

Journal of Educational and Behavioral Statistics

<http://jeps.aera.net>

Outliers in Questionnaire Data : Can They Be Detected and Should They Be Removed?

Wobbe P. Zijlstra, L. Andries van der Ark and Klaas Sijtsma

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2011 36: 186

originally published online 8 December 2010

DOI: 10.3102/1076998610366263

The online version of this article can be found at:

<http://jeb.sagepub.com/content/36/2/186>

Published on behalf of



American Educational Research Association

and



<http://www.sagepublications.com>

Additional services and information for *Journal of Educational and Behavioral Statistics* can be found at:

Email Alerts: <http://jeps.aera.net/alerts>

Subscriptions: <http://jeps.aera.net/subscriptions>

Reprints: <http://www.aera.net/reprints>

Permissions: <http://www.aera.net/permissions>

Outliers in Questionnaire Data: Can They Be Detected and Should They Be Removed?

Wobbe P. Zijlstra
L. Andries van der Ark
Klaas Sijtsma
Tilburg University

Outliers in questionnaire data are unusual observations, which may bias statistical results, and outlier statistics may be used to detect such outliers. The authors investigated the effect outliers have on the specificity and the sensitivity of each of six different outlier statistics. The Mahalanobis distance and the item-pair based outlier statistics were found to have the best combination of specificity and sensitivity. Next, it was investigated how outliers influenced the bias in the percentile rank score, Cronbach's alpha, and the validity coefficient. Outliers due to random responding and faking produced considerable bias, and outliers due to extreme responding produced little bias. Finally, the influence of removing discordant observations on bias was studied. Removing observations due to random responding identified by means of the Mahalanobis distance, the local outlier factor, and the item-pair based outlier statistic reduced bias.

Keywords: *contaminant observations; discordancy testing; outlier detection in questionnaire data; outlier detection methods for questionnaire data; suspected patterns of item scores*

Introduction

The purposes of this study were to investigate the specificity and the sensitivity of six different methods for the detection of outliers in multi-item questionnaire data of the rating scale type; how contaminated observations affected important questionnaire statistics, which are the percentile rank score, Cronbach's alpha, and the validity coefficient; and how removal of observations identified by a particular outlier detection method from the data affected these questionnaire statistics.

Hawkins (1980, p. 1) defines an outlier as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.” Outliers may bias results from statistical analysis. This may lead to wrong conclusions. Thus, it is important that outliers are detected

before the data are analyzed. Barnett and Lewis (1994, pp. 34–43) recommend two ways of dealing with outliers. First, the outliers are identified and the data are analyzed without the outliers. The outliers may be analyzed separately as interesting cases. Second, robust statistics may be used to accommodate the outliers by minimizing their effect on the analysis (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986).

The current study focuses on outlier detection in data collected by means of multi-item questionnaires, typical of the measurement of attributes such as introversion (psychology), alienation (sociology), and health-related quality of life (medicine). Items used typically consist of a statement, such as “I feel uncomfortable among other people,” which may be used for measuring the personality trait of introversion, and a response format, which often has the form of an ordered scale with two or more answer categories. Item scores are discrete and ordered, and reflect the degree to which someone endorsed the statement. Attributes are measured using multiple items and there is no distinction between independent and dependent variables.

Outliers may arise when not only the intended attribute (e.g., introversion) drives responses to items but responses are also influenced by unintended attributes (e.g., vocabulary), causes related to group characteristics (e.g., gender), or response tendencies (e.g., extreme responding). These unwanted influences may contaminate the adequate measurement of the attribute of interest. Respondents’ item scores that were also affected by unwanted influences constitute the group of contaminant observations. The other respondents’ item-score vectors constitute the group of regular observations. The purpose of outlier detection typically is the identification of contaminant item-score vectors.

Many outlier statistics have been proposed for continuous data (e.g., Barnett & Lewis, 1994) and for linear regression models in which a distinction is made between independent and dependent variables (e.g., Atkinson & Riani, 2000; Chatterjee & Hadi, 1986; Rousseeuw & Leroy, 2003). Such methods cannot be used in multi-item questionnaire data, which are discrete and do not distinguish independent and dependent variables. Identifying outliers by means of contingency table analysis has been done for reasonably filled two-way tables (Yick & Lee, 1998), but a 12-item, 5-answer-category questionnaire requires a 12-way contingency table with $5^{12} = 2.44\text{e}8$ cells, most of which are empty. Hence, this approach will fail, and this procedure was also ignored here.

Hence, we investigated six outlier detection methods that seemed better suited here: the Mahalanobis distance (Mahalanobis, 1936), the local outlier factor (LOF; Breunig, Kriegel, Ng, & Sander, 2000), the item-based outlier statistic (Zijlstra, van der Ark, & Sijtsma, 2007), the item-pair based outlier statistic (Zijlstra et al., 2007), the intraindividual variance (e.g., Baumeister & Tice, 1988), and the extreme response style score (Bachman & O’Malley, 1984). An *outlier detection method* is the combination of an outlier statistic and a discordancy test. Each outlier detection method is based on a unique definition of an outlier.

Four simulation studies were done to investigate the performance of the six outlier detection methods. Three different types of contamination were simulated. Study 1 investigated the specificity of the outlier detection methods in questionnaire data that were not contaminated. Study 2 investigated the specificity and the sensitivity in questionnaire data sets that contained a mixture of regular and contaminant observations. Study 3 investigated the effect of contaminated observations on the bias in the percentile rank scores of the questionnaire's norm distribution, Cronbach's alpha, and the questionnaire's validity coefficient. Study 4 investigated the effect removal of discordant observations has on the bias in these dependent variables.

The article is organized as follows. First, we introduce the percentile rank scores, Cronbach's alpha, the validity coefficient, three types of contamination, and the six outlier detection methods. Second, for each of the four studies, we separately discuss the method and the results. Third, we used the outlier detection methods to analyze a real data set. Finally, we discuss the consequences of the results of this study.

Theory

Terminology

Because the literature does not unambiguously define the term "outlier," we use the following terminology. Sample observations stemming from the population of interest are *regular* observations, and sample observations stemming from another population, analyzed together with the regular observations, are *contaminant* observations. Each observation is assigned a score on an *outlier statistic*. Observations with high outlier scores are called *suspect* observations. A cutoff value is determined to test whether a suspect observation should be considered contaminant or regular. Such a test is a *discordancy test*. Observations that are tested positive are *discordant*, and observations that are tested negative are *not discordant*.

Notation

Let the questionnaire contain J items for measuring a particular attribute. Let X_j denote the random variable for the score on item j ($j \in \{1, \dots, J\}$), and let x_j ($x_j \in \{0, \dots, m\}$) be the realization of X_j . Let N be the sample size, and let \mathbf{X} be an $N \times J$ data matrix that consists of N item-score vectors of length J . Respondents are indexed v ($v \in \{1, \dots, N\}$). Henceforth, unless stated otherwise the term *observation* refers to a respondent's item-score vector, \mathbf{x}_v .

Quality Indices for Questionnaires

In practical questionnaire use, the J item scores, which measure different aspects of the same attribute, are added to obtain the total score, defined as

$X_+ = \sum_{j=1}^J X_j$. Interesting quality indicators for a questionnaire often quantify an aspect of total score, X_+ . Examples are the questionnaire's norm distribution and aspects of this distribution such as the percentile rank scores, the total score reliability estimated by means of Cronbach's alpha, and the questionnaire's validity.

The norm distribution is the distribution of total scores, which serves as benchmark for total-score interpretation (e.g., the Neuroticism-Extraversion-Openness Five-Factor-Inventory [NEO-FFI]; Costa & McCrae, 1992), but score transformations are also used regularly. For example, percentile rank scores provide information about the percentage of the norm sample, which falls at or below the respondent's total score. A contaminated norm sample may bias a respondent's percentile rank score. Consequently, the respondent may not be well diagnosed.

