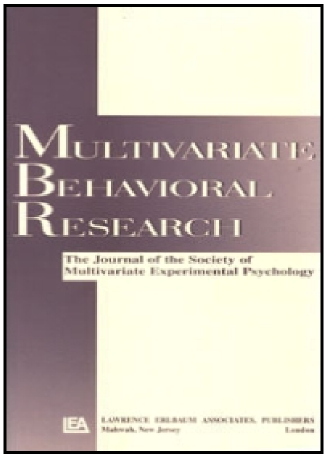
Outlier detection in test and questionnaire data

Zijlstra, W.P.; van der Ark, L.A.; Sijtsma, K.

Published in:
Multivariate Behavioral Research

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record



Multivariate Behavioral Research
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t775653673

Outlier Detection in Test and Questionnaire Data

Wobbe P. Zijlstra [a]; L. Andries van der Ark [a]; Klaas Sijtsma [a]
[a] Tilburg University,

Online Publication Date: 10 October 2007
To cite this Article: Zijlstra, Wobbe P., van der Ark, L. Andries and Sijtsma, Klaas (2007) 'Outlier Detection in Test and Questionnaire Data', Multivariate Behavioral Research, 42:3, 531 - 555
To link to this article: DOI: 10.1080/00273170701384340
URL: http://dx.doi.org/10.1080/00273170701384340

MULTIVARIATE BEHAVIORAL RESEARCH, *42*(3), 531–555

Outlier Detection in Test and Questionnaire Data

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

Classical methods for detecting outliers deal with continuous variables. These methods are not readily applicable to categorical data, such as incorrect/correct scores (0/1) and ordered rating scale scores (e.g., $0, \ldots, 4$) typical of multi-item tests and questionnaires. This study proposes two definitions of outlier scores suited for categorical data. One definition combines information on outliers from scores on all the items in the test, and the other definition combines information from all pairs of item scores. For a particular item-score vector, an outlier score expresses the degree to which the item-score vector is unusual. For ten real-data sets, the distribution of each of the two outlier scores is inspected by means of Tukey's fences and the extreme studentized deviate procedure. It is investigated whether the outliers that are identified are influential with respect to the statistical analysis performed on these data. Recommendations are given for outlier identification and accommodation in test and questionnaire data.

INTRODUCTION

Outliers are often identified as observations or subsets of observations which appear to be inconsistent with the remainder of the data (Barnett & Lewis, 1994, p. 7). Such observations are of interest in particular when they exercise a disproportionate influence on the outcome of the statistical analysis of one's data. For example, compared to a data analysis without the outlying observations, one that includes these outliers may result in means that shift further to the left or the right, correlations that are higher or lower, and regression coefficients that are biased. Obviously, such influential observations should be identified and a

Correspondence concerning this article should be addressed to: Wobbe P. Zijlstra, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. E-mail: w.p.zijlstra@uvt.nl

decision should be taken about their role in the statistical data analysis. In this paper, we discuss outliers in the context of test and questionnaire data, that are typically collected in psychological, sociological, educational, and political science research.

An obvious sequence for identifying and dealing with outliers is the following. First, one starts by noting that several observations in the data are unusual or suspected (cf., Barnett & Lewis, 1994, p. 1) given what one would expect. We refer to such observations as *suspected observations* (Iglewicz & Hoaglin, 1993, p. 30). Notice that many authors use different terms; for example, Barnett and Lewis (1994, p. 297) use the term *suspicious observations* and Hadi and Simonoff (1993) use the term *potential outliers*. Second, a formal decision is made whether the suspected observations are indeed different from the remainder of the data. Such a decision may be based on a *discordancy test* (Barnett & Lewis, 1994, chap. 4). When it is decided that a suspected observation is discordant it is called an *outlier*. Third, a decision should be made what to do with outliers. Three possibilities are the following. The first possibility is to investigate the influence of the outliers by analyzing the data with and without them. Outliers exercising a disproportionate influence on the statistical results are referred to as *influential observations*. For example, it may be decided that the influential observations are deleted from the data. The second possibility is to accommodate the outliers. This entails the construction of statistics that are robust with respect to outliers, or a transformation of the data. The third possibility is to conclude that the outlying cases are representative of a group that was misrepresented in the sample, and it may be decided that a new sample should be collected based on the appropriate stratification. Alternatively, outliers may be studied as separate interesting cases.

Barnett and Lewis (1994, pp. 33–34) distinguish three ways for outliers to arise in a sample. In their terminology these are:

1. Measurement error: Outliers arise for deterministic reasons, for example, due to a reading error, a recording error, or a calculation error in the data;
2. Execution error in collecting the data: Individuals that do not belong to the population envisaged are included in the sample (such outliers are called *contaminants*); and
3. Inherent variability: Outliers are merely rare events that are perfectly reasonable given the model at hand.

