Detecting careless respondents in web-based questionnaires

Niessen, A. Susan M.; Meijer, Rob R.; Tendeiro, Jorge N.

Published in:
Journal of Research in Personality

DOI:
10.1016/j.jrp.2016.04.010

Publication date:
2016

Journal of Research in Personality 63 (2016) 1–11

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Detecting careless respondents in web-based questionnaires: Which method to use?

A. Susan M. Niessen *, Rob R. Meijer, Jorge N. Tendeiro

Department Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 7 July 2015
Revised 30 March 2016
Accepted 29 April 2016
Available online 30 April 2016

Keywords:
Data collection
Careless response
Data screening
Response patterns
Person fit

ABSTRACT

High data quality is an important prerequisite for sound empirical research. Meade and Craig (2012) and Huang, Curran, Keeney, Poposki, and DeShon (2012) discussed methods to detect unmotivated or careless respondents in large web-based questionnaires. We first discuss these methods and present multi-test extensions of person-fit statistics as alternatives. Second, we applied these methods to data collected through a web-based questionnaire, in which some respondents received instructions to respond quickly which can result in more careless responding. In addition, we conducted a simulation study. We compared sensitivity and specificity of different methods and concluded that multi-test extensions of person-fit statistics are a good alternative as compared to other methods, although the sensitivity to detect careless respondents using empirical data was lower than using simulated data.

1. Introduction

High data quality is a very important prerequisite for sound empirical psychological research. In applied studies in the social and behavioral sciences, the Internet is increasingly used for data collection. Although web-based questionnaires are an efficient way to collect data and some studies have shown that using the web may have no detrimental effect on data quality (e.g., Beach, 1989), other studies called for screening tools to check data quality. In particular when a questionnaire is being administered to specific groups, like students, unmotivated or careless respondents may pose a serious threat to data quality (Johnson, 2005; Maniaci & Rogge, 2014). Maniaci and Rogge (2014) have shown, for example, that inattentive responses lead to poor data quality and can have negative effects on effect sizes and power depending on the proportion of inattentiveness in the sample. Recently, Meade and Craig (2012) and Huang, Curran, Keeney, Poposki, and DeShon (2012) studied methods for detecting careless respondents and recommended several methods to identify these respondents. However, in both studies questionnaires containing 300 items and many subscales were used. As we discuss below, the usefulness of some of the methods they recommended relies heavily on the availability of a large number of items. The use of these methods is limited when shorter scales, scales that predominantly measure one factor, or scales with a limited number of subscales are used. In the present study, we first discuss existing methods in the literature to detect careless response behavior and second we propose the use of person-fit statistics as alternative methods. Third, we compared the sensitivity and specificity of the methods suggested in previous studies (Huang et al., 2012; Meade & Craig, 2012) with person-fit statistics using web-based administered data and simulated data. Finally, we provide some practical advice as to which statistics to use.

2. Methods to detect careless responding

Careless responding is characterized by "random responding, leaving many answers blank, misreading items, answering in the wrong areas of the answer sheet and/or using the same response category repeatedly without reading the item" (Johnson, 2005, pp. 104–105).[1] Two main research contexts are often mentioned as prone to careless responding: Research that uses web-based questionnaires and research where students receive course credit for their participation (Kurtz & Parrish, 2001; Meade & Craig, 2012). Johnson (2005) mentioned several possible reasons for the

* Corresponding author at: Department of Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands.
E-mail address: a.s.m.niessen@rug.nl (A. Susan M. Niessen).

[1] As one reviewer noted "careless responding captures the more benign form of aberrant responding" where participants misinterpret items due to inattention (see Schmitt & Stults, 1985), whereas random responding is more blatant and intentional. In most part of this study we study careless responses, but in the simulation study we will discuss the differences between these types of aberrant responses in more detail.

http://dx.doi.org/10.1016/j.jrp.2016.04.010
0092-6566/

occurrence of careless responding in web-based questionnaires. Psychological distance from the administrator and responding anonymously could lead to decreased accountability experienced by the respondents. In addition, the effortless process of completing an internet-based questionnaire could induce rushing through the questions. For students that receive course credit for their participation concerns have been raised about the validity of data obtained due to disturbing levels of non-cooperation and motivational factors (Kurtz & Parrish, 2001; Meade & Craig, 2012; Sprock, 2000).

Meade and Craig (2012) evaluated methods for detecting careless responding using a web-based questionnaire administered to a student sample. In addition, they conducted a simulation study to test the sensitivity of the methods they studied. These methods included bogus items, consistency indices, response-pattern analysis, outlier analysis, response time analysis, and self-report measures (to be discussed in more detail below). Based on their findings they made the following recommendations. First, as minimum screening tools, they recommended screening for extremely short response times and using a self-report measure of response quality. An example of the latter method is asking: "In your opinion, should we use your data?" at the end of a questionnaire. Second, they recommended using bogus items, preferably in the form of an instructed response, such as: "Respond with *strongly agree* for this item". Finally, they recommended using three post-hoc measures, an even-odd consistency index, a maximum longstring index, and a multivariate outlier statistic: the Mahalanobis distance (Mahalanobis, 1936). When the even-odd consistency index cannot be applied due to unavailability of multiple subscales, the psychometric antonyms or psychometric synonyms measures were recommended.

In a web-based personality questionnaire administered to students Huang et al. (2012) also studied two consistency indexes (the even-odd consistency index and the psychometric antonyms index), the longstring, and response times. The first half of the questionnaire was completed with the instruction to answer honestly, whereas in the second half there were different conditions. Students in the first condition received again the instruction to answer honestly, students in second condition were instructed to answer the items without much effort but without running the risk of being caught (the cautionary careless condition), and in the third condition students were instructed to respond without effort, with no risk of consequences (careless condition). Huang et al. (2012) studied the sensitivity of these indices using different cutoff scores. Based on their findings, they recommended using the even-odd or psychometric antonyms consistency indices and a short response time as indicators for careless response.

Although these are useful and practical recommendations, Meade and Craig (2012) already mentioned at the end of their study that some of these methods are only useful for questionnaires with many items. To illustrate this, below we discuss the recommended methods and their limitations. In addition, we propose and discuss detection of careless responding using person-fit statistics based on item response theory (IRT) models. These statistics are often discussed in more 'technical' psychometric papers in which their statistical properties are researched. However, there are few studies that apply person-fit statistics to detect careless responses. Johnson (2005) mentioned person-fit statistics to detect careless responses but discarded them for being too stringent, complex, and computationally intensive. Although IRT methods may, as we will show below, have some drawbacks, it is also true that in the IRT literature many different statistics have been proposed to detect inconsistent response behavior that differ with respect to the strictness and complexity of the underlying model. Therefore, we included person-fit statistics in this study and investigated how they performed compared to the methods suggested in the papers cited above.

2.1. Consistency indices

Meade and Craig (2012) and Huang et al. (2012) studied the even-odd consistency index and the psychometric antonyms index. In addition, Meade and Craig (2012) also used a psychometric synonyms index. They recommended using the even-odd consistency index. However, these indices are not very useful for questionnaires that do *not* consist of many items and scales. To explain why this is the case, let us first consider how these indices are computed.