Cronbach's (1951) alpha coefficient is a much used lower bound to the reliability of the total score. Nunnally and Bernstein (1994, p. 265) suggested that for making decisions about individuals, total-score reliability should be at least .90, whereas for comparing groups, a reliability of at least .80 is adequate. A contaminated sample may bias Cronbach's alpha. For example, Barnette (1999) showed that Cronbach's alpha was estimated too low due to extreme responding and random responding.

In general, the validity coefficient is expressed by the correlation between the total score and a criterion score. Validity coefficients express many different types of validity, for example, convergent validity, discriminant validity, predictive validity, and concurrent validity (e.g., Nunnally & Bernstein, 1994, pp. 94–101). Contaminated samples may bias the validity coefficient (e.g., Mischel, 1968, pp. 83–87).

Contaminated Questionnaire Data

We investigated three influences known to have a contaminating effect on questionnaire data beyond the researcher's intentions (Nunnally & Bernstein, 1994, pp. 380–386). They are extreme response style (e.g., Bachman & O'Malley, 1984), random response style (e.g., Emons, 2008), and faking (e.g., McFarland & Ryan, 2000; Zickar & Drasgow, 1996). These influences have also been investigated in the context of person-fit analysis (Meijer & Sijtsma, 2001) but such methods are ignored here for several reasons. First, these methods assume the fit of an item response model (Van der Linden & Hambleton, 1997) to the data, and even though they can be quite flexible, these models impose a particular structure that captures only particular kinds of contamination but not others. Second, the wrong item response model may have been hypothesized, which may result in a confounding of misfit either due to the person being a contaminant or to the item response model being misspecified. Third, a technical limitation of person-fit methods is that they have not been developed well for polytomous items. Fourth, another technical limitation is that item-score vectors

consisting of only extreme item scores 0 or m cannot be analyzed but may well represent cases of contamination and should not be ignored. The methods studied here sometimes rest on assumptions about distributions but are less restrictive than item response models and thus provide a greater opportunity for studying contamination.

Respondents characterized by extreme responding have a tendency to choose the extreme answer categories scored 0 and m . Extreme responding is a stable trait (Bachman & O'Malley, 1984) and known to occur independent of item content and the respondent's trait level (Emons, 2008). Therefore, extreme responding is expected to affect many and sometimes all item scores. The researcher may confuse observations due to extreme responding with those characterized by many 0s or many scores equal to m driven by the attribute of interest.

Respondents characterized by random responding have a tendency to randomly pick an answer category. The resulting item-score vector is meaningless. Causes of random responding may be lack of motivation, careless responding, confusion, lack of comprehension (e.g., Nunnally & Bernstein, 1994, pp. 380–382), or lack of traitedness (e.g., Baumeister & Tice, 1988).

Respondents are faking when they try to present themselves more or less favorable than they really are. Faking is closely related to social desirability, which is the tendency to choose answer categories that reflect socially approved behaviors (Nunnally & Bernstein, 1994, p. 382). McFarland and Ryan (2000) asked respondents to fake all item scores as much as possible in a desired direction and found that total scores increased by almost two standard deviations. These authors suggested that respondents with lower trait values are more inclined to fake item scores. Zickar and Drasgow (1996) argued that high-scoring honest respondents are difficult to differentiate from respondents faking on many items. Faking is difficult to detect when few items are involved.

Outlier Detection Methods

Each of the six outlier statistics studied uniquely defines suspect behavior. Probably, the best known outlier statistic is the Mahalanobis distance (Mahalanobis, 1936). The other five outlier statistics are the LOF (Breunig et al., 2000), the item-based outlier statistic (Zijlstra et al., 2007), the item-pair based outlier statistic (Zijlstra et al., 2007), the intraindividual variance (e.g., Baumeister & Tice, 1988), and the extreme response style score (Bachman & O'Malley, 1984). High values identify suspect observations. Probably, the best known discordancy test is the boxplot or Tukey's fences (Tukey, 1977, pp. 43–44), but Zijlstra, van der Ark, and Sijtsma (2010) found that the *extreme studentized deviate* (ESD; Rosner, 1983) identified more contaminant observations than Tukey's fences and two other discordancy tests. Thus, we use the ESD in this study. In the next section, we briefly discuss the six outlier statistics and the ESD discordancy test.

Mahalanobis Distance

Let $\hat{\boldsymbol{\mu}}$ be the vector of J sample item means, and $\hat{\boldsymbol{\Sigma}}$ the sample covariance matrix for the J items. The Mahalanobis distance quantifies the distance between an observation \mathbf{x}_v and the center of the data when the correlational structure of the data is taken into account. It is denoted by MD_v^2 , and defined as

$$MD_v^2 = (\mathbf{x}_v - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_v - \hat{\boldsymbol{\mu}}).$$

If the J variables have a multivariate normal distribution, MD^2 follows a chi-square distribution with J degrees of freedom. An observation \mathbf{x}_v is classified as discordant when $MD_v^2 > \chi_{J,\alpha}^2$. Because item scores are discrete, this approach may not be appropriate here. Zijlstra et al. (2010) showed that for discrete item scores, the ESD discordancy test (Rosner, 1983), to be discussed shortly, produces specificity and sensitivity results similar to results obtained using the chi-square distribution. Contrary to the chi-square distribution, the ESD can be used for all outlier statistics investigated. To facilitate comparisons across outlier statistics, the ESD was used for discordancy testing.

LOF

According to the LOF (Breunig et al., 2000), suspect observations \mathbf{x}_v lie in a region with a relatively sparse density compared to the densities of the regions of the neighboring observations. Because it is unknown in the social and behavioral sciences, we explain the computation of the LOF in six steps; see Breunig et al. (2000) for more details.

First, for discrete item scores the *city-block* distance (e.g., Kaufman & Rousseeuw, 2005, p. 12) between observations \mathbf{x}_v and \mathbf{x}_w is determined, so as to leave the distance between, for example, item-score vectors (0, 0) and (0, 4) the same as between (0, 0) and (2, 2). It is defined as

$$d(v, w) = \sum |\mathbf{x}_v - \mathbf{x}_w|. \tag{1}$$

The *city-block* distance is symmetric, implying $d(v, w) = d(w, v)$. Without loss of generality, we choose $v = N$, and index the other $N - 1$ distances $w = 1, 2, \dots, N - 1$ in such a way that the ordering of distances $d(v, w)$ is defined as

$$d(v, 1) \leq d(v, 2) \leq \dots \leq d(v, N - 1). \tag{2}$$

Second, the *k-nearest neighborhood* of \mathbf{x}_v , denoted as $N_k(v)$, is defined. The k th nearest neighbor of \mathbf{x}_v is the observation with the k th smallest distance to \mathbf{x}_v ; that is, $d(v, k)$ (see Equation 2). For given k ($k = 1, \dots, N - 1$), observation \mathbf{x}_w belongs to neighborhood $N_k(v)$, if $d(v, w) \leq d(v, k)$. Hence, $N_k(v)$ is the set of $\#N_k(v)$ observations that are closest to \mathbf{x}_v : If two or more observations have a distance to \mathbf{x}_v equal to $d(v, k)$, then $\#N_k(v) > k$; and $\#N_k(v) = k$ otherwise.

Third, the *city-block* distance $d(v, w)$ is replaced by the *reachability distance*, $rd_k(v, w)$. For continuous variables and using the Euclidean distance for $d(v, w)$, Breunig et al. (2000) argued that this step reduces statistical fluctuations of $d(v, w)$ and, hence, of the mean and the standard deviation of $LOF_k(v)$; see the fifth step. We followed this line of reasoning for discrete item scores and using the *city-block* distance, even though we expect the statistical fluctuation of $d(v, w)$ to be smaller. The reachability distance between \mathbf{x}_v and \mathbf{x}_w is

$$rd_k(v, w) = \max[d(w, k), d(v, w)]. \quad (3)$$

This is an asymmetric distance measure. If \mathbf{x}_v is in a dense region and \mathbf{x}_w in a sparse region, then $rd_k(v, w) > rd_k(w, v)$.