Much research has been done into outlier detection for continuous variables and variables with many categories. Barnett and Lewis (1994) and Rousseeuw and Leroy (2003) provide many references until 1994 (for more recent sources, see, e.g., Atkinson & Riani, 2000; Chambers, Hentges, & Zhao, 2004; and Cook & Critchley, 2000). This is different when a variable has only few, say, no more than five, categories. Such variables are typical of psychological tests and questionnaires, and are called "items" in that context. Let X_j denote the random variable for the score on item j ($j = 1, \ldots, J$), and let x_j be a realization of X_j. For example, many educational tests contain J items that are dichotomously scored as either correct ($x_j = 1$) or incorrect ($x_j = 0$). Similar correct/incorrect scoring can be found in intelligence testing. With only two score categories, defining observations as suspected and testing for discordancy may be problematic or, at least, awkward. For example, if 60% of the respondents give a correct answer to item j and 40% an incorrect answer, would one conclude that people in this latter group give a surprising response and that the group consists entirely of suspected observations? Similarly, many questionnaires used for personality testing in psychology or attitude testing in sociological or political science research contain rating scales to which ordered polytomous scores are assigned. Ordering means that a higher score indicates a higher level of endorsement with a statement about one's personality or the attitude under investigation. Polytomous items typically are scored $x_j = 0, \ldots, m$. The well known Likert items have five ordered answer categories ($m = 4$). In general, $2 \leq m \leq 6$, but larger values of m are sometimes encountered in practice, and other scoring schemes also may be used. If for one item with $m = 4$ a score distribution is found like $(.20, .42, .18, .12, .08)$, are the 8% 4-scores all suspected observations? Or the 20% 3- and 4-scores together? Or the 20% 0-scores and the 8% 4-scores?

In the context of categorical variables outliers have not been studied frequently. One exception is the study of outliers in contingency tables (e.g., Kotze & Hawkins, 1984; Lee & Yick, 1999; Simonoff, 1988; Yick & Lee, 1998). The J item scores produced by N respondents may be collected in a J-way contingency table. Thus far, only outliers in reasonably filled two-way (i.e., $J = 2$) contingency tables have been studied. Most psychological tests have $J > 10$, resulting in sparse J-way contingency tables. For example, if $J = 10$ and $m = 4$, then the contingency table has $5^{10} = 9,765,625$ cells. Even with a large sample most cells are empty and the available approaches for outlier detection in contingency tables fail. Hodge and Austin (2004) called this the 'curse of dimensionality'. An elaboration of the contingency table approach is used in data mining techniques in computer sciences (see Hodge & Austin, 2004, for an overview), where this approach is applicable to continuous and categorical data. The approach is based on the distances between the observations but also suffers from the 'curse of dimensionality'.

Another exception is person-fit analysis. An early attempt was due to Levine and Rubin (1979), who studied the appropriateness of a vector of J binary item scores by means of its likelihood in the 2-parameter logistic model (e.g., Van der Linden & Hambleton, 1997). More generally, person-fit analysis studies the fit of item response models to an individual's item-score vector or evaluates the fit of an item-score vector in a group under consideration (see Meijer & Sijtsma, 2001, for an extensive overview). The decision to categorize sets of observations such as item-score vectors as either fitting or misfitting is known as outlier identification (cf., Barnett & Lewis, 1994, p. 7). Recent examples can be found in certification testing (Meijer, 2002) and adaptive testing (Bradlow & Weiss, 2001; Bradlow, Weiss, & Cho, 1998). In person-fit analysis, the interest is mainly with identifying aberrant item-score vectors and inferring the cause of this aberrance, for example, for diagnostic reasons (e.g., did the respondent understand the test instruction? Did he or she suffer from test anxiety?). Furthermore, person-fit analysis is model based and therefore its application is more involved. In the present study, the interest is with the sample and making valid inferences about the population by using simple indices.

We propose a new approach to outlier analysis in which we use *outlier scores* as indices for identifying suspected observations. The first outlier score is defined as an individual's frequency of unpopular item scores in his/her vector of J item scores; for polytomous items this definition is a little more involved than for binary items. This is explained later on. The rationale for this outlier score is that for some tests or questionnaires, a respondent's item scores are suspected if he or she often chooses unpopular answer categories. The second outlier score is the number of weighed Guttman (1950) errors; such an error in combinations of binary item scores occurs each time a respondent answers a relatively difficult item correctly and an easier item incorrectly. The rationale for this outlier score is that a respondent's item scores are suspected if he or she has many score combinations that contradict the order of the items according to difficulty. This idea is also useful with polytomous items. For ten real-data sets, the distributions of the two outlier scores were inspected using both Tukey's fences and the extreme studentized deviate procedure. Also, the influence of the identified outliers on several statistics was investigated. Recommendations are given for the use of outlier detection methods in the analysis of real test and questionnaire data.

METHODS OF OUTLIER DETECTION

Outlier Scores

Item-based outlier score. The idea behind the item-based outlier score, O_+, is that responses to the modal (most popular) score categories of items are not suspected, responses to the next less popular score category are more suspected, and so on; and responses to the least popular score category are the most suspected. We assume that each item in the test or questionnaire has an equal number of ordered answer categories, and that adjacent ordered integer scores

$x = 0, \ldots, m$ represent this ordering. Note that for dichotomous item scores $m = 1$. Proportions of answers in score categories are denoted by $P(X_j = x)$, and the score distribution of item j is denoted by $[P(X_j = 0), \ldots, P(X_j = m)]$.