To compute the even-odd index, each scale in a questionnaire is divided into two subscales consisting of the even and odd numbered items, respectively. For each subscale, total scores are computed and then for each respondent the Pearson's product-moment correlation is computed between the total scores on the even and the odd items across the subscales. Then this correlation is corrected for test length using the Spearman-Brown formula (Johnson, 2005). That is, the correlation is calculated for the whole test because the correlation between the odd and even items only provides an estimate for half of the test. The idea is that the correlation between the total scores on the even and odd items is high when a respondent fills out a questionnaire seriously. When this correlation is low, response behavior is inconsistent. Jackson (1977) proposed that scores on the even-odd index lower than 0.30 may indicate inconsistent response behavior.

Now, consider a questionnaire consisting of 100 items, each item scored 1–5, with five subscales each consisting of 20 items. To compute the even-odd index for respondent j with mean subscale scores ($X_{j1Even},\ldots,X_{j5Even}$ and $X_{j1Odd},\ldots,X_{j5Odd}$), a correlation coefficient between the mean scores on the even sections and the odd sections is computed and corrected for test length. As an example (taken from the dataset discussed below), we take the mean even and odd subtest scores of person 192 for the 5 subscales: this resulted in (3.7, 3.4; 3.6, 3.4; 3.3, 3.9; 2.9, 3.8; 2.9, 3.2) and a correlation of $r = -0.35$, indicating low consistency. However, note that the "sample size" in this procedure equals the number of subscales used in the analysis, in this case five, which results in a very unreliable estimation of the correlation coefficient. The 95% confidence interval of the score for this specific respondent was between −0.94 and 0.77. Even when, say, thirty subscales are used, the reliability of the estimated consistency seems problematic.

The psychometric antonyms and/or the psychometric synonyms indices can serve as alternatives when multiple subscales are not available. The psychometric antonyms index is determined by first identifying item pairs with opposite content as indicated by negative inter-item correlations (referred to as "opposite item pairs" below). Huang et al. (2012) selected 30 item pairs with the highest negative correlations to compute this statistic. Meade and Craig (2012) considered items truly opposite when the inter-item correlation was largely negative, that is, they considered item pairs with an inter-item correlation between $r = -1$ and $r = -0.60$. For each respondent, a correlation coefficient was computed between the scores on the opposite item pairs. In the case of consistent behavior, the correlation will be largely negative. The procedure for the psychometric synonyms index was similar, where item pairs with an inter-item correlation larger than $r = 0.60$ were considered synonyms.

For both of these measures, the number of item pairs functions as the sample size for the computed correlation coefficient. This results in the same problem as in the even-odd index: Unless the number of item pairs used is large, the confidence intervals around the correlation estimate will be very large. Meade and Craig (2012) could identify 27 item pairs meeting the criteria for the synonyms index and only five for the antonyms index in a 300-item questionnaire. Therefore, this method will often not be suitable for shorter

questionnaires. Moreover, it is debatable whether questionnaires should contain many items that are so strongly positively or negatively correlated. Asking persons very similar questions might stimulate careless behavior. Huang et al. (2012) took a fixed number of item pairs with the largest negative and positive inter-item correlations to guarantee a sufficient number of item pairs. However, a drawback then may be that items are not related strongly enough to be true antonyms or synonyms and to serve as good indictors of inconsistency.

2.2. Mahalanobis distance

The Mahalanobis distance is a multivariate outlier statistic (Mahalanobis, 1936) that portrays the distance between observations, here the person's item scores, and the center of the data, here the vector of sample means for the items, while taking the correlational structure between the items into account. The distance is smallest when the vector of a person's responses is similar to the vector of sample means. Meade and Craig (2012) computed the Mahalanobis distance for each of five trait subscales and averaged the five outcomes, resulting in a mean Mahalanobis distance over all trait scales. As a result, a low Mahalanobis distance on one subscale can compensate a high Mahalanobis distance on another subscale. The Mahalanobis distance can also be computed across an entire questionnaire. Probability values can be computed by converting Mahalanobis distance to chi-square *p*-values (Zijlstra, van der Ark, & Sijtsma, 2011), and a critical *p*-value can be selected to flag suspicious respondents. However, this only holds when the assumption of multivariate normality of the item scores is met. Considering the number of response options that are used in survey data, it is unlikely that this assumption will hold.

2.3. Maximum longstring

This measure is sensitive to a different type of careless response behavior than the previous methods. Instead of random or inconsistent responding, the longstring method is sensitive to extremely consistent responding. The maximum longstring equals the number of times a respondent chooses the same response option consecutively. Although this seems to be a very useful measure to detect this particular type of careless responding, a cutoff score is difficult to establish (Johnson, 2005). Costa and McCrae (2008) provided some guidelines for maximum longstrings based on the NEO-PI-R, which were also used by Huang et al. (2012). These guidelines were conditional on the number and types of items selected and varied from six to fourteen consecutive responses. In particular for short questionnaires it may be difficult to distinguish conscientiously and consistently responding persons from careless respondents using a long string measure. A method suggested by Johnson (2005) to find a clear cutoff point was to conduct a scree-like test and find an 'elbow' in the data.

2.4. Response time

Meade and Craig (2012) and Huang et al. (2012) studied short response time as an indicator of carelessness. It is very unlikely that respondents read the item content and answer items seriously when the response time is very short. In contrast, a long response time can, for example, be due to taking a break, which should not necessarily be labeled as careless. A concern when using response time is, again, establishing a suitable cutoff score. Meade and Craig (2012) did not provide any guidelines and suggested to look at clear outliers in the response time distribution, whereas Huang et al. (2012) used an "educated guess". That is, given their knowledge of the questionnaire they decided that it was unlikely that a response to an item could be given in less than 2 s. Responses to sets of items that reflected much faster response behavior were classified as careless.

2.5. Bogus items

Bogus items are questions that have an obvious correct answer, so that an incorrect answer can be regarded as not paying attention to the item. An example of a bogus item is "I am always wearing the same clothes" with answer options yes/no. Respondents that choose here the option "yes" may fill out the questionnaire without paying attention to the content of the items. In their questionnaire, Meade and Craig (2012) added 10 bogus items. They recommended using no more than one bogus item per 50–100 questions because it may irritate respondents if there are too many of such items.

2.6. Explicit instructed response item

Specific types of bogus items are items with an instruction to provide an extreme response ("Please respond with *strongly agree* to this item"). These items are considered sensitive in detecting careless respondents because it is most unlikely that the keyed response is given spontaneously without reading the question. Although it is relatively easy to add a few items like this to a questionnaire, there are some challenges that may invalidate the use of these indices. First, respondents may not endorse such items in the keyed direction because they think these items are funny and as a result respondents willingly provide a non-keyed response. Second, it is difficult to determine how many items should be included in a questionnaire and it is even more problematic to establish a cutoff score that can be used to consider a respondent as careless. Also, as far as we know, there is not much known about what kind of item format or item content works best and when investigated results are mixed. Huang et al. (2012) originally included instructed response items in their questionnaire, but found that they flagged an unrealistically high number of respondents. They suspected that respondents might have interpreted these items idiosyncratically. In a recent paper, however, Huang, Bowling, Mengqiao, and Li (2015) investigated the validity of a scale consisting of eight items that referred to counterfactual statements, deviation from "common sense", and improbable events and they found promising results for the future use of these methods.