Fourth, the *local reachability distance*, which is the mean of the reachability distances between observation \mathbf{x}_v and its k nearest neighbors, is defined as

$$lrd_k(v) = \frac{\sum_{w \in N_k(v)} rd_k(v, w)}{\#N_k(v)}.$$

A large value of $lrd_k(v)$ indicates that the observation \mathbf{x}_v is located in a sparse region.

Fifth, the LOF for a given value k is defined as

$$LOF_k(v) = \frac{\sum_{w \in N_k(v)} \frac{lrd_k(v)}{lrd_k(w)}}{\#N_k(v)}.$$

$LOF_k(v)$ compares the local reachability distance of observation \mathbf{x}_v to the local reachability distances of its k nearest neighbors. If $LOF_k(v)$ is close to 1, then $lrd_k(v)$ is approximately equal to the local reachability distances of its k nearest neighbors and, as a result, \mathbf{x}_v is not suspect. If $LOF_k(v) > 1$, then $lrd_k(v)$ is larger than the local reachability distances of its k nearest neighbors and \mathbf{x}_v may be suspect. The higher the $LOF_k(v)$ value, the more suspect observation \mathbf{x}_v .

Sixth, $LOF_k(v)$ is computed for a range of k . Breunig et al. (2000) proposed at least $k = 10$. In addition, the minimum of k for which $LOF_k(v)$ is computed should exceed the number of identical observations; else, $lrd_k(v) = 0$, which causes $LOF_k(v)$ to be uncomputable. We chose k maximally equal to 100, which is the maximum number of contaminant observations in this study. The maximum value of $LOF_k(v)$ for these values of k was the reported outlier statistic $LOF(v)$; that is,

$$LOF(v) = \max[LOF_{10}(v), LOF_{11}(v), \dots, LOF_{100}(v)].$$

Item-Based Outlier Statistic

Item-based outlier statistic O_+ (Zijlstra et al., 2007) counts the frequency of responses in unpopular answer categories. The modal answer category has outlier score 0, the next less popular answer category has outlier score 1, and so on; and

the least popular answer category has outlier score m . Let $P(X_j = x)$ be the proportion of responses in answer category x of item j , and let x_{vj} be the score of respondent v on item j . The outlier item score of respondent v , denoted by O_{vj} , is determined using the rank number of $P(X_j = x_{vj})$, denoted $\text{rank}[P(X_j = x_{vj})]$, such that

$$O_{vj} = (m + 1) - \text{rank}[P(X_j = x_{vj})]. \quad (4)$$

Respondent v 's score on the item-based outlier statistic, O_{v+} , is defined as

$$O_{v+} = \sum_{j=1}^J O_{vj}. \quad (5)$$

The distribution of O_+ is unknown, but Zijlstra et al. (2007) found it to be positively skewed in several real data sets. The association between outlier statistic O_+ and total score X_+ has a U-shape; that is, respondents with high O_+ values tend to have a low or a high total score (Zijlstra et al., 2007). As a result, suspect observations are likely to be found in the tails of the total score distribution.

Item-Pair Based Outlier Statistic

For two dichotomous items, a Guttman (1950) error occurs when a respondent responds positively to the most unpopular of the 2 items and negatively to the most popular item. For polytomous items, each item is assumed to be composed of m item steps (Molenaar, 1991, 1997), and Guttman errors are defined at the level of item step scores. Item-pair based outlier statistic G_+ (Zijlstra et al., 2007) counts the frequency of Guttman errors that occur in an item-score vector. Because it is relatively unknown, we explain the determination of G_+ for $J = 3$ and $m = 2$.

Item step popularities are defined as cumulative proportions $P(X_j \geq g)$; $g = 1, \dots, m$ because $P(X_j \geq 0) = 1$ by definition; thus, it is uninformative (Molenaar, 1991, 1997). Item steps of the same item have a fixed order. For 3 items, the $3m$ item step popularities are arranged by decreasing magnitude. A possible order is

$$P(X_3 \geq 1) \geq P(X_1 \geq 1) \geq P(X_1 \geq 2) \geq P(X_2 \geq 1) \geq P(X_3 \geq 2) \geq P(X_2 \geq 2). \quad (6)$$

Molenaar (1997) assumes that respondents take item steps in the order from most popular to most unpopular. Thus, a limited number of score patterns are allowed whereas the others represent violations of the ordering. These violations are the Guttman errors, and the exact number of these errors depends on the degree to which the item step ordering was violated. For example, in $\mathbf{x}_v = (1, 2, 0)$ score $X_1 = 1$ was obtained by passing item step $X_1 \geq 1$ and failing the less popular step $X_1 \geq 2$; score $X_2 = 2$ by passing both steps of Item 2; and score $X_3 = 0$ by failing

both steps of item 3. Following the ordering in Equation 6, \mathbf{x}_v thus represents the following pattern of passed and failed item steps:

$$\text{failed, passed, failed, passed, failed, passed.} \tag{7}$$

Outlier statistic G_+ can be computed as follows. Equation 7 is written as vector \mathbf{Z}_v with $Z_{vi} = 1$ if an item step was passed, and $Z_{vi} = 0$ if an item step was failed. Thus, we obtain $\mathbf{Z}_v = (0, 1, 0, 1, 0, 1)$. Next, for all pairs of item steps, it is evaluated whether the less popular item step has been passed and the more popular item step has been missed. G_+ counts these errors (Meijer, 1994; Meijer & Sijtsma, 2001); that is

$$G_{v+} = \sum_{i=2}^{J \times m} \left\{ Z_{vi} \times \left[\sum_{j=1}^{i-1} (1 - Z_{vj}) \right] \right\}. \tag{8}$$

For observation $\mathbf{x}_v = (1, 2, 0)$, we thus find that $G_{v+} = 6$. The properties of the distribution of G_+ are unknown. Zijlstra et al. (2007) found the distribution to be positively skewed in several real data sets. Respondents with either a small or a large total score X_+ cannot have large G_+ values. As a result, such respondents are not likely to be identified as suspect.

Intraindividual Variance

Austin, Deary, Gibson, McGregor, and Dent (1998) and Baumeister and Tice (1988) considered an item-score vector with many different scores to be suspect and proposed to identify such item-score vectors by means of the variance of the J item scores. For respondent v , let the mean item score be denoted by \bar{X}_v , then the variance of the item scores is

$$S_v^2 = \frac{1}{J} \sum_{j=1}^J (X_{vj} - \bar{X}_v)^2.$$

Baumeister and Tice (1988) argued that respondents with low or high total scores X_+ cannot have a large intraindividual variance S^2 . As a result, they are likely not identified as suspect. One may argue that item-score vectors for which $S^2 = 0$ also are suspect. Because it is readily checked whether this kind of response behavior has occurred, we focus on the case of positive variance.

Extreme Response Style Score

Bachman and O'Malley (1984) suggested assigning an extremity score $E_j = 1$ if item scores are in one of the two extreme answer categories, 0 and m , and else $E_j = 0$. For respondent v , the extreme response style score on J items is defined as

$$E_{v+} = \sum_{j=1}^J E_{vj}.$$

Respondents having a low or high total score by definition have many extreme scores, but without additional information, it cannot be decided whether this is due to an extreme response style or to little or much endorsement of the items.

ESD Discordancy Test

The ESD (Rosner, 1983) tests whether suspect values of an outlier statistic are discordant. Let the generic notation U_v denote the value of an outlier statistic, \bar{U} the sample mean, and S_U the sample standard deviation; then, the ESD is defined as

$$\text{ESD} = \frac{\max |U_v - \bar{U}|}{S_U}, \quad (9)$$

and is assumed to have a standard normal distribution. Barnett and Lewis (1994, p. 131) recommend the *outward consecutive testing* procedure. Following Zijlstra et al. (2007), we implemented this procedure as follows. So as not to miss discordant observations, we selected the highest $(N - 1)/2$ values of U —just under half the sample size—as the suspect observations to be tested for discordancy. Testing started with the least deviating suspect observation. If this observation was tested discordant, all other, more extreme suspect observations were also labeled discordant, and the procedure stopped. If the first observation was tested not discordant, it was not considered suspect anymore, and the next larger suspect observation was tested for discordancy. Testing was continued while proceeding further into the tail of the U distribution until a suspect observation was tested discordant or the suspect observation that deviates the most was tested not discordant. When a particular value of U was observed multiple times, only one of these scores was tested for discordancy.