Outlier item-score, O_j, equals 0 for the modal (i.e., the most popular) category, $O_j = 1$ for the next less popular category, and so on; and $O_j = m$ for the least popular category. Assume that respondent v has item score x_{vj}. Then, his/her outlier score, O_{vj}, is determined using the rank number of $P(X_j = x_{vj})$, denoted rank$[P(X_j = x_{vj})]$, such that

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

For respondent v, the outlier item-scores are added across items to obtain item-based outlier score O_{v+}:

$$O_{v+} = \sum_{j=1}^{J} O_{vj}. \tag{2}$$

As an example, for $J = 5$ and $m + 1 = 3$, Table 1 shows the frequency distributions for each of the items. Let $\mathbf{X}_v = (X_{v1}, \ldots, X_{vJ})$ and let $\mathbf{x}_v$ contain the J item scores of respondent v. Assume that respondent v has item-score vector $\mathbf{x}_v = (2, 2, 2, 1, 1)$. For item 1, the third category ($X_{v1} = 2$) is modal and thus has rank 3. Using Equation 1, it follows that $O_{v1} = (2 + 1) - 3 = 0$. For item 2, the third category ($X_{v2} = 2$) is the least popular and thus has rank 1; hence $O_{v2} = (2 + 1) - 1 = 2$. Similarly, it follows that $O_{v3} = 0$, $O_{v4} = \frac{1}{2}$, and $O_{v5} = 0$. Using Equation 2, respondent v has item-based outlier score $O_{v+} = 2\frac{1}{2}$ (see the last column of Table 1). The item-score vector that produces the maximum value of O_+ is denoted $\mathbf{x}_{max}$; here, $\mathbf{x}_{max} = (1, 2, 0, 0, 0)$ and the corresponding outlier score equals $O_+ = 10$.

Item-pair based outlier score. Another approach to outlier detection uses the information contained in pairs of items. Consider polytomously scored items indexed j and k. Define the proportion of respondents that have at least a score of g on item j, $P(X_j \geq g)$; likewise, define proportion $P(X_k \geq h)$. Because by definition, for $g = h = 0$ the proportions $P(X_j \geq 0) = P(X_k \geq 0) = 1$ (see Table 1), they do not contain useful information and are left out of consideration.

For item pair (j, k), determine the common, decreasing ordering of the proportions $P(X_j \geq g)$ and $P(X_k \geq h)$, for $g, h = 1, \ldots, m$. For example, for items 1 and 2 ($m = 2$) in Table 1 the common ordering of the proportions is,

$$P(X_1 \geq 1) \geq P(X_1 \geq 2) \geq P(X_2 \geq 1) \geq P(X_2 \geq 2). \tag{3}$$

TABLE 1
Examples of Item Category Proportions [$P(X_j = x)$] of Five Items With Three Answer Categories Each, the Rank of the $P(X_j = x)$'s, the Item-Based Outlier Score (O_j) (Equation 1) for Each Answer Category, the Cumulative Item Category Proportions [$P(X_j \geq x)$], and the O_{vj} Scores for Item-Score Vectors $\mathbf{x}_v = (2, 2, 2, 1, 1)$ and $\mathbf{x}_{max} = (1, 2, 0, 0, 0)$. The Last Column Shows O_{v+}

	Item 1			*Item 2*			*Item 3*			*Item 4*			*Item 5*			
x	*0*	*1*	*2*	*0*	*1*	*2*	*0*	*1*	*2*	*0*	*1*	*2*	*0*	*1*	*2*	O_{v+}
$P(X_j = x)$	.3	.2	.5	.6	.3	.1	.1	.4	.5	.2	.4	.4	.2	.5	.3	
rank[$P(X_j = x)$]	2	1	3	3	2	1	1	2	3	1	$2\frac{1}{2}$	$2\frac{1}{2}$	1	3	2	
O_j	1	2	0	0	1	2	2	1	0	2	$\frac{1}{2}$	$\frac{1}{2}$	2	0	1	
$P(X_j \geq x)$	1.0	.7	.5	1.0	.4	.1	1.0	.9	.5	1.0	.8	.4	1.0	.8	.3	
O_{vj}		0			2			0			$\frac{1}{2}$			0		$2\frac{1}{2}$
$O_{max,j}$		2			2			2			2			2		10

Item-pair based outlier scores use weighed Guttman errors in polytomous item scores (Molenaar, 1991). Such errors are defined on the common ordering of proportions from different items as in Equation 3. Based on this ordering, item-pair scores can represent either Guttman errors (i.e., score pairs that disagree with the Guttman model) or conformal patterns (i.e., score pairs that agree with the Guttman model). For example, score pair $(X_1, X_2) = (1, 0)$ is a conformal pattern: Given the ordering in Equation 3, the event that one has a score of at least 1 on item 1 is more likely than the event of having a score of at least 1 on item 2, because $P(X_1 \geq 1) = .7$ exceeds $P(X_2 \geq 1) = .4$ (Table 1). Following the same line of reasoning, score pair (0, 1) is a Guttman error because having at least a score of 1 on item 2 is less likely than having a score of at least 1 on item 1, and $X_1 = 0$ contradicts this ordering. Taking the common ordering of the proportions into account, one may check that the conformal patterns are $(0, 0), (1, 0), (2, 0), (2, 1)$, and $(2, 2)$, and that the Guttman errors are $(0, 1), (0, 2), (1, 1)$, and $(1, 2)$.