All in all, from our discussion above it is clear that the use of several existing indices is not always straightforward, especially for scales that do not consist of hundreds of items. For scales of moderate test length, in the psychometric literature person-fit statistics are sometimes advocated to identify inconsistent item score patterns. Below we discuss these statistics and provide arguments why these statistics may be potentially useful as alternative statistics or may be used complementary to the statistics discussed above.

3. Person-fit statistics

Person-fit statistics have been proposed to detect inconsistent item score patterns given the other score patterns in a sample or given that an IRT model fits the data (Meijer, Niessen, & Tendeiro, 2016; Meijer & Sijtsma, 2001). For some statistics it is necessary to first estimate the parameters of an IRT model, whereas there are also statistics that do not assume any parametric IRT model, but instead are based on sample properties. Karabatsos (2003) showed that simple nonparametric statistics performed as well as other statistics (see also Tendeiro & Meijer, 2014). In the present study, we used a multi-test extension (Conijn, Emons, & Sijtsma, 2014; Drasgow, Levine, & McLaughlin, 1991; Meijer, Egberink, Emons, & Sijtsma, 2008) of the nonparametric number

of Guttman errors for polytomous items (G^p; Meijer, Molenaar, & Sijtsma, 1994; Molenaar, 1991) and a multi-test extension of the l_z statistic for polytomous items (l_z^p; Drasgow, Levine, & Williams, 1985).

3.1. Number of Guttman errors

Almost all person-fit statistics are sensitive to the number of Guttman errors. Consider a test consisting of five dichotomously scored items ordered from easy (or most popular) to difficult (or least popular) according to their item proportion-correct (or mean score) score. If a respondent has a total score of three, then we expect that the three easiest items are answered correctly and the remaining two most difficult items are answered incorrectly. When there is a reversal of item scores, that is when a difficult item is answered correctly and the easier item is answered incorrectly, this is counted as a Guttman error. In general, the more errors, the more inconsistent an item score pattern is. Thus, assuming that items are ordered from easy to difficult the pattern [1, 1, 1, 0, 0] contains zero errors, whereas the pattern [1, 0, 0, 1, 1] contains four errors, because there are four (0, 1) item pairs.

For polytomous items the concept of item steps is used (Sijtsma & Molenaar, 2002). An item step is a psychological threshold between ordered response options. As an example we use the personality item "I see myself as someone who is talkative", with five response options: *strongly disagree*, *disagree*, *neither agree nor disagree*, *agree*, and *strongly agree*. A respondent first considers whether he or she agrees enough with the statement to take the first item step from *strongly disagree* to *disagree*. If not, the response will be *strongly disagree*. The next consideration is to move from *disagree* to *neither agree nor disagree*. This process continues until the respondent decides not to take the next item step or reaches the last response option. In a sample, we can determine the proportion of respondents that took each item step, for each item. Then the item steps are ordered from high to low popularity. Finally, for each respondent the observed item responses can be compared to the ordering of the item steps. A Guttman error occurs when a less popular item step was taken, whereas a more popular item step was not taken. The sum of the Guttman errors results in the G^p statistic. In Appendix A we provide a numerical example using three items.

3.2. L_z^p statistic

IRT models specify the probability of given a specific answer to an item as a function of a respondent's latent characteristic or trait (i.e., personality traits, intelligence, but also the amount of knowledge of a particular subject matter). Using an IRT model it is possible to calculate the likelihood that a respondent with a particular trait level answers an item correctly (dichotomous items) or chooses a particular option (for polytomous items). The central idea is that when this likelihood is very low, an item score pattern can be classified as inconsistent. To illustrate this: assume that we have 3 polytomous items with three response categories (0, 1, and 2) and the popularity of choosing these answer categories for a person with a particular trait level equals for item 1 (0.3, 0.5 and 0.2) for item 2 (0.7, 0.2, and 0.1) and for item 3 (0.1, 0.5, and 0.4). Then the item score pattern (1, 0, 1) is the most likely pattern and the score pattern (2, 2, 0) is the most unlikely pattern. Several fit statistics have been proposed that are sensitive to the degree an item score pattern is unlikely. One of these statistics is the l_z^p statistic (Drasgow et al., 1985). This statistic is defined as the standardized log-likelihood of an item score vector under an IRT model. We define l_z^p under the graded response model (Samejima, 1969). The exact computation of this statistic can be found in Drasgow et al. (1985).

3.3. Multi-test extension of Guttman errors and l_z^p

A main concern when using person-fit statistics is their sensitivity to detecting aberrant response patterns when scales have few items. Since one of the prerequisites for applying IRT models is unidimensionality of the scale, IRT models cannot simply be applied across multiple short subscales. The sensitivity of these person-fit statistics depends on the discrimination parameters, the spread of the item difficulties, and the test length (e.g., Meijer et al., 1994). Two possible solutions for this sensitivity problem are suggested in the literature. The first solution is to use scales with strongly discriminating items. Such scales yield similar detection rates as a larger scale with items weaker in discrimination (Meijer et al., 1994). Another method to increase sensitivity is to use a multi-test extension (Conijn et al., 2014; Drasgow et al., 1991; Meijer et al., 2008). In a multi-test extension, information on person fit across multiple (short) subscales is pooled. A simple way to do so is by summing the person-fit statistics across different subscales as in Meijer et al. (2008) and Conijn et al. (2014). In the present study, we use multi-test extensions of G^p (G_m^p; Meijer et al., 2008) and l_z^p (l_{zm}^p; Conijn et al., 2014). G_m^p is defined as the sum of the number of Guttman errors per subscale and l_{zm}^p is the sum of the standardized log-likelihood statistics of the subscales. A researcher or practitioner can calculate these statistics using the R program PerFit (Tendeiro, 2015; Tendeiro, Meijer, & Niessen, in press).

4. Method

Investigating the sensitivity and specificity of different methods to detect careless responding is not simple. Often two conditions are needed: a careless condition where respondents answer the items in some kind of careless way and a normal condition where respondents answer the items "normally". Then, sensitivity equals the proportion of score patterns correctly flagged in the careless condition, whereas specificity equals the proportion of score patterns correctly flagged in the normal condition.

Thus, what is needed is information about who answered carelessly and who answered honestly, or "normally". In practice, this is almost never known. Therefore, in many studies data are simulated and researchers manipulate normal response behavior and careless behavior. A drawback of simulated data is, however, that the data may be too artificial and do not mimic real "careless" behavior, and as a result provide inflated or deflated sensitivity results. For example, a researcher may randomly generate item scores from a uniform distribution to mimic careless response behavior. This will result in perfect random response patterns that, in practice, probably are seldom observed. In the real world, respondents may only be careless on part of the questionnaire, and more importantly, respondents do not behave as randomly as a random number generator. It is well known that persons find it very difficult to generate completely random score patterns, even if instructed to do so (e.g., Nickerson, 2002). These same arguments apply mutatis mutandis when careless response behavior is mimicked by, for example, generating long strings. What realistic long strings are is difficult to determine.

