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.

© 2016 Elsevier Inc. All rights reserved.

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/© 2016 Elsevier Inc. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/951214

Download Persian Version:

https://daneshyari.com/article/951214

Daneshyari.com