A helpful metaphor may result from considering the ordering in Equation 3 as a staircase which is climbed from left ("easy") to right ("difficult"). A respondent who produced a Guttman error on the items j and k is assumed to have missed one or more steps, which expresses the idea that he or she partly "ignored" the common ordering. Molenaar (1991) proposed weighing each Guttman error for the number of steps missed. For each step respondent v takes, previously missed steps—if any—are counted, and the total number of steps missed equals the weight assigned to the Guttman error; this weight is denoted w_{vjk}.

As a first example of counting errors, consider the Guttman error $(X_1, X_2) = (1, 1)$. Starting from conformal pattern (0, 0), the steps taken to achieve $(1, 1)$ are $X_1 \geq 1$ and $X_2 \geq 1$. Given the ordering in Equation 3, all steps preceding $X_1 \geq 1$ have been taken, so the number of previous steps missed equals zero. However, one step preceding $X_2 \geq 1$ [i.e., step $X_1 \geq 2$] should have been taken but was missed. As a result, the weight given to Guttman error $(1, 1)$ is $w_{v12} = 0 + 1 = 1$.

As a second example, consider the Guttman error $(X_1, X_2) = (0, 2)$. Starting from conformal pattern (0, 0), the steps taken to achieve $(0, 2)$ are $X_2 \geq 1$ and $X_2 \geq 2$. Given the ordering in Equation 3, two steps preceding $X_2 \geq 1$ should have been taken but were missed [i.e., steps $X_1 \geq 1$ and $X_1 \geq 2$]. The same two steps preceding $X_2 \geq 2$ should have been taken but were also missed. Thus, the weight assigned to Guttman error $(0, 2)$ is $w_{v12} = 2 + 2 = 4$.

Respondent v may either produce or not produce a Guttman error on item pair (j, k). This results in Guttman score $G_{vjk} = 1$ or $G_{vjk} = 0$, respectively. Weighing G_{vjk} by error count w_{vjk} and adding across all (j, k) combinations, yields for a given respondent v the item-pair based outlier score

$$G_{v+} = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} w_{vjk} G_{vjk}. \qquad (4)$$

For dichotomously scored items, it is readily checked that $w_{vjk} = 1$. Index G_+ also plays an important role in person-fit analysis (Meijer & Sijtsma, 2001).

Relationships between outlier scores and test score. Test score X_+ is defined as the sum of the J item scores, such that $X_+ = \sum_{j=1}^{J} X_j$. Some relationships between the outlier scores O_+ and G_+, and test score X_+ are the following (but notice that many other possibilities exist depending on the properties of the test and the resulting data).

One example is a questionnaire that measures a relatively rare phenomenon, such as a particular pathology. As a result, the distribution of X_+ is skewed to the right. Respondents that have relatively high X_+ scores are expected to have high O_+ scores because they have many high item scores which are rare among the majority of the group. Thus, in such questionnaires we expect a strong positive linear relationship between X_+ and O_+ and suggest that observations in the right-tail of the X_+ distribution may be outliers. Another example is that the distribution of X_+ on a relatively easy educational test may be skewed to the left. As a result, the X_+ and O_+ are expected to have a strong negative linear relationship which suggests possible outliers in the left tail. Obviously, the thinner the tail and the more distant observations in the tail are from the central tendency of the distribution, the more likely they are outliers.

Respondents having low or high X_+ scores cannot have many Guttman errors; thus, their G_+ scores are low by definition and an inverse U-shaped relationship is expected between X_+ and G_+.

Finally, notice that a strong positive linear relationship between O_+ and G_+ suggests that they quantify similar outlier concepts even though their definitions are different.

Identifying Suspected Observations and Testing for Discordancy

Respondents with a surprisingly high outlier score, O_+ or G_+, or both, are considered suspected. *Tukey's fences* (Tukey, 1977, pp. 43–44), also known as the *boxplot* method (e.g., Vanderviere & Hubert, 2004), may be used to identify suspected observations as follows. The interquartile range (*IQR*) is the difference between the 75th percentile (Q_3) and the 25th percentile (Q_1) of the outlier score. The (inner) fences are at $Q_3 + 1\frac{1}{2} \times IQR$ and $Q_1 - 1\frac{1}{2} \times IQR$ (e.g., see the boxplots in the first column of Figure 1). For the proposed outlier scores, observations smaller than $Q_1 - 1\frac{1}{2} \times IQR$ are not suspected and, as a consequence, they are not considered any further. Observations greater than $Q_3 + 1\frac{1}{2} \times IQR$ are considered to be *suspected observations*.