As an alternative, a researcher may instruct respondents to respond carelessly or normally. This will result in real careless and normal response behavior as far as the respondents followed the instructions. If this is not the case, and one can never be sure, sensitivity, for example, will probably be underestimated because there will be less careless respondents in the careless group than expected on the basis of the instructions. To explain this: Note that sensitivity is defined as the proportion of score patterns correctly flagged in the careless condition. Assume now that we instruct

40 persons to respond carelessly. A perfect statistic would classify all these patterns as careless, and sensitivity equals 1. However assume now that only 20 out of these 40 persons would respond carelessly, then the maximum sensitivity is only 20/40 = 0.5. Following a similar reasoning, specificity can be underestimated when respondents in the normal condition respond carelessly.

In this study we followed both approaches, although it is true that both approaches have their strong and weak points. We used empirical data pertaining to students instructed to provide honest responses to a questionnaire. Then we compared the score patterns of these students with (a) students who were instructed to provide careless responses, and with (b) simulated careless responses. Doing so, we tried to get a good picture of the usefulness of the different methods to detect careless response behavior.[2]

4.1. Participants and measures

Sophomore psychology students who took a course in test theory and test construction were asked to voluntarily complete the 100 item International Personality Item Pool (IPIP) questionnaire (Goldberg et al., 2006) with five response options. This questionnaire is a freely available personality inventory that was designed to measure the Big Five personality traits extraversion, agreeableness, conscientiousness, neuroticism, and openness, and was constructed based on the NEO-PI-R (Costa & McCrae, 1992). Each of these subscales consists of 20 items. The questionnaire was administered through an online survey tool. The students could not obtain course credit or any other incentive for their participation.

There were 652 students in the course and 266 of these students completed the questionnaire (41%). Seventy-three percent was female and the mean age was $M = 21$ ($SD = 3.4$). The questionnaire was set up so that approximately 25% randomly drawn respondents would receive instructions to respond carelessly, resulting in 70 participants in the 'careless' condition and 196 participants in the 'normal' condition. Students in the 'normal' condition received instructions to complete the questionnaire honestly:

> *The following statements apply to your perceptions about yourself in different situations. You are asked to indicate to which extent you agree with each statement, using a scale from strongly disagree to strongly agree. There are no right or wrong answers, so please select the answer that suits you best for each statement. Please take your time to think about each answer.*

Students in the 'careless' condition were instructed as follows:

> *Imagine that you have to complete a questionnaire in order to get course credit. You do not feel like completing the questionnaire and you are only interested in getting the course credit. You would like to do this task as quickly as possible. You will only receive course credit if you finish the entire questionnaire. Complete the following questionnaire like you would have done if you were in this situation.*

Note that we asked the students to respond as quickly as possible, assuming that this may result in careless response behavior. We would like to emphasize here that it is possible that some students may actually respond as quickly as possible and still provided accurate answers. So, although we tried to mimic the situation that may lead to careless response behavior the current condition may not completely coincide with careless responding. Yet, for the sake of simplicity we will denote this as the careless condition. A bogus item, an instructed response item and a question about response quality were added to the questionnaire. An item from the Big Five Inventory (John, Donahue, & Kentle, 1991; John, Naumann, & Soto, 2008) was used followed by the instruction "Respond with strongly agree for this item". Respondents who ignored this instruction were flagged. The bogus item was taken from Meade and Craig (2012), with the content "I have never brushed my teeth". Students who did not respond with *strongly disagree* or *disagree* were flagged. Following Meade and Craig's advice, we only used two items to flag aberrant respondents, so that respondents were not irritated by the presence of these types of items.[3] At the end of the questionnaire, the following question about response quality was asked: "If the researchers of this study were really interested in your personality structure, would you recommend using your responses in their research? It is very important for the researchers to have valid data." The students could answer *yes* or *no*, were *no* was taken as an indicator of careless response. For each student, response time in minutes was also recorded. The raw data are available via Niessen, Meijer, and Tendeiro (2016).

4.2. Procedure

We assessed sensitivity and specificity using rational cutoff scores based on the literature, and empirical cutoff scores based on the empirical data. As independent variables we used: the response quality item, the bogus item and instructed response item, response time, maximum longstring, the even-odd index, Mahalanobis distance, and the two multi-test extensions person-fit methods discussed above. Rational cutoff scores were used for all measures except Mahalanobis distance and the two person-fit statistics, since none were available. Empirical cutoff scores based on the values found in the sample who received normal instructions were used for all measures except the response quality question and the added items. For these latter measures we used 0/1 coding.

Sensitivity and specificity based on rational cutoff scores were assessed by analyzing the proportion of respondents from the entire careless condition ($n = 70$) identified as aberrant, and the proportion of respondents from the normal condition ($n = 196$) identified as normal. Response time in minutes was recorded for each respondent. To establish a rational cutoff we followed a procedure similar to that of Huang et al. (2012). We estimated how many seconds it would take to read the instructions/demographic questions and to read each item (this was 3 s per item, thus 1 s longer than in the Huang et al., 2012 study). This resulted in a minimum response time of 6 min to be classified as an attentive respondent. The rational cutoff for the maximum longstring was based on a scree-like plot using only the respondents from the normal condition (Jackson, 1977). The scree plot did not show a clear cutoff, so we were conservative (a cutoff value between five and seven seemed reasonable based on the plot) and chose a cutoff of more than seven consecutive responses. The even-odd index scores were computed for each respondent using the five subscales of the IPIP. Items were assigned to even and odd scales randomly, since not all subscales had an equal number of even and odd numbered items. The cutoff suggested by Jackson (1977) of values smaller than 0.30 was used to flag respondents.

To analyze empirically derived cutoff scores we generated samples with different proportions of careless respondents. Based on Meade and Craig's (2012) estimate of the prevalence of careless

[2] One reviewer asked why we did not use a separate condition where we warned the respondents that not filling out a questionnaire seriously is against the rules of the university. In some studies this was found to be an effective method to avoid careless responding. However, at our university it is not allowed to do this because participating in a study is the only requirement to get credits. Therefore, we concentrate here on statistics that can be used when these instructions are not given.

[3] As one of the reviewers remarked this number of bogus items or instructed items might be too conservative in practical test use and may increase error in the sensitivity of these methods as compared to using more bogus items. Therefore, our conclusions using this method will be interpreted cautiously.

Table 1
Descriptive statistics for all detection measures.