Statistics \bar{U} and S_U were computed anew in each step, using the sample consisting of the unsuspect observations, the suspect observations that appeared not discordant, and the observation to be tested for discordancy. Because \bar{U} and S_U were computed in a subsample having at most one discordant observation, these statistics may be regarded as robust estimators. For each step in the outward consecutive testing procedure, the ESD classified a respondent v as either discordant if $\text{ESD}_v > \text{ESD}^*$, or not discordant if $\text{ESD}_v \leq \text{ESD}^*$. The critical value of the ESD was set at $\text{ESD}^* = 2.5758$. Provided that U has a normal distribution and the sample consists of regular observations only, $\text{ESD}^* = 2.5758$ results in a nominal specificity of .99. This nominal specificity seems to be reasonable for most applications of questionnaires. Measurement using questionnaire data may not be accurate enough to justify a larger nominal specificity; for example, for a larger specificity, ESD^* may exceed the largest possible value of an outlier

statistic. A smaller nominal specificity may be chosen depending on the researcher's intentions.

Study 1: Specificity in Regular Data

Data without contaminants were used as worst case for investigating the specificity of the outlier detection methods by means of a Monte Carlo study. By definition, each observation that was tested discordant is a misclassification.

Method

In each design cell, the graded response model (Samejima, 1997) was used to generate 1,000 data sets, each containing the J item scores of $N = 500$ regular respondents. Latent variable values θ were sampled from $N(0, 1)$. For item j , a_j is the discrimination parameter and b_{jx} the location parameter of answer category x ; then, the graded response model is

$$P(X_j \geq x|\theta_v) = \frac{\exp [a_j(\theta_v - b_{jx})]}{1 + \exp [a_j(\theta_v - b_{jx})]}, \text{ for } x = 1, \dots, m. \quad (10)$$

For latent variable value θ_v and item parameters a_j and b_{jx} , we obtained $m + 1$ probabilities, $P(X_j = x|\theta_v)$, for $x = 0, \dots, m$. Item scores were randomly sampled from a multinomial distribution with probabilities $P(X_j = x|\theta_v)$. This was repeated for all sampled θ s and all J items, and resulted in a complete data matrix.

Based on the NEO-FFI (Costa & McCrae, 1992), we chose $J = 12$ items with $m + 1 = 5$ answer categories. The item parameters were based on estimates reported by Embretson and Reise (2000, p. 101). The discrimination parameter was fixed for all items: $a_j = a = 1.3$. The location parameters were split into two parts; that is, $\mathbf{b}_j = (b_{j1}, b_{j2}, b_{j3}, b_{j4})$, $b_{jx} = \lambda_j + \varepsilon_{jx}$, with $\lambda_j = -1$ for $j = 1, \dots, 4$; $\lambda_j = 0$ for $j = 5, \dots, 8$; and $\lambda_j = 1$ for $j = 9, \dots, 12$. The values of ε_{jx} were: $\boldsymbol{\varepsilon}_j = (-1.25, -0.25, 0.25, 1.25)$ for $j = 1, 5, 9$; $\boldsymbol{\varepsilon}_j = (-1.875, -0.375, 0.375, 1.875)$ for $j = 2, 6, 10$; $\boldsymbol{\varepsilon}_j = (-2.5, -0.5, 0.5, 2.5)$ for $j = 3, 7, 11$; and $\boldsymbol{\varepsilon}_j = (-3.125, -0.625, 0.625, 3.125)$ for $j = 4, 8, 12$. To avoid confounding effects of asymmetry, the mean location parameters and the distances between two adjacent answer categories were chosen to be symmetric.

The independent variable was outlier detection method with the six methods as levels. The dependent variable was the specificity, which equaled the ratio of the number of regular observations that were tested not discordant and the total number of regular observations (Selvin, 2004, pp. 69–74). Given the nominal specificity of .99, we used the following ad hoc rules of thumb for interpretation: $> .999$ is "very large," $.995$ – $.999$ is "large," $.980$ to $.995$ is "expected," $.925$ to $.980$ is "small," and $< .925$ is "very small." Analysis of variance (ANOVA) and Tukey's honestly significant difference (HSD) procedure were used to test whether the mean specificity was equal for all outlier detection methods.

Specificity and sensitivity in Studies 1 and 2 and Cronbach's alpha and the correlation coefficient in Studies 3 and 4 are bounded, which may result in violations of the assumptions of normality and homogeneity of variance in ANOVA. However, for none of the studies did visual inspection show gross deviations of either assumption. Moreover, our studies had large sample size and equal sample size per design cell and under these conditions ANOVA is robust to violations of both assumptions (Maxwell & Delaney, 1990, pp. 111–114).

Effect sizes were classified as follows (Cohen, 1988, pp. 284–288): $\eta^2 > .01$ is a small effect, $\eta^2 > .06$ a medium effect, and $\eta^2 > .14$ a large effect. Cohen (1988) derived the effect qualifications for the comparison of two groups but this is different from the current research. However, consistent with much of the literature and for easy communication, we prefer to maintain the qualifications but notice that an effect is small relative to a medium effect, which is medium relative to a large effect but that these qualifications do not have an absolute interpretation. These relative interpretations were also used in the other three studies.

Results and Discussion

The mean specificity of the statistics LOF (.939), E_+ (.960), MD^2 (.968), and G_+ (.973) was small, and the mean specificity of O_+ (.982), and S^2 (.993) was as expected. ANOVA showed that the mean specificity was not the same for the six methods, $F(5, 5999) = 2455, p < .001$. The effect of outlier detection method was large ($\eta^2 = .672$). All pairwise comparisons (Tukey's HSD) were significant ($p < .001$). We conclude that S^2 had the best specificity.

Study 2: Specificity and Sensitivity in Contaminated Data

We investigated the specificity and the sensitivity of the outlier detection methods for contaminated data by means of a Monte Carlo study.

Method

We used samples of regular observations and samples consisting of $(1 - \pi) \times N$ item-score vectors of regular respondents and $\pi \times N$ item-score vectors of contaminant respondents (π is a proportion). Regular item scores were generated as in Study 1. Contaminant item scores were the results of either extreme responding, random responding, or faking, which were three different conditions in the design.

Extreme Response Style. Contaminated item scores were generated using the graded response model (Equation 10). The distances between location parameters ε_{jx} and $\varepsilon_{j,x+1}$ were reduced by 60% (also, see Emons, 2008); thus, respondents are more likely to score 0 when $\theta < \lambda_j$ and m when $\theta > \lambda_j$. The values of ε_{jx} were: $\varepsilon_j = (-0.5, -0.1, 0.1, 0.5)$ for $j = 1, 5, 9$; $\varepsilon_j = (-0.75, -0.15, 0.15, 0.75)$

for $j = 2, 6, 10$; $\boldsymbol{\epsilon}_j = (-1.0, -0.2, 0.2, 1.0)$ for $j = 3, 7, 11$; and $\boldsymbol{\epsilon}_j = (-1.25, -0.25, 0.25, 1.25)$ for $j = 4, 8, 12$.

Random response style. Contaminated item scores were drawn from a multinomial distribution with score probabilities $P(X_j = x) = .2$ for $x = 0, \dots, 4$.

Faking. Contaminated item scores were generated using the graded response model (Equation 10). We only simulated faking that makes one look more favorable. A constant value of 2 was subtracted from λ_j (for $j = 2, 3, 6, 7, 10, 11$); thus, faking respondents are likely to score higher on these items than regular respondents with the same θ s. Faking respondents had $\theta < 0$.

The three independent variables were *outlier detection method* (6 levels), *type of contamination* (3 levels: extreme responding, random responding, and faking), and *proportion of contamination* (3 levels: $\pi = .05, .10, .20$). In each design cell, 1,000 data sets with sample size $N = 500$ were generated.