In what follows, L denotes the number of suspected observations in the sample (e.g., based on Tukey's fences or another heuristic) and K the number

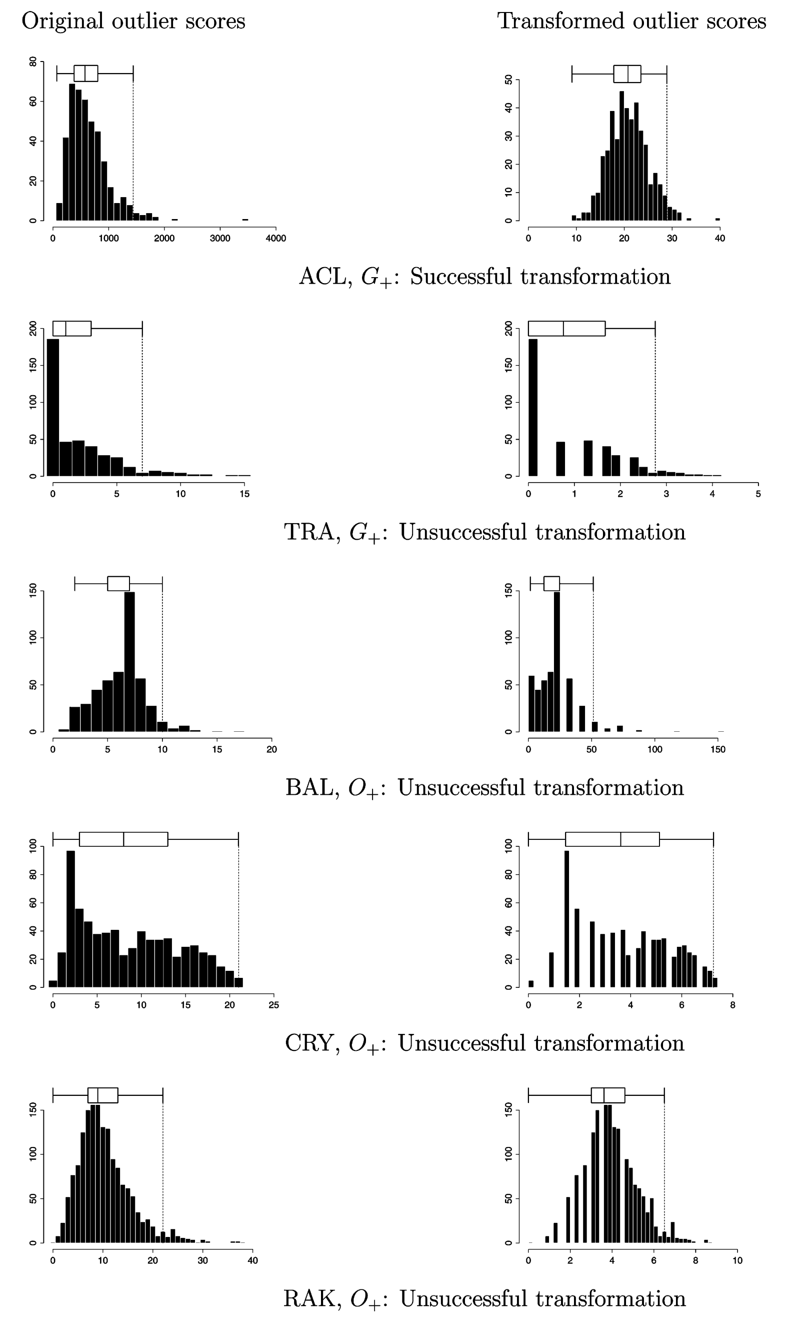


FIGURE 1 Examples of Box-Cox power transformations of outlier scores for data sets ACL, TRA, BAL, CRY, and RAK. *Note*. For the transformed outlier scores the boxplots with Tukey's fences are based on the transformation of the non-transformed boxplots and Tukey's fences, and not on the transformed outlier scores.

of observations judged to be outliers (e.g., based on a formal statistical test). We use two methods to judge whether observations are outliers. First, we adopt Tukey's fences as an informal test; all L scores greater than $Q_3 + 1\frac{1}{2} \times IQR$ are considered outliers (this implies that $K = L$). Second, we use Tukey's fences as a heuristic device to identify suspected observations and use a formal test—called a discordancy test—to decide which suspected observations are outliers (note that this implies that $K \leq L$).

As a formal discordancy test the generalized *extreme studentized deviate* (ESD) procedure is used (e.g., Barnett & Lewis, 1994, pp. 221–222; Iglewicz & Hoaglin, 1993, pp. 32–33; Rosner, 1983). The generalized ESD procedure tests the null hypothesis that the scores have a normal distribution with mean μ and variance σ^2 against the alternative that the scores are contaminated by scores from a normal distribution with mean $\mu + a$ ($a > 0$) and variance σ^2. Let the generic notation U denote an outlier score with realization u, sample mean $\overline{U}$ and sample standard deviation S_U. The ESD is defined as

$$ESD_v = \frac{\max |U_v - \overline{U}|}{S_U}. \tag{5}$$

Rosner (1983; also see Barnett & Lewis, 1994, p. 221) approximated the significance probability (SP) of the test by

$$SP(ESD_v) \leq N \times P\left(t_{N-2} > \sqrt{\frac{N(N-2)ESD_v^2}{(N-1)^2 - N \times ESD_v^2}}\right), \tag{6}$$

where N is the number of observations and $P(t_{N-2} > c)$ is the probability that an observation from a Student's t distribution with $N - 2$ degrees of freedom exceeds c. Among the abundance of discordancy tests for univariate samples (Barnett & Lewis, 1994, chap. 6), the ESD procedure is the most powerful test when the remainder of the scores is normally distributed and the number of genuine outliers does not exceed the number of suspected observations (Iglewicz & Hoaglin, 1993, pp. 38–41; Jain, 1981). This means that for the ESD procedure to be powerful, the number of suspected observations that is tested has to be at least as large as the number of genuine outliers. Also, the ESD procedure has the advantage that the p-value can be approximated well using Equation 6. Equation 6 includes a minor practical adjustment proposed by Simonoff (1984), which is that the significance probability is calculated as if only one suspected observation is tested for discordancy.