Method	Sample												ES
	Total				Normal				Careless				
	M	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.	
Use Me[a]	0.23				0.14				0.49				5.83[*,c]
Instructed item[a]	0.73				0.71				0.81				1.82[c]
Bogus item[a]	0.43				0.39				0.51				1.62[c]
Response time[b]	11	856	3	9750	12	980	3	9750	8	311	3	2310	−0.08[d]
Max. longstring	6.5	11	2	100	4.8	3	2	34	11.3	20	3	100	0.62[*,d]
Mahalanobis D	9.88	1.40	3.52	13.51	9.80	1.12	4.69	13.34	10.12	1.98	3.52	13.51	0.23[d]
Even-odd index	0.75	0.44	−1.0	1.0	0.80	0.34	−1.0	1.0	0.58	0.62	−1.0	1.0	−0.51[*,d]
G_{fm}^{p}	408	216	85	1392	376	160	133	1392	498	310	85	1314	0.58[*,d]
l_{zm}^{p}	0.72	6.2	−27.6	8.8	1.69	4.5	−26.7	8.8	−1.99	8.9	−27.6	8.2	−0.62[*,d]

ES is the effect size for the difference between respondents in the normal condition and respondents in the careless condition.
[a] Proportion flagged.
[b] Median reported instead of mean.
[c] Odds ratio.
[d] Cohen's *d*.
[*] $p < 0.05$.

response (10–12%), we created conditions in which there were 5%, 10%, or 20% careless respondents, also representing conditions with milder and more extreme prevalence of carelessness than discussed in Meade and Craig (2012). That is, 11 (5%), 22 (10%), or 49 (20%) respondents from the careless condition were added to the sample of 196 respondents who received normal instructions. The 5%, 10%, or 20% respondents were sampled randomly from the careless respondents and this procedure was repeated 1000 times per condition.

Before computing Mahalanobis distance and the multi-test extensions of the number of Guttman errors and the l_z statistic we recoded all items because for many items the response option strongly disagree was never chosen. All items were recoded to a four point scale, merging strongly disagree and disagree responses. This may result in some loss of information. To check this we compared the measurement accuracy[4] using four and five options and we found only differences for persons with low total scores. Only for the neuroticism scale differences were somewhat larger. After recoding the item scores, model fit for the graded response model was checked by inspecting the S-X^2 item diagnostics (Orlando & Thissen, 2003), the marginal fit and standardized local dependence X^2 statistics. These statistics showed an overall good fit.

Mahalanobis distance was computed across all items. The number of Guttman errors and the l_z statistic for polytomous items were computed for each subscale using the PerFit R package (Tendeiro, 2015; Meijer et al., 2016). A sum score across subscales was computed for both statistics to obtain the multitest extension of the person-fit statistics. Since person-fit statistics and Mahalanobis distances cannot be computed when missing item responses are present, we used the nonparametric imputation method available in the PerFit package to handle missing responses before computing these statistics. Thirteen percent of the students had at least one missing item response. Most of the students (89%) with missing item responses had one or two items missing responses, with a maximum of seven missing responses, so the effect of imputation on response pattern consistency was very small. It is important to note that we did not consider these missing values as a sign of careless responding because for almost all students the number of missing values was very small. Finally, empirical cutoff scores were computed for each condition, depending on the proportion of careless respondents in the sample. To obtain these cutoffs for response time, the maximum longstring, the even-odd index, Mahalanobis distance, and the two person-fit statistics, the 5%, 10%, and 20% most extreme scores obtained by respondents in the normal condition, determined the cutoff values. Because values for Mahalanobis distance and the person-fit statistics depended on the other respondents in the sample and were computed on samples including careless respondents, these cutoffs were computed after each sampling procedure and differed slightly over replications.

4.2.1. Simulated data

Because we were not sure whether students followed our instruction to respond quickly, and whether quick responses resulted in careless responses, we also simulated careless response behavior and compared the sensitivity of the two person-fit statistics and the Mahalanobis distance with the sensitivity using the empirical data. Simulated responses were added to the response patterns from the empirical data. Careless respondents were operationalized as respondents who answered items randomly. The proportion of random respondents was, as in the empirical example, 5%, 10%, and 20% of the total sample, and the proportion of items that was answered randomly equaled 10%, 25%, and 50%. For each simulated item score pattern normal item responses were randomly selected from the empirical dataset of respondents in the normal condition. We then changed scores for 10%, 25%, and 50% of the items by drawing an item score between 1 and 4 randomly from both a uniform distribution and a normal distribution N ($\mu = 1.5$, $\sigma = 1$) with 100 replications per condition. These distributions represent completely random response behavior (uniform distribution) and random response behavior with a higher probability to choose the middle response options (normal distribution). We did not assume that for careless behavior the probability to choose a particular response option was related to that person's normal response behavior. Specificity equaled the proportions of the simulated random item score patterns that belonged to the 5%, 10%, or 20% most extreme values.

5. Results

5.1. Empirical data

Table 1 shows descriptive statistics for all detection measures for the complete empirical data sample, and for the subsamples with respondents who received normal and careless instruction, respectively. It also shows effect sizes for the differences between respondents in the normal condition and respondents in the careless condition. For the bogus item, instructed response item,

[4] Measurement accuracy was determined using IRT-based information functions that provide information about the standard error of the estimated trait level across different trait levels.

and the data quality question more respondents were flagged in the careless condition than in the normal condition. However, these differences were rather small and not significant for both items. Furthermore, there were some very high response times, up to several days, so we reported the median response time. Students in the careless condition responded faster than students in the normal condition when looking at the medians, but this difference was not significant. Very long strings of the same consecutive responses were rare, but they were more common in the careless condition. Two respondents, both in the careless condition, gave the same response on all items. Also, respondents in the careless condition had lower even-odd index scores, more Guttman errors, and lower l^p_{zm} statistics than respondents in the normal condition. The differences were small and non-significant for Mahalanobis distance.

To check the reliability of the even-odd index with a scale containing a limited number of subscales, we computed 95% confidence intervals around the even-odd index value for each respondent. The intervals had a mean width of $M = 0.85$ ($SD = 0.51$). The mean width was even larger in the group that had an even-odd value below the rational cutoff score of 0.30 ($M = 0.91$, $SD = 0.53$). The cutoff score of 0.30 fell within the confidence interval of 56% of all respondents. For the careless sample this was true for 60% of the respondents and for the normal sample for 55% of the respondents. This indicated that the numerical values of the even-odd index should be interpreted carefully. We also inspected the distributions of the scores using the entire sample and found a left-skewed distribution for the even-odd index and the l^p_{zm} statistic, and a right-skewed distribution for response time, the maximum longstring, and the G^p_m statistic. Only Mahalanobis distance was distributed close to normal, with a slight left skew.

Table 2 shows the sensitivity and specificity for each method using rational cutoff scores and Table 3 shows sensitivity using empirical cutoff scores. Let us walk through the most salient results. Remember that in the normal condition (Table 1), according to the response quality question 14% of the respondents indicated that their responses should not be used (abbreviated as "Use me"); in each careless response condition this was around 50% (Table 2). This latter percentage was low, perhaps in part because the students were instructed to fill out the questionnaire quickly. That is, quick response behavior may not always have resulted in careless behavior.

However, anecdotal evidence suggested that some of these students responded to this question randomly as well. The bogus item and instructed response item had high sensitivity, but also showed low specificity; this was especially the case for the instructed response item ("Respond with *strongly agree*"). Sensitivities for response time, the maximum longstring and the even-odd index were comparable, and were around 20%, with high specificity when rational cutoff scores were used.

When empirical cutoff scores were used, sensitivity was higher for all conditions. However, the increase in sensitivity as compared to using rational cutoff scores was relatively small, especially for the maximum longstring and the even-odd index. Response time had the highest sensitivity of all measures when empirical cutoffs were used. Mahalanobis distance showed low sensitivity for the condition with few careless respondents and relatively high sensitivity (43%) when there were 20% careless respondents in the sample. The sensitivity of the person-fit statistics G^p_m and l^p_{zm} were comparable, with slightly increasing sensitivity when the proportion of careless respondents became larger. Mahalanobis distance performed equally well as the person-fit statistics when the proportion of careless respondents was large.