For the definition of specificity, see Study 1. The sensitivity equaled the ratio of the number of contaminant observations that were tested discordant and the total number of contaminant observations (Selvin, 2004, pp. 69–74). Sensitivity was interpreted as follows: $> .8$ is very large, $.6 - .8$ large, $.4 - .6$ moderate, $.2 - .4$ small, and $< .2$ very small. Full factorial ANOVAs were done to investigate the effects of the outlier detection methods, and the type and proportion of contamination on the specificity and the sensitivity.

Type and proportion of contamination did not have a “no contamination” level, consisting of regular observations only, because this would result in a confounded experimental design. Instead, planned comparisons were made on the specificity between regular observations and each level of type of contamination. Three levels of type of contamination combined with one dependent variable resulted in three planned comparisons. Without contamination, sensitivity cannot be determined; in this condition, planned comparisons were not performed. The total sum of squares used for the computation of the effect sizes (η^2) in the planned comparisons was computed from the cells involved in the planned comparison. As a result, the total sums of squares may differ among planned comparisons.

Results

For both the specificity and the sensitivity, all full factorial ANOVA effects were significant ($p < .001$). Therefore, we only discuss the effects with an effect size of $\eta^2 > .01$.

Specificity

The results for the ANOVA were as follows. The main effect of outlier detection methods ($\eta^2 = .503$) was similar to the effect found in Study 1. This large effect dominated all interaction effects in which outlier detection methods were

TABLE 1
Mean Sensitivity of Six Outlier Detection Methods for Type and Proportion of Contamination (π)

Type	π	MD ²	LOF	O_+	G_+	S^2	E_+
Extreme	.05	.519	.527	.158	.386	.439	.684
	.10	.468	.454	.112	.333	.407	.633
	.20	.340	.304	.055	.230	.290	.365
Random	.05	.775	.864	.336	.788	.188	.122
	.10	.744	.837	.269	.764	.157	.095
	.20	.618	.619	.109	.651	.094	.056
Faking	.05	.085	.122	.007	.117	.038	.023
	.10	.051	.079	.005	.079	.032	.022
	.20	.032	.057	.006	.042	.025	.020

Note: LOF = local outlier factor.

included. For all types of contamination and all proportions of contamination, S^2 had the largest specificity (moderate for $\pi = .05$ and large for $\pi = .10$ and $\pi = .20$). Statistic LOF had the lowest specificity, which was small in all cases.

The interaction effect of proportion of contamination and type of contamination was small ($\eta^2 = .026$). For extreme responding and random responding, specificity increased as proportion of contamination increased. For faking, this effect was absent. The main effect of type of contamination was small ($\eta^2 = .054$). When aggregated over outlier detection methods and proportions of contamination, specificity for extreme responding and random responding was as expected, but for faking it was small. The main effect of the proportion of contamination was also small ($\eta^2 = .044$). In general, specificity increased as proportion of contamination grew.

The planned comparisons showed that the inclusion of extreme responding ($\eta^2 = .070$) and random responding ($\eta^2 = .062$) resulted in a higher specificity compared to fully regular samples. Including faking had no discernible effect on specificity ($\eta^2 = .001$).

Sensitivity

The results for the ANOVA were as follows. Table 1 shows the mean sensitivity for each cell in the design. For extreme responding, statistic E_+ had the largest sensitivity, which was either small or large. Statistics MD², LOF, G_+ , and S^2 had small to moderate sensitivity, and statistic O_+ very small sensitivity. For random responding, sensitivity for statistic LOF was large to very large, for statistics MD² and G_+ large, for statistic O_+ small to very small, and for statistics

S^2 and E_+ very small. For faking, all outlier detection methods had very small sensitivity.

A large two-way interaction effect ($\eta^2 = .277$) was found between type of contamination and outlier detection method. Sensitivity for outlier statistics S^2 and E_+ was largest for extreme responding and much lower for random responding and faking (Table 1). Sensitivity for statistics MD^2 , LOF, G_+ , and O_+ was largest for random responding and smallest for faking. Of all statistics, statistic E_+ identified extreme responding best and statistic LOF identified random responding and faking best.

Random responding had the largest sensitivity, followed by extreme responding and faking ($\eta^2 = .395$). Sensitivity was largest for statistic LOF, followed by statistics MD^2 , G_+ , E_+ , S^2 , and O_+ , respectively ($\eta^2 = .182$). Sensitivity decreased as proportion of contamination grew ($\eta^2 = .036$). This effect is due to masking (e.g., Barnett & Lewis, 1994, p. 97); that is, due to the presence of a large number of contaminant observations, the suspect observation under investigation often is incorrectly tested not discordant.

Discussion

Not surprisingly, the specificity in contaminated samples was larger than in regular samples. The outlier detection methods had a larger effect on specificity than proportion of contamination. This effect was similar to that for regular observations only (Study 1). Hence, also in the presence of contaminated observations did S^2 have the largest and LOF the smallest specificity. Sensitivity varied greatly among the design cells. No method was best for all contamination types. For extreme responding, statistic E_+ may be used, and for random responding G_+ , and not LOF due to its small specificity. None of the methods detected faking well. Statistic O_+ in general failed detecting contamination. In real data analysis, when the researcher does not know whether data are contaminated, G_+ and MD^2 have the best specificity and sensitivity.

Study 3: Effect of Contamination on Statistics

The effect of type and proportion of contamination on percentile rank scores, Cronbach's alpha, and the validity coefficient was investigated using a Monte Carlo study.

Method

The samples contained both regular and contaminant item scores; see the previous studies for details. Planned comparisons were made between regular observations and each type of contamination on bias in percentile rank scores, bias in

Cronbach's alpha, and bias in the validity coefficient. These dependent variables were defined as follows.

Percentile Rank Scores. Let $PR(X_+ = x_+)$ be the percentage of scores in the norm sample that fall at or below score x_+ ; $PR^R(X_+ = x_+)$ the percentile rank score in a sample of regular observations; and $PR^C(X_+ = x_+)$ the percentile rank score in a contaminated sample. The mean of absolute differences of $PR^R(X_+ = x_+)$ and $PR^C(X_+ = x_+)$ across all values of $x_+ = 0, \dots, J \times m$, and denoted MAD, expresses the mean absolute effect of the contamination on a percentile rank score,

$$MAD = \frac{1}{J \times (m + 1)} \sum_{x_+=0}^{J \times m} |PR^R(X_+ = x_+) - PR^C(X_+ = x_+)|. \quad (11)$$

The bias is obtained by comparing the MAD in Equation 11 with the MAD in a completely regular sample, which equals 0 by definition. Hence, the mean value of the MAD across replicated samples, denoted \overline{MAD} , may be interpreted as the estimated bias in MAD. We reported \overline{MAD} across 1,000 replications.

Cronbach's Alpha. Let $Cov(X_j, X_k)$ denote the sample covariance between items j and k , and $S_{X_+}^2$ the sample variance of X_+ ; then Cronbach's alpha is defined as

$$\text{alpha} = \frac{J}{J - 1} \frac{\sum_{j \neq k} Cov(X_j, X_k)}{S_{X_+}^2}.$$

The mean alpha in the regular sample was .836. We reported the estimated bias in alpha, which equals the mean difference across 1,000 replications of alpha in the contaminated sample and alpha in the regular sample.

Validity Coefficient. The validity coefficient was the correlation between total score X_+ and criterion measure Y . Criterion measure Y was generated from a standard normal distribution. We assumed that Y was unaffected by contamination. The population correlation between Y and latent variable θ was .5, which is a value commonly found in psychological research (Cohen, 1988, p. 80). The mean validity coefficient in the regular sample was .459. We reported the estimated bias, which equaled the mean difference across 1,000 replications of the validity coefficient in the contaminated sample and the validity coefficient in the regular sample.

The total number of paired comparisons was 3 (levels of type of contamination) \times 3 (dependent variables) = 9. If an effect was found, we did post hoc comparisons using Tukey's HSD procedure among the three proportions of contamination ($\pi = .05, .10, .20$). In each design cell, 1,000 data sets with sample size $N = 500$ were generated.