When multiple outliers are present in the sample, problems of *masking* and *swamping* may occur (e.g., Barnett & Lewis, 1994, pp. 109–110; Iglewicz & Hoaglin, 1993, p. 30). Masking occurs when a small cluster of outliers attracts the mean $\overline{U}$ and inflates the standard deviation S_U (Hadi, 1992); this results in the presence of one or more less extreme outliers masking the presence of the more extreme outliers. As a result, neither the more extreme nor the less extreme outliers may be identified. Swamping happens when a cluster of observations are tested simultaneously (called block-testing), some of which are non-outlying scores and others outlying scores, and the whole block is found to be significant; then the non-outlying observations are labelled discordant due to the presence of one or more outliers (Hadi, 1992). To minimize masking and swamping, *outward consecutive testing* is advocated (Barnett & Lewis, 1994, p. 131; Simonoff, 1984), meaning that the suspected observation that deviates the least is tested first. If this observation is judged to be discordant, all observations that are more extreme are also judged to be discordant. If this observation is not judged to be discordant, the next suspected observation is tested for discordancy, and this is repeated until a suspected observation is judged to be discordant or the suspected observation that deviates the most is found not to be discordant. When a particular outlier score is observed multiple times, only one of these observations (called the pivot observation) is tested. If the pivot observation is judged to be discordant, all observations that are equal or greater than the pivot observation are judged to be discordant. If the pivot observation is not judged to be discordant, none of the same observations are discordant, and the next extreme suspected observation is tested.

A suspected observation is judged to be discordant if $p < .05$ (Equation 6). The significance probabilities are based on the assumption that the outlier scores follow a normal distribution, and may be incorrect if this assumption is not satisfied. Hence, observations may be incorrectly declared to be outliers due to the non-normality of the population (Tietjen & Moore, 1972). In general, the distribution of the outlier scores is unknown and depends on the test or the questionnaire that produced the data. In our data examples discussed shortly, we found that the observed outlier score distributions were often skewed to the right and sometimes bounded by zero.

In order to render the p-values resulting from the ESD procedure (based on Equation 6) more trustworthy, outlier scores are transformed to an approximately normal distribution using the *Box-Cox power transformation* (Box & Cox, 1964; Iglewicz & Hoaglin, 1993, pp. 50-53). The Box-Cox power transformation changes the relative distances between the scores and is especially useful for skewed distributions with a relatively large range (Hoaglin, Mosteller, & Tukey, 1983, chap. 4). Let λ be a parameter defining a particular transformation, and $Y(\lambda)$ the transformed outlier score, then for $U > 0$ the Box-Cox power transformation is defined as

$$Y(\lambda) = \begin{cases} \dfrac{U^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(U) & \text{if } \lambda = 0. \end{cases} \tag{7}$$

The following points may be noted with respect to the application of the Box-Cox power transformation in this study:

1. In general, parameter λ is chosen such that the distribution of $Y(\lambda)$ approximates a normal distribution as closely as possible. In this study, λ was chosen such that it maximized the correlation between the proportions of the transformed outlier scores and the ordinates of the transformed outlier scores when they are normal (NIST/SEMATECH, 2006).
2. The estimates for λ were found by computing this correlation for $\lambda = -1.00, -0.99, -0.98, \ldots, 2.50$ and choosing the λ value that produced the highest correlation. More accurate estimates of λ do not necessarily improve the Box-Cox power transformation (cf., Box & Cox, 1964).
3. In this study, suspected observations were disregarded for the estimation of λ because the ESD procedure assumes that such observations come from a different distribution than the non-suspected observations.
4. If an outlier score had a value of zero the Box-Cox power transformation could not be applied (Equation 7); therefore a constant was added to all observations so that all outlier scores were positive (i.e., $U' = U + 1$).
5. The Shapiro-Wilk test (Shapiro & Wilk, 1965) was used to test whether the transformed data without the suspected observations followed a normal distribution (using a significance level of $\alpha = .05$). The Shapiro-Wilk test is an omnibus test known to have excellent power when testing for normality (e.g., Henderson, 2006, pp. 124–125).

