unmotivated answering behavior, validity scales may be used or special person-fit statistics that are aimed at detecting long strings of unexpected responses (see Tendeiro & Meijer, 2012). Another limitation is that the power of person-fit statistics depends on test length and item...
characteristics. Longer tests will result in higher power to detect aberrant response behavior than shorter tests. In general, for short scales, say less than 10 items, it is not advisable to use person-fit statistics. Also, high item discrimination and large spread between the difficulties of the items will result in high power. Fortunately, however, many tests that consist of items with low discrimination power are often long (educational tests), whereas tests that are relatively short consists of items with relatively high discriminating items (stand-alone clinical scales, such as the Beck Depression Inventory).

Discussion

As several authors have emphasized, it is important to check for the validity of individual test scores. This can be done using different methods; in this article, we discussed the use of person-fit statistics. Of course, no method or even combination of methods can be used as final evidence that irregular activities have taken place. Nevertheless, we believe that ignoring the information that these methods can provide is like a detective “ignoring the footprints in the garden because they might not belong to the burglar” (Funder, 2007, p. 22). When psychological and educational tests are used for individual decision making, it is important to check for the individual validity of test scores. The methods and the software package that we discussed in this study can be of great help for researchers and practitioners.

In the literature, there is some discussion about the “validity of validity indices.” For example, McGrath, Mitchel, Kim, and Hough (2010) did not find much evidence for the influence of individual response configurations (“response styles”) on the validity of test scores, whereas others (e.g., Holden, Wheeler, & Marjanovic, 2012; Sotaridona & Meijer, 2003) showed that it is important to investigate these configurations of item scores. An important distinction between authors in favor or against the use of validity indices lies in the way these effects are evaluated. McGrath et al. (2010) used moderated multiple regression analysis in evaluating the presence of an effect of aberrant responding. In their study, criterion scores are regressed onto predictor scores (clinical scale scores, like BDI scores) and moderator scores (e.g., scores on the validity scale or person-fit scores) and their product term. The statistical significance of this product term then determines whether there is a moderator effect of validity scores. Above we discussed the study by Schmitt et al. (1993) who showed that the product of person-fit scores and total scores is significant. However, even if this product is not significant, several authors have argued (e.g., Ben-Porath & Waller, 1992; Holden et al., 2012) that aberrant responding can still influence the validity of individual test scores.

Also note that McGrath et al. (2010) “focus exclusively on studies that evaluated whether response bias indicator suppress or moderate the validity of substantive indicators” (p. 454). They found that “the evidence was simply insufficient to draw firm conclusions” (p.436) or “the evidence failed to corroborate the hypothesis, although careless responding represents a possible exception” (p. 463). However, their final conclusion was rather nuanced saying that “in the absence for or against the measurement of response bias, a reasonable argument can be made for the continuing to use response bias indicators,” but they warned against the detrimental effects of false positives, that is, erroneously concluding that a test taker has cheated whereas in practice this is not the case. Especially, when scores are corrected on the basis of bias indicators this may affect the rank ordering in, for example, personnel selection contexts.

We fully agree with their remark and we are certainly not in favor of correcting scores on the basis of, for example, person-fit scores. At the same time, their remark shows that manipulating responses may result in different ordering of persons and may have an important effect on selection results. Also note that person-fit statistics are especially useful to detect careless or random responding. Furthermore, Holden et al. (2012) showed that although moderated multiple regression may indicate that criterion validity can vary as a function of another variable, this technique does “not identify specific cases where another variable has manifested its effect” (p. 19). They showed that discriminant analysis was better able to identify biases responses than moderated multiple regression techniques.

On a more general level, ignoring information about the validity of individual test scores because on a group level there seems to be no effect between predictor scores and validity indices is not to be recommended. For example, in high-stakes educational testing, individual test scores are checked and students with very unlikely score patterns or very similar score patterns as other students have to redo the exam (Belov, personal communication, April 4, 2013). Note that in this example it is important to check the total scores, irrespective of the effect on the group level. Also in clinical research, several studies discussed above showed that total scores for persons with extreme person-fit scores are difficult to compare. For example, in the study by Wardenaar et al. (2014), clients with extreme person-fit scores endorse items that refer to mood/cognition and do not endorse the somatic items. Depression scores for persons with extreme person-fit scores should thus be interpreted differently than total scores for persons with normal person-fit scores.

Finally, note that person-fit statistics define invalid scores with respect to “group-level” patterns. In person-fit research, it is always assumed that for a large majority of persons the test model can be used as a valid indicator of their trait level. However, for a small percentage the response patterns may be very unlikely given the IRT model and further inspection may be required. This is also in agreement with Dawes, Faust, and Meehl’s (1989) observation that in general it would be absurd to rely on an actuarial model (i.e., in our
case an IRT model) if there is overwhelming evidence that this model will not apply because of particular observations. This is known as the “broken leg problem” where an actuarial formula can be successfully used to predict an individual’s attendance to the movie is highly successful in predicting an individual’s weekly attendance at a movie but should be discarded upon discovering that the subject is in a cast with a fractured femur. The clinician may beat the actuarial method if able to detect the rare fact and decide accordingly.  
(Dawes et al., 1989, p.1670)

As discussed above, person-fit statistics may help identify such rare cases.

**Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**

The author(s) received no financial support for the research, authorship, and/or publication of this article.

**References**