Results and Discussion

The first column containing results in Table 2 (No removal) shows $\overline{\text{MAD}}$, the bias in Cronbach's alpha, and the bias in the validity coefficient. The remainder of Table 2 contains results for Study 4, to be discussed later.

Percentile Rank Scores

In all three planned comparisons, type of contamination had a large effect on $\overline{\text{MAD}}$. For the extreme response sample, the overall result was $\overline{\text{MAD}} = 0.555$ (this is the mean of 0.271, 0.489, 0.906; Table 2, first column, 1st to 3rd row; $\eta^2 = .47$). For the random response sample, the overall result was $\overline{\text{MAD}} = 0.737$ (this is the mean of 0.347, 0.651, 1.214; Table 2, first column, 4th to 6th row; $\eta^2 = .46$). Finally, for the faking sample, the overall result was $\overline{\text{MAD}} = 1.836$ (this is the mean of 0.786, 1.576, 3.147; Table 2, first column, 7th to 9th row; $\eta^2 = .46$). All post hoc tests were significant ($p < .001$), and the corresponding effect sizes were large, showing that $\overline{\text{MAD}}$ increased as proportion of contamination grew.

The $\overline{\text{MAD}}$ is a summary statistic, which obscures information that is relevant for individual diagnosis. A closer look at the effect of contamination on percentile rank score revealed the following. For $\pi = .20$, Figure 1 shows the effect of type of contamination on percentile rank score. The horizontal axis shows the percentile rank score in the regular sample, and the vertical axis shows the bias in the contaminated samples. In the extreme response sample (solid line), the 50th percentile rank score was unbiased but lower percentile rank scores were positively biased and higher percentile rank scores were negatively biased. For example, a respondent who had a percentile rank score of 25 in the regular sample (Figure 1, vertical dashed line) got an expected percentile rank score in the extreme response sample of $25 + 1.2 = 26.2$.

Compared to the extreme response sample, the effect in the random response sample (dashed line) was the opposite and larger. For example, a respondent who had a percentile rank score of 25 in the regular sample received an expected percentile rank score in the random response sample of $25 - 2.7 = 22.3$. In the faking sample (dotted line), respondents obtained smaller percentile rank scores, especially when X_+ was close to the median. For example, for a respondent who received a percentile rank score of 25 in the regular sample, the expected percentile rank score in the faking sample was $25 - 7.9 = 17.1$. For $\pi = .05$ and $\pi = .10$, the effects of type of contamination were smaller.

The three response styles had large effects on the distribution of X_+ , and effects increased as proportion of contamination was larger. In research in which X_+ is compared to a norm distribution, the respondent may have a different percentile rank score when the norm distribution is contaminated.

TABLE 2

MAD, Bias in Cronbach's Alpha, and Bias in Validity Coefficient for "No Removal" and Six Removal Procedures for Type and Proportion of Contamination (π)

Type	π	No Removal	MD ²	LOF	O_+	G_+	S^2	E_+
MAD								
None	0	0	0.173	0.286	0.438	0.188	0.070	0.853
Extreme	.05	0.271	0.393	0.473	0.387	0.443	0.386	0.817
	.10	0.489	0.619	0.696	0.441	0.693	0.690	0.757
	.20	0.906	1.100	1.140	0.741	1.161	1.274	0.631
Random	.05	0.347	0.258	0.335	0.549	0.261	0.290	1.016
	.10	0.651	0.339	0.383	0.683	0.332	0.529	1.158
	.20	1.214	0.590	0.710	1.178	0.551	1.107	1.564
Faking	.05	0.786	0.808	0.795	1.188	0.795	0.788	1.376
	.10	1.576	1.612	1.565	2.052	1.580	1.577	2.018
	.20	3.147	3.249	3.126	3.680	3.179	3.155	3.425
Cronbach's alpha								
None	0	.836 ^a	.008	.012	-.011	.009	.003	-.026
Extreme	.05	-.001	.011	.014	-.012	.011	.007	-.026
	.10	-.002	.013	.017	-.011	.013	.010	-.026
	.20	-.005	.015	.015	-.012	.013	.013	-.023
Random	.05	-.017	.005	.010	-.019	.006	-.011	-.040
	.10	-.036	.000	.006	-.031	.002	-.027	-.054
	.20	-.077	-.017	-.020	-.068	-.013	-.066	-.091
Faking	.05	-.008	.000	.004	-.020	.001	-.006	-.035
	.10	-.018	-.009	-.005	-.032	-.009	-.016	-.045
	.20	-.040	-.033	-.026	-.058	-.031	-.038	-.068
Validity coefficient								
None	0	.459 ^a	.001	.001	-.018	.002	.000	-.034
Extreme	.05	-.002	.001	.002	-.015	.001	.001	-.033
	.10	-.003	.004	.005	-.013	.004	.002	-.034
	.20	-.006	.001	.003	-.012	.002	.001	-.027
Random	.05	-.016	-.005	.000	-.025	-.004	-.013	-.044
	.10	-.034	-.009	-.007	-.035	-.010	-.028	-.056
	.20	-.066	-.028	-.029	-.063	-.030	-.062	-.082
Faking	.05	-.015	-.011	-.010	-.028	-.010	-.012	-.048
	.10	-.024	-.023	-.022	-.044	-.021	-.026	-.059
	.20	-.046	-.045	-.042	-.059	-.041	-.045	-.074

Note: LOF = local outlier factor.

^a The mean value over the 1,000 replications.

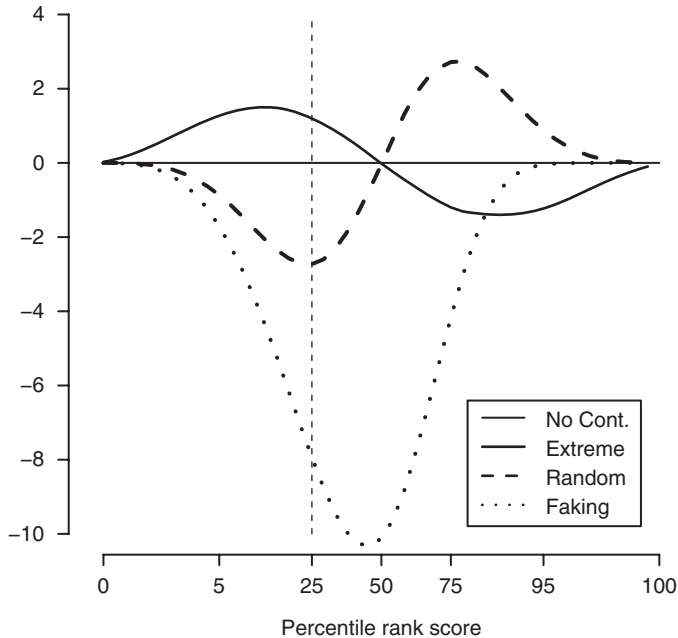


FIGURE 1. Influence of the three different response styles on the percentile rank score for $\pi = .20$ (see text for further explanation).

Cronbach's Alpha

Table 2 (middle panel, first column) shows that Cronbach's alpha decreased as proportion of contamination increased. The mean bias in Cronbach's alpha was $-.003$ in the extreme response sample ($\eta^2 = .013$); $-.043$ in the random response sample ($\eta^2 = .358$); and $-.022$ in the faking sample ($\eta^2 = .248$). For random responding and faking, all post hoc tests were significant ($p < .001$), and corresponding effect sizes were large. For extreme responding, the post hoc tests were significant ($p < .05$ for $\pi = .05$ vs. $\pi = .10$; and $p < .001$ for the two remaining post hoc tests). Corresponding effect sizes were small.

The presence of respondents exhibiting random responding or faking producing a negative bias in Cronbach's alpha may lead the researcher to conclude incorrectly that the questionnaire is too unreliable for comparing groups. The results for random responding were similar to those found by Barnette (1999). Extreme responding hardly affected Cronbach's alpha.