Investigating the Influence of Outliers

Leaving out observations from a data set will likely change the outcome of the statistical analysis. When omission of outliers results in a larger change than omission of an equal number of random observations, the outliers are considered to be influential observations. Given that K cases were judged to be outliers, a useful research strategy may be to compare the effect of omitting the K outliers with omitting K randomly selected cases on the same statistical analysis. When the omission of K randomly selected cases is repeated a great number of times, each time omitting K randomly selected cases that are replaced after the computations in that round have been completed, confidence intervals for the outcome of the statistical analysis may be constructed. It may then be checked whether the result based on the data without the outlying cases lies outside this interval. If it does, the outliers were influential with respect to this particular statistical analysis. To determine whether outliers were influential, a distribution of the statistic of interest, generically denoted S, can be determined as follows:

1. Compute S after the K outliers have been deleted from the sample. The resulting statistic is denoted $S^*_{(K)}$.
2. Compute S after K different observations have been deleted at random from the sample. Repeat this 1000 times, and denote the resulting statistics by $S_{(K)b}$ ($b = 1, \ldots, 1000$; b indexes repetitions). The 1000 values of $S_{(K)b}$ were used to determine the 2.5th and the 97.5th percentile of the sampling distribution.
3. Under the null hypothesis that the influence of the K outliers is equal to the influence of K randomly selected cases, $S^*_{(K)}$ is expected to lie within the 2.5th and the 97.5th percentile boundaries of the distribution. If $S^*_{(K)}$ lies outside these boundaries, the null hypothesis is rejected, and the outliers are considered to be influential.

INVESTIGATION OF OUTLYING OBSERVATIONS IN REAL-DATA SETS

Method

First, the outlier scores O_+ and G_+ and the methods for identifying outliers, Tukey's fences and the ESD procedure, were used for inspecting ten real-data sets (Table 2) with respect to the presence of outliers. The data sets were chosen from studies in which the authors had been involved. The data sets were collected with tests and questionnaires that differed with respect to the attributes measured, the number of items and the sample size, and the number of answer categories.

Second, we investigated whether a statistical analysis of the complete data leads to other results than a similar analysis of the data excluding the identified outliers. If the results are different, the omitted cases are considered influential. For example, the statistic of interest may be Cronbach's (1951) alpha coefficient, which is a well known lower bound to the reliability of test score X_+. The question is whether another value of alpha is found in the complete data than in the data without the identified outliers.

Four well known statistics (including Cronbach's alpha) that are often used as quality indices for test scores and individual items were used for determining the possible influence of deleting the identified outliers. They were:

- *Cronbach's alpha*. Let $Cov(X_j, X_k)$ denote the sample covariance between the scores on items j and k, and let $S^2_{X_+}$ denote the sample variance of total score X_+; then

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

TABLE 2
Data Sets Used for Outlier Identification and Accommodation; Attribute Measured, Sample Size, Test Length, Number of Answers Categories, and Reference

Data Set	*Attribute*	*N*	*J*	$m+1$	*Reference*
1 VER	Verbal intelligence by means of verbal analogies	990	32	2	Meijer, Sijtsma, & Smid (1990)
2 BAL	Intelligence by balance scale problem-solving	484	25	2	Van Maanen, Been, & Sijtsma (1989)
3 CRY	Tendency to cry	705	23	2	Vingerhoets & Cornelius (2001)
4 IND	Inductive reasoning	478	43	2	De Koning, Sijtsma, & Hamers (2003)
5 RAK	Word comprehension	1641	60	2	Bleichrodt, Drenth, Zaal, & Resing (1985)
6 TRA	Transitive reasoning	425	10	2	Verweij, Sijtsma, & Koops (1999)
7 COP	Strategies for coping with industrial malodor	828	7	4	Cavalini (1992)
8 ACL	Personality traits	433	52	5	Gough & Heilbrun (1980)
9 WIL	Willingness to participate in labor union action	496	6	5	Van der Veen (1992)
10 SEN	Sensation seeking tendency	441	13	7	Van den Berg (1992)

- *Item-rest correlation*. The correlation between the score on item j and the total score on the other $J-1$ items, defined as $R_{(-j)} = X_+ - X_j$, is often used as an index for the degree to which item j is a measure of the same construct as the other $J-1$ items. In SPSS (2005) output, the item-rest correlation is called corrected item-total correlation.
- *Loevinger's/Mokken's H*. Loevinger's (1948; also, see Mokken, 1971) scalability coefficient H may be interpreted as an index for the accuracy of a person ordering with respect to X_+. It is used in the context of ordinal measurement (Sijtsma & Molenaar, 2002, chap. 4). Let $Cov_{max}(X_j, X_k)$ denote the maximum covariance of the scores on the items j and k given the marginal distributions of the cross-table of X_j and X_k; then

$$H = \frac{\sum\sum_{j<k} Cov(X_j, X_k)}{\sum\sum_{j<k} Cov_{max}(X_j, X_k)}.$$

- *Item scalability coefficient* H_j. The item scalability coefficient H_j gives the scalability of item j with respect to the other $J - 1$ items, and is defined as

$$H_j = \frac{\sum_{k \neq j} Cov(X_j, X_k)}{\sum_{k \neq j} Cov_{max}(X_j, X_k)}.$$

The higher H_j, the more item j contributes to an accurate person ordering as expressed by the overall H.

Results

Association between outlier scores and test score. Figure 2 shows three examples of the association between the test score X_+ (abscissa) and the outlier scores (ordinate), O_+ (first column) and G_+ (second column). The regression curve was obtained using the LOESS fitting method (e.g., Chambers & Hastie, 1992). The association between O_+ and X_+ was approximately linear for data sets CRY ($r = .98$; Figure 2d), TRA ($r = -.81$), COP ($r = .79$), and SEN ($r = -.61$). For data sets VER (Figure 2a), BAL, IND, RAK, ACL (Figure 2g), and WIL the association can be best characterized as a U-shape. Irrespective of the form of the association, for all data sets the item-based outliers were found in the tails of the X_+ distribution.