5.1.1. Correlations between detection methods

Table 4 shows correlations between the different detection methods. These correlations were calculated using all empirical data from both the normal and the careless condition. From this table it can be concluded that correlations between the different methods were generally low. Moderate correlations were found between the two items and the even-odd index score and the Mahalanobis distance, G^p_m and l^p_{zm}. High correlations were found between Mahalanobis distance and the two person-fit statistics.

5.2. Simulated data

Sensitivity for the Mahalanobis distance, G^p_m, and l^p_{zm} are shown in Fig. 1, for different proportion of random respondents (5%, 10%, 20%) and proportion of random responses (10%, 25%, 50% of the items). In general, sensitivity was higher when responses were simulated uniformly random than when simulated normally random. Sensitivity increased when the proportion of random respondents and the proportion of random responses increased. For G^p_m, and l^p_{zm} the sensitivity was higher than for the Mahalanobis distance, but results were similar when the proportion of random respondents and random response behavior was large, which is in

Table 2
Cutoff scores, sensitivity and specificity based on rational cutoff scores.

Measure	Cutoff	Sensitivity	Specificity
Use me	1	0.49	0.86
Instructed item	1	0.81	0.29
Bogus item	1	0.51	0.60
Response time	6	0.21	0.97
Maximum longstring	7	0.19	0.94
Even-odd index	0.30	0.20	0.95

Note. Sensitivity is the mean proportion of respondents from the careless condition correctly identified as careless respondents. Specificity is the proportion of respondents from the normal condition correctly identified as normal respondents.

Table 3
Cutoff scores, sensitivity and specificity based on empirically derived cutoff scores.

Measure	Proportion careless					
	5%		10%		20%	
	Cutoff	Sensitivity	Cutoff	Sensitivity	Cutoff	Sensitivity
Response time	7	0.46 (0.14)	7	0.46 (0.09)	8	0.51 (0.04)
Maximum longstring	6	0.31 (0.13)	6	0.31 (0.09)	5	0.36 (0.04)
Even-odd index	0.14	0.20 (0.11)	0.52	0.25 (0.08)	0.75	0.30 (0.04)
Mahalanobis D	14.92 (0.18)	0.12 (0.09)	11.73 (0.10)	0.22 (0.08)	10.88 (0.04)	0.43 (0.04)
G^p_m	1.35 (0.09)[a]	0.31 (0.13)	0.84 (0.08)[a]	0.33 (0.09)	0.36 (0.04)[a]	0.40 (0.04)
lz^p_m	−1.09 (0.09)[a]	0.32 (0.13)	−0.76 (0.08)[a]	0.36 (0.09)	−0.38 (0.04)[a]	0.39 (0.04)

Note. Specificity was equal to the proportion of respondents from the normal condition by design. Exceptions were response time and the maximum longstring, because of the many ties in values. For response time specificity equaled 0.90, 0.90, 0.80 and for the maximum longstring 0.90, 0.90, 0.80, respectively. Standard deviations are between brackets.

[a] Based on standardized scores.

Table 4
Correlations between values on all detection methods.

Measure	1.	2.	3.	4.	5.	6.	7.	8.
1. Use me								
2. Instructed item	0.1*							
3. Bogus item	0.2*	0.4*						
4. Response time	0.2*		−0.1					
5. Max. longstring	0.3*	0.1	0.1*					
6. Even-odd index	−0.3*	−0.1*	−0.2*	−0.1*	−0.1			
7. Mahalanobis D	0.1		0.1	−0.1	−0.4*	−0.4*		
8. G_m^p	0.3*		0.2*	−0.1	−0.1	−0.5*	0.8*	
9. l_{zm}^p	−0.3*		−0.2*			0.4*	−0.7*	−0.9*

Note. Empty cells were correlations smaller than 0.1 or −0.1.
* $p < 0.05$.

agreement with the results from the empirical data. The differences between the two person-fit statistics were small, but l_{zm}^p yielded slightly higher sensitivity than G_m^p. In general, sensitivity for the simulated data was much higher than for the empirical data indicating that uniform random responses may not be representative for careless responses provided by actual respondents. This is interesting information because in many psychometric studies that investigate the sensitivity of person-fit statistics aberrant responses are simulated in a similar way. Results from these studies may thus overestimate the sensitivity as compared to when empirical data are used. Only when the simulated randomness was very low, we found lower sensitivity for the simulated data than for the empirical data. Sensitivity was also much higher than in the empirical data for simulated responses under a normal distribution when 50% of the responses were simulated randomly.

6. Discussion

The aim of this study was to investigate the usefulness of different methods to detect careless responses when web-based questionnaires of medium length were used. These types of questionnaires are often used for research purposes. We applied methods that were suggested in the literature on careless response data and we used person-fit statistics that were suggested in the psychometric literature. Doing so, we also tried to bring together different research fields in psychology that are both aimed at identifying "aberrant" response data. What did we learn from this study?

The bogus item, instructed response item, and self-reported data quality showed either low sensitivity or low specificity and did not perform well in our study, although we selected a bogus item that was previously found to function well (Meade & Craig, 2012). An explanation that was suggested by Meade and Craig (2012) is that respondents may find it amusing to answer these items in unexpected ways. Furthermore, as already noted above, perhaps we should have used more bogus item or instructed items so that accidental answering behavior on these items was diminished. However, using many of these items in a relatively short scale may result in a questionnaire that shows distrust in the honesty of a respondent's behavior and this may affect the way a respondent is approaching the questionnaire. So we think that the jury is still out on the use of many bogus items or instructed items in questionnaires of modest length.

The maximum longstring and response time showed relatively low sensitivity and specificity when rational cutoff scores were used; better results were obtained when cutoffs were estimated based on the data in the sample. Furthermore, our results using both the empirical data and the simulated data showed that the multi-test extensions of person-fit statistics could be used as an alternative to the other methods used to detect careless respondents with inconsistent response patterns. The sensitivity was somewhat higher than, for example, the even-odd index and Mahalanobis distance. There are, however, also some disadvantages to using person-fit statistics. We had to collapse response options before calculating person-fit statistics; doing so we may have lost some information in the data. Person-fit statistics are based on stricter models than other techniques and most applications using person-fit statistics assume that the number of response options is equal across items, although this is not strictly necessary. In addition, in order to use these statistics with missing values, some sort of imputation method must be used. On the positive side, in this paper we used empirically based cutoff score based on sample properties. Very recently, statistical properties of the distribution of person-fit statistics for polytomous data have been studied that can be used in future research (Sinharay, 2015). Researchers then can use these distributions to obtain a cutoff score and to decide when an item score pattern is very unlikely given the assumed underlying test model.