Validity Coefficient

Extreme responding did not have a discernible effect on the validity coefficient ($\eta^2 = .002$). Random responding had a large effect (bias = $-.039$;

$\eta^2 = .144$), and faking a medium effect (bias = $-.028$; $\eta^2 = .093$). The post hoc tests for random responding and faking were significant ($p < .001$), and the corresponding effect sizes ranged from small ($\pi = .05$ vs. $\pi = .10$) to large ($\pi = .05$ vs. $\pi = .20$). The validity coefficient decreased as proportion of contamination increased.

Study 4: Effect of Removing Discordant Observations

The effect of removing discordant observations on the percentile rank scores, Cronbach's alpha, and the validity coefficient was investigated using a Monte Carlo study.

Method

First, for the regular observations (see Studies 1–3), planned comparisons were made between the “no removal” condition and the removal conditions for each of the six outlier detection methods. The dependent variables were \overline{MAD} , bias in Cronbach's alpha, and bias in the validity coefficient (see Study 3). This resulted in 6 (Independent Variables) \times 3 (Dependent Variables) = 18 planned comparisons. Second, for each level of proportion of contamination and type of contamination, we compared the “no removal” condition to the removal conditions for each of the six outlier detection methods. Again, the dependent variables were bias in Cronbach's alpha and bias in the validity coefficient. This resulted in 3 \times 3 \times 6 (Independent Variables) \times 3 (Dependent Variables) = 162 planned comparisons. In each design cell, 1,000 data sets with sample size $N = 500$ were generated. We only discuss the planned comparisons showing interesting results.

Results and Discussion

Percentile Rank Score

No contamination. (Table 2, 1st row, upper panel). All effects were large ($\eta^2 > .6$). Removing observations identified by statistic E_+ resulted in the largest bias ($\overline{MAD} = 0.853$), followed by statistics O_+ ($\overline{MAD} = 0.438$), LOF ($\overline{MAD} = 0.286$), G_+ ($\overline{MAD} = 0.188$), MD^2 ($\overline{MAD} = 0.173$), and S^2 ($\overline{MAD} = 0.070$). Obviously, removing discordant observations from a regular sample may have a damaging effect and should be avoided.

Extreme response style. (Table 2, 2nd to 4th row, upper panel). Removal procedures MD^2 , LOF, G_+ , and S^2 had large effects resulting in larger bias than under condition “no removal.” Statistic E_+ explicitly identifies observations due to extreme responding, and produced the largest bias for $\pi = .05$ and $\pi = .10$ but the smallest bias for $\pi = .20$. Removal procedure O_+ resulted in smaller bias

for $\pi = .10$ and $\pi = .20$ than “no removal.” In general, removal of discordant observations likely results in larger bias. We advocate not to remove discordant observations in this case.

Random response style. (Table 2, 5th to 7th row, upper panel). For all levels of π , removing observations identified by MD^2 and G_+ produced a smaller bias, which represented a large effect size. Removal procedure E_+ produced a larger bias for all levels of π (i.e., a large effect size). For the remaining removal procedures, the effects were much smaller. Hence, both removal procedures MD^2 and G_+ can be used to reduce bias due to random responding.

Faking. (Table 2, 8th to 10th row, upper panel). All removal procedures resulted in a larger bias than “no removal.” The effect of removal conditions O_+ and E_+ was large, the effect of the other removal conditions was small or negligible.

Cronbach's Alpha

No contamination. (Table 2, 1st row, middle panel). Removal procedures MD^2 , LOF, G_+ , and S^2 produced positive bias, and O_+ and E_+ produced negative bias. The effect of S^2 was small, the effect of MD^2 medium, and the remaining effects were large. Removal procedure E_+ resulted in a bias ($-.026$) that may be too large to be acceptable to researchers. For the remaining removal procedures, the resulting bias was negligible. Zijlstra et al. (2007) found similar effects for O_+ and G_+ .

Extreme response style. (Table 2, 2nd to 4th row, middle panel). Removal of discordant observations always resulted in a larger absolute bias than “no removal.” Three of the effects were medium, and the remaining 15 effects were large.

Random response style. (Table 2, 5th to 7th row, middle panel). Only removal procedure E_+ resulted in a larger bias than “no removal,” and all its effects were large. The largest bias in Cronbach's alpha ($-.091$) was due to procedure E_+ for 20% contamination. For removal procedures MD^2 , LOF, and G_+ , all effects were also large but these procedures resulted in the smallest bias.

Faking. (Table 2, 8th to 10th row, middle panel). Removal procedures E_+ and O_+ resulted in a larger bias than “no removal,” MD^2 , LOF, and G_+ resulted in a smaller bias, and S^2 had no effect. For E_+ , O_+ , and LOF, the effects were large, for G_+ the effects were medium, and for MD^2 , the effects were either small or medium.

Validity Coefficient

No contamination. (Table 2, 1st row, lower panel). Without contamination, removal procedures MD^2 , LOF, G_+ , and S^2 did not influence the validity coefficient. Removing observations identified by O_+ and E_+ caused the validity coefficient to decrease by .018 (medium effect) and .034 (large effect), respectively.

Extreme response style. (Table 2, 2nd to 4th row, lower panel). Removal procedures MD^2 , LOF, O_+ , G_+ , and S^2 had no or small effects on the validity coefficient compared to “no removal.” For removal procedure E_+ , the effects were medium and large, which resulted in a larger bias than “no removal.”

Random response style. (Table 2, 5th to 7th row, lower panel). As with Cronbach’s alpha, only removal procedure E_+ resulted in a larger bias than “no removal.” The effects ranged from small to medium. Removal procedures MD^2 , LOF, and G_+ resulted in a smaller bias than “no removal.” The effects were small for $\pi = .05$, medium for $\pi = .10$, and large for $\pi = .20$. We recommend removal procedures MD^2 , LOF, and G_+ for bias reduction in samples contaminated by random responding. Removal procedures O_+ and S^2 had no discernible effects.

Faking. For removal procedures O_+ and E_+ , the effects ranged from small to large and resulted in a larger bias. None of the other removal procedures had discernible effects.

Discussion

In general, removal procedures MD^2 , LOF, and G_+ produced the smallest bias, followed by “no removal” and S^2 . Removal procedures MD^2 , LOF, and G_+ can have a large bias-reducing effect. Removal procedures O_+ and E_+ usually do more harm than good. Removal of data due to faking or extreme responding usually does not contribute to bias reduction, but removal of data due to random responding produces data for which procedures MD^2 , LOF, and G_+ produce a large bias reduction. Proportion of contamination did not have a substantial effect. Only removal of discordant observations due to random responding increasingly reduces bias as proportion of contaminants increases.

Real Data Example

The cognitive subscale of the Health Complaint Scale (HCS; Denollet, 1994) is a 12-item self-report questionnaire measuring cognitive health complaints. Each item contains a statement expressing an anxious concern about health (e.g., “being afraid of illness”) or the extent to which illness interferes with one’s

life (e.g., “not being able to work fluently”). Respondents express the degree to which they have been bothered lately by the concern in one of the five ordered answer categories, *not at all*, *a little bit*, *moderately*, *quite a bit*, and *extremely*. Item scores are 0, 1, 2, 3, and 4, respectively. The total score, X_+ , which ranges from 0 to 48, quantifies cognitive health complaints. The sample consisted of $N = 633$ patients who underwent cardioverter-defibrillator (ICD) implantation. Missing values (0.42% of all scores) were replaced by scores resulting from two-way imputation (Van Ginkel & Van der Ark, 2005). Cronbach’s alpha of X_+ was .947. The correlation of X_+ and the total score on *negative affectivity* (Denollet, 2005), which was used as a coefficient for construct validity, was $r = .629$. The six outlier detection methods were used to study the effect of removal of discordant observations on the percentile rank scores, Cronbach’s alpha, and the validity coefficient.