From our empirical and simulation study there are two take-home messages. The first take-home message is that the sensitivity of all methods discussed in this study is not high. Although this may be due to students that do not follow strict instructions, this may also reflect the fact that normal and careless response behavior is often not that different as often is assumed in many simulation studies. In addition, Huang et al. (2012) made an interesting remark about the sensitivity of methods to detect careless response behavior. They discussed that "Although the sensitivity of these (. . .) indices may appear unimpressive, we like to highlight the decision making context, which falls under what Swets (1992) described as industrial quality-control approach, where the cutoff is set at the tail of the distribution to yield extremely few false positives (i.e., normal responses misidentified as IER)". In light of this remark Huang et al. (2012) argued that the sensitivity and specificity was reasonably comparable to other administered examination such as breast cancer screening and employee drug testing.

A second take-home message related to the first is that some of the statistics that were advised in Meade and Craig (2012) and Huang et al. (2012) may function less optimal when personality questionnaires of, say, 100 items are used. The even-odd index was unreliable for questionnaires that do not contain many subscales. The maximum longstring is useful to detect excessively long strings of the same response option, but we found that this type of response pattern was rare. Furthermore, the bogus item, the instructed response item and a self-report measure of response quality did not function well in this study, as we explained above. An easy to apply statistic that seems to working rather well, also for questionnaires of moderate length, is response time. Although in this study we used response time that was required to complete the whole questionnaire, an alternative may be to use response time, for example, per page of a questionnaire. Person-fit statistics may be an alternative but are more complex to apply.

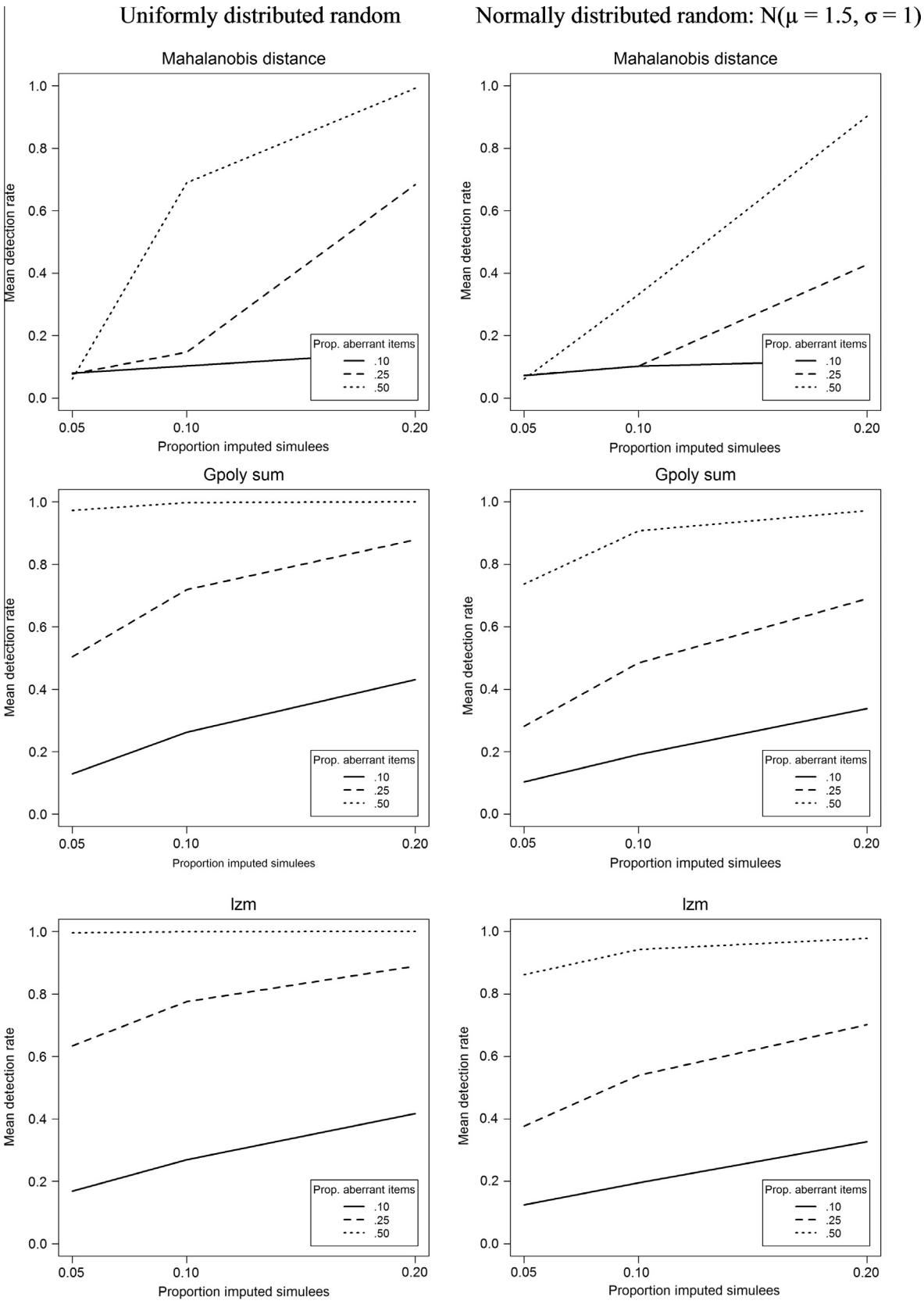


Fig. 1. Sensitivity for Mahalanobis distance, G_m^p and l_{zm}^p for each condition, with simulated responses under a uniform distribution and a normal distribution.

6.1. How should we handle careless respondents?

Detection of careless respondents is not easy. Different methods can be used that are sensitive to different types of careless behavior, and still the sensitivity may be low. We recommend using maximum longstrings to identify extreme cases of consecutive responses, and response time to identify unrealistically fast respondents as quick checks (although using a rational cutoff score yielded lower detection rates in our data). In addition, using person-fit seems to yield the best results to detect carelessness in the form of inconsistency in data collected with questionnaires of medium length. The summed number of Guttman errors and the summed l_z^p statistic performed similarly, but the summed number of Guttman errors poses fewer restrictions on the data. Therefore, this person-fit statistic may be preferred in future studies.

The context in which questionnaires are being used determines to a large extent how we can deal with careless responses. When tests or questionnaires are used for individual diagnostics, such as in clinical personality assessment, the knowledge that individuals have produced unexpected or careless responses can be discussed during interviews. For example, Meijer et al. (2008) described the measurement of self-concept of children between 8 and 12 years of age through self-report questionnaires. In that study, after having collected statistical information about the consistency of children's response behavior, the authors collected evidence through interviews about why some children responded in an idiosyncratic way. From these interviews it became clear that inconsistent responding was due to misunderstanding some of the instructions.