Method LOF identified most discordant observations (190). This is consistent with the simulation studies in which LOF had the lowest specificity. Method E_+ did not identify discordant observations. This is caused by the large number of 0 scores in the sample (34%), yielding many high E_+ values. Methods MD^2 , G_+ , S^2 , and O_+ identified 78, 76, 70, and 20 discordant observations, respectively. MD^2 , G_+ , and S^2 correlated highly ($r > .89$) and the corresponding outlier detection methods identified 48 discordant item-score vectors containing many 0s and 4s without much structure. This lack of structure may be due to a combination of extreme and random responding, when respondents indicate that they are extremely bothered by some concerns but not by highly related concerns. Such observations may be candidates for removal from the main analysis. Outlier score O_+ and X_+ correlated $r = .98$: Respondents who report serious cognitive health complaints are relatively rare and, therefore, O_+ tends to identify them as discordant, but because the questionnaire measures cognitive health complaints O_+ should not be used here.

Based on Study 4, discordant observations should be removed when resulting from random responding but preferably not extreme responding. We had no way to tell the difference in the HCS data, so we decided to ignore the difference here. Removal was based on MD^2 and G_+ because the simulations showed they were effective methods. Removal had a negligible effect on Cronbach’s alpha, which may be due to a ceiling effect. Validity increased by a respectable .04 units (Table 3). Percentile rank score shifted on average by 2 points. For low total scores (6–24; not tabulated), the shift was positive, up to 5 percentage points, and for total scores outside this range the shift was small negative, up to 0.5 percentage points.

Discussion and Conclusions

The results of the Studies 1 and 2, in which the specificity and the sensitivity of six outlier detection methods were investigated, are summarized in Table 4 (first and second row). Statistics MD^2 and G_+ had the best combination of

TABLE 3

Effect of Removing #D Discordant Observations Identified by Outlier Detection Methods MD^2 and G_+ , and the Combination of These Methods on MAD, Cronbach's Alpha (Alpha), and the Validity Coefficient (r_{xy})

Outlier Detection Method	#D	MAD	Alpha	r_{xy}
MD^2 and G_+	64	1.867	.013	.033
MD^2	78	2.490	.013	.042
G_+	76	2.017	.015	.041

TABLE 4

Summary of Performance of Six Outlier Detection Methods With Respect to Specificity, Sensitivity, and Bias Reduction

	MD^2	LOF	O_+	G_+	S^2	E_+
Specificity		-			+	
Sensitivity	+	+	-	+	-	-
Bias reduction	+	+	-	+		-

Note: LOF = local outlier factor; + = method advocated; - = method not advocated; a blank indicates a neutral advice.

specificity and sensitivity. In Study 3, we investigated the effect of three types of contamination on important statistics used in assessing the psychometric quality of questionnaires. The results of Study 3, in which the effect of three types of contamination on important psychometric statistics was investigated, are summarized in Table 5. In general, random responding and faking resulted in more bias than extreme responding. Bias increased as proportion of contamination grew.

Only in simulated data does one know for certain whether an observation is a contaminant. In real data, one can only determine the number of discordant observations and the influence of removing these observations on several statistics (cf. Zijlstra et al., 2007), but one does not know what caused the discordant observations. Thus, the results obtained from this simulation study should be interpreted with care. Some of our results may be compared to results obtained with the real data example and real data sets by Zijlstra et al. (2007), who found, for example, that removal of item-score vectors based on O_+ and G_+ had the same effect on Cronbach's alpha.

The results of Study 4, in which influence of removal of discordant observations on interesting statistics was investigated, are summarized in Table 4 (last row). Statistics MD^2 , LOF, and G_+ performed best; that is, sometimes removal of observations resulted in a small but negligible increase in bias and sometimes

TABLE 5

Summary of Bias of Three Types of Unusual Response Behavior on the Percentile Rank Scores, Cronbach's Alpha, and Validity Coefficient

	Extreme	Random	Faking
Percentile rank scores	–	–	--
Cronbach's alpha	0	--	–
Validity coefficient	0	--	–

Note: 0 = small to no bias; – = medium bias; -- = large bias.

in a large reduction of bias. Removing discordant observations based on O_+ and E_+ increased the bias.

The outlier statistics are based on useful definitions of suspect observations, and their primary objective is to identify such observations but not to reduce bias in a particular statistic. The researcher must decide whether discordant observations should be removed. It may be noted that, only if both specificity and sensitivity equal 1 does removal of discordant observations reduce bias for sure. In all other cases, misclassification (i.e., regular observations that were erroneously removed and contaminant observations that were erroneously maintained) may affect bias of statistics in different ways.

An alternative to removing discordant observations is the use of robust statistics (e.g., Hampel et al., 1986; Rousseeuw & Leroy, 2003) to accommodate statistics for the presence of outliers. Assuming a multivariate continuous distribution of the data, robust versions of Cronbach's alpha (e.g., Christmann & Van Aelst, 2006) and the correlation coefficient (e.g., Devlin, Gnanadesikan, & Kettenring, 1975) are available. The development of robust statistics and their comparison with the methods discussed in this study is a topic of future research.

Acknowledgment

The authors are grateful to Krista C. van den Broek for making available the data for the real data example.

References

- Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York, NY: Springer.
- Austin, E. J., Deary, I. J., Gibson, G. J., McGregor, M. J., & Dent, J. B. (1998). Individual response spread in self-reported scales: Personality correlations and consequences. *Personality and Individual Differences*, 24, 421–438.
- Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, 48, 491–509.
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York, NY: John Wiley.

- Barnette, J. J. (1999). Nonattending respondent effects on internal consistency of self-administered surveys: A Monte Carlo simulation study. *Educational and Psychological Measurement, 59*, 38–46.
- Baumeister, R. F., & Tice, D. M. (1988). Metatracts. *Journal of Personality, 56*, 571–598.
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *SIGMOD Record, 29*, 93–104.
- Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science, 1*, 379–416.
- Christmann, A., & Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. *Journal of Multivariate Analysis, 97*, 1660–1674.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Costa, P. T., & McCrae, R. R. (1992). *The NEO personality inventory and the NEO five factor inventory professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
- Denollet, J. (1994). Health complaints and outcome assessment in coronary heart disease. *Psychosomatic Medicine, 56*, 463–474.
- Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and type D personality. *Psychosomatic Medicine, 67*, 89–97.
- Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlational coefficients. *Biometrika, 62*, 531–545.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Emons, W. H. M. (2008). Person-fit analysis of polytomous items. *Applied Psychological Measurement, 32*, 224–247.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 60–90). Princeton, NJ: Princeton University Press.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York, NY: John Wiley.
- Hawkins, D. M. (1980). *Identification of outliers*. London, England: Chapman and Hall.
- Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley.
- Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*, 49–55.
- Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.
- McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812–821.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*, 311–314.
- Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.
- Mischel, W. (1968). *Personality and assessment*. New York, NY: John Wiley.

- Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multcategory items. *Kwantitatieve Methoden*, 12(37), 97–117.
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 369–380). New York, NY: Springer.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.
- Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25, 165–172.
- Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection*. Hoboken, NJ: John Wiley.
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 85–100). New York, NY: Springer.
- Selvin, S. (2004). *Statistical analysis of epidemiological data*. New York, NY: Oxford University Press.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of item response theory*. New York, NY: Springer.
- Van Ginkel, J. R., & Van der Ark, L. A. (2005). SPSS syntax for missing value imputation in test and questionnaire data. *Applied Psychological Measurement*, 29, 152–153.
- Yick, J. S., & Lee, A. H. (1998). Unmasking outliers in two-way contingency tables. *Computational Statistics & Data Analysis*, 29, 69–79.
- Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, 20, 71–87.
- Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research*, 42, 531–555.
- Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2010). *Discordancy tests for outlier detection in multi-item surveys*. Manuscript submitted for publication.

Authors

- WOBBE P. ZIJLSTRA is post doctoral researcher at the Department of Methodology and Statistics and the Center of Research on Psychology in Somatic diseases (CoRPS), Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands; e-mail: w.p.zijlstra@uvt.nl. His research interests are outliers and medical psychological research.
- L. ANDRIES VAN DER ARK is associated professor at the Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands; e-mail: a.vdark@uvt.nl. His research interests are item response theory, latent class analysis, and missing data analysis.
- KLAAS SIJTSMA is full professor at the Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands; e-mail: k.sijtsma@uvt.nl. His research interest is measurement of individual differences.

Manuscript submitted June 11, 2009

Revised November 13, 2009

Accepted January 15, 2010