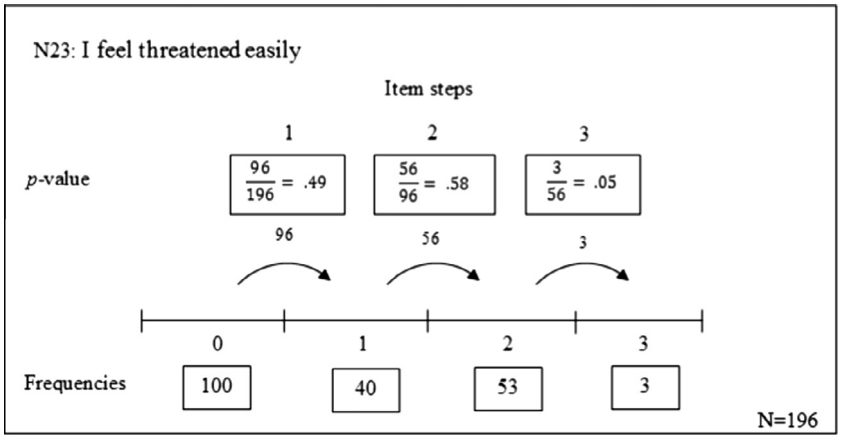
As a final warning we would like to stress that when the methods discussed in this study are used to detect careless responses for research purposes it is very important not to use these methods to blindly remove patterns to "clean" the data. Instead results should be reported on the entire dataset and on a cleaned dataset, so that any discrepancies can be discussed and readers can make a well-informed judgment about the quality of the dataset.

Perhaps a good strategy is to have researchers pre-register their plans for handling careless responding. So researchers in advance should then indicate which statistics they will use to handle careless responding. Furthermore, they should indicate which cutoff scores they will use. This will, however, not always be easy because the distribution of the statistic is not known in advance. For some person-fit statistics this distribution is known for dichotomous data (see, e.g., Meijer & Sijtsma, 2001; see also a recent paper by Sinharay, 2015, for polytomous data). Whether researchers use cutoff scores obtained from the sample or from a theoretical distribution when removing item score patterns from the data, researchers should always make it very explicit why they removed score patterns from the dataset.

Appendix A

Below we provide an example that illustrates how the number of Guttman errors is computed for polytomous data using the concept of item steps. This example uses three items from the neuroticism scale that is used in the empirical study and it is based on the sample of respondents who received normal instructions.

The figure below illustrates how for an individual item the popularity of the item steps (p-values) are calculated. Consider item N23 "I feel threatened easily". First, we look at the frequencies of the number of persons that obtained the different scores 0 through 4. So 100 respondents obtained a 0 score, 40 respondents obtained a 1 score, etc. Next, we compute the frequency from the data that each item step was taken (that is the number of persons who took the imaginary step from response option 0 to 1, from 1 to 2, etc.). Finally, we compute the proportion of respondents who took each item step (p-value). The p-value of an item step equals the number of persons who took the item step, divided by the number of people who took the previous item step.



A similar table can be constructed for each item in a questionnaire. Table A1 shows three neuroticism items and the p-values for each item step. After computing the p-values of each item step within each item, the item steps can be ordered according to the p-values across all items (Table A2). Next, for each respondent we determine whether each item step was taken (1) or not (0). For each respondent we can count the number of times that a less 'popular' item step was taken after not taking a more popular item step, resulting in the number of Guttman errors. Table A2 demonstrates this for two respondents. Respondent 12 never took a less popular item step while not taking a more popular item step,

Table A1
p-values for the item-steps of three neuroticism items.

Item	Content	Item step 1 1 - 2	Item step 2 2 - 3	Item step 3 3 - 4
N23	I feel threatened easily	0.49	0.58	0.05
N74	I worry about things	0.85	0.77	0.12
N81	I have frequent mood swings	0.53	0.58	0.12

Table A2
Item-step patterns for two respondents.

Person	Response pattern	Item steps									Gut. errors
		74-1	74-2	23-2	81-2	81-1	23-1	74-3	81-3	23-3	
12	1 - 2 - 1	1	0	0	0	0	0	0	0	0	0
161	4 - 3 - 1	1	1	1	0	0	1	0	0	1	6

Note. Item steps are ordered based on their *p*-value in descending order.

resulting in 0 Guttman errors. Respondent 161 took some less popular item steps and did not take some more popular item steps, resulting in 6 Guttman errors.

References

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology: Interdisciplinary and Applied, 123*, 101–103. http://dx.doi.org/10.1080/00223980.1989.10542966.

Conijn, J. M., Emons, W. H. M., & Sijtsma, K. (2014). Statistic lz based person-fit methods for non-cognitive multiscale measures. *Applied Psychological Measurement, 38*, 122–136. http://dx.doi.org/10.1177/0146621613497568.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. R., & McCrae, R. R. (2008). The revised NEO personality inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment. Personality measurement and testing* (Vol. 2, pp. 179–198). Thousand Oaks, CA, US: Sage Publications Inc.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171–191. http://dx.doi.org/10.1177/014662169101500207.

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. http://dx.doi.org/10.1111/j.2044-8317.1985.tb00817.x.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96. http://dx.doi.org/10.1016/j.jrp.2005.08.007.

Huang, J. S., Bowling, N. A., Mengqiao, L., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology, 30*, 299–311. http://dx.doi.org/10.1007/s10869-014-9357-6.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99–114. http://dx.doi.org/10.1007/s10869-011-9231-8.

Jackson, D. N. (1977). *Jackson vocational interest survey manual.* Port Huron, MI: Research Psychologists Press.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory-versions 4a and 54.* Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (pp. 114–158). New York, NY: Guilford Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103–129. http://dx.doi.org/10.1016/j.jrp.2004.09.009.

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

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76*, 315–332. http://dx.doi.org/10.1207/S15327752JPA7602_12.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*, 49–55.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. http://dx.doi.org/10.1016/j.jrp.2013.09.008.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437–455. http://dx.doi.org/10.1037/a0028085.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. 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. http://dx.doi.org/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. http://dx.doi.org/10.1177/014662169401800202.

Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A Practical guide to check the consistency of tem response patterns in clinical research through person-fit statistics: Examples and a computer program. *Assessment, 23*, 52–62. http://dx.doi.org/10.1177/1073191115577800.

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

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12*, 97–117.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109*, 330–357. http://dx.doi.org/10.1037/0033-295X.109.2.330.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Careless response in web-based questionnaires [Data file]. Available via:<http://hdl.handle.net/10411/20718>.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289–298. http://dx.doi.org/10.1177/0146621603027004004.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*.

Schmitt, N. W., & Stults, D. M. (1985). Factors defined by negatively keyed items: The results of careless respondents? *Applied Psychological Measurement, 9*, 367–373. http://dx.doi.org/10.1177/014662168500900405.

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

Sinharay, S. (2015). Asymptotically correct standardization of person-fit statistics beyond dichotomous items. *Psychometrika*. http://dx.doi.org/10.1007/s11336-015-9465-x.

Sprock, J. (2000). Invalid response sets in MMPI and MMPI-2 profiles of college students. *Educational and Psychological Measurement, 60*, 956–964. http://dx.doi.org/10.1177/00131640021971014.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47*, 522–532. http://dx.doi.org/10.1037/0003-066X.47.4.522.

Tendeiro, J. N. (2015). PerFit (R package version 1.4) [computer software] Available from:<http://cran.r-project.org/web/packages/PerFit/index.html>.

Tendeiro, J. N., & Meijer, R. R. (2014). The usefulness of simple nonparametric statistics. *Journal of Educational Measurement, 51*, 239–259. http://dx.doi.org/10.1111/jedm.12046.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (in press). PerFit: An R package for conducting person-fit analysis in IRT. *Journal of Statistical Software* (in press).

Zijlstra, W. P., van der Ark, L., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36*, 186–212. http://dx.doi.org/10.3102/1076998610366626.