For the data sets VER (Figure 2b), CRY (Figure 2e), RAK, COP, WIL, and SEN the association between X_+ and G_+ can be best characterized by an inverse U-shape. The mean and the variance of G_+ were larger in the middle of the range of X_+ scores and smaller when the X_+ scores were low or high. Data sets BAL, IND and ACL (Figure 2h) showed only part of the inverse U-shape association, because only part of the X_+ range was observed. In general, item-pair based outliers were found in the middle of the X_+ distribution. An exception was data set TRA, which showed an approximate linear association ($r = -.60$), with the item-pair based outliers found in the lower tail of the X_+ distribution.

Figure 2 (third column) shows three examples of the association between the item-based outlier score O_+ (abscissa) and item-pair based outlier score G_+ (ordinate). The associations were all positive, and appeared in three ways. First, data sets BAL, IND, and TRA showed approximately linear relationships characterized by correlations of .77, .72, and .71, respectively. Second, data set CRY (Figure 2f) showed an inverse U-shape association, which was the same as the association between X_+ and G_+ because $r(X_+, O_+) = .98$. Third, data

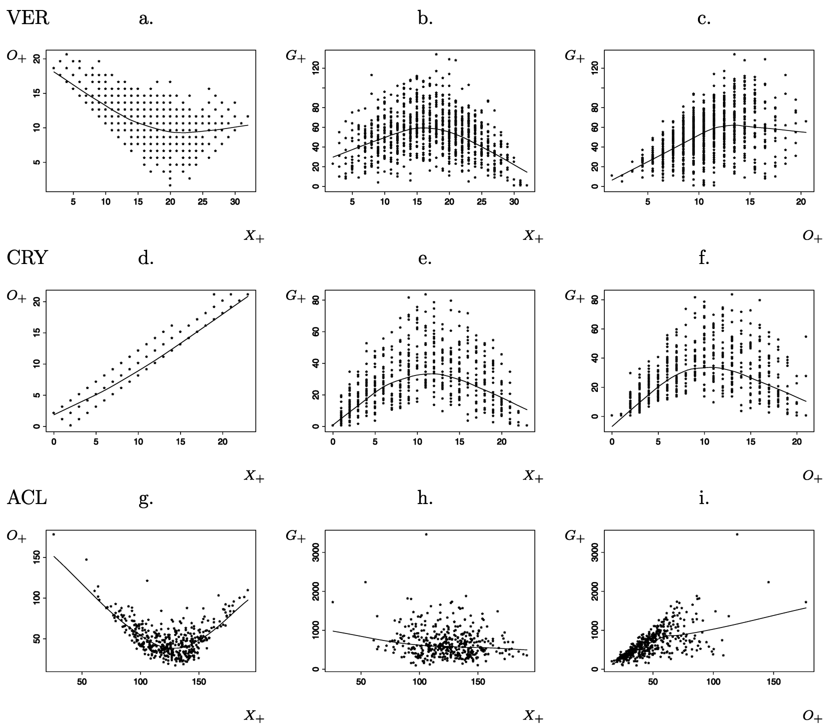


FIGURE 2 Examples of scatter plots (with smoothed association curves using LOESS fitting method) among the two outlier scores (O_+, G_+) and test scores (X_+) for data sets VER, CRY, and ACL. *Note.* First column: association between X_+ (abscissa) and O_+ (ordinate); Second column: association between X_+ (abscissa) and G_+ (ordinate); Third column: association between O_+ (abscissa) and G_+ (ordinate).

sets VER (Figure 2c), RAK, COP, ACL (Figure 2i), WIL, and SEN showed heteroscedastic associations, which can be described as follows. Larger O_+ values were associated with a wide range of G_+ values, and smaller O_+ values were associated with small G_+ values, but smaller G_+ values were associated with a wide range of O_+ values. This suggests that the two outlier scores quantify different concepts and may be used complementary.

Outlier detection. For each data set, Table 3 shows the number (L) and the percentage ($L\%$) of suspected observations identified by Tukey's fences, the number of outliers (K) identified by the ESD procedure, and details of the Box-Cox power transformation using the item-based outlier score O_+ and the item-pair based outlier score G_+.

TABLE 3
Suspected Observations, Outliers, and Information on the Box-Cox Power Transformation for Ten Data Sets

Data Set	*Outlier Score*	*L*	*L%*	*K*	λ	*S-W p-value*	*Skewness Before*	*Skewness After*	*Comments*
VER	O_+	8	0.8%	0	0.59	< .001	0.19	−0.17	B(18)
	G_+	6	0.6%	0	0.93	.033	0.17	0.08	
BAL	O_+	15	3.1%	11	2.02	< .001	−0.46	0.30	B(10), C(7)
	G_+	28	5.8%	0	0.52	< .001	0.57	−0.36	C(15)
CRY	O_+	0	0%	0	0.49	< .001	0.38	−0.00	B(22), C(2), D
	G_+	2	0.3%	0	0.74	< .001	0.59	0.16	C(0)
IND	O_+	8	1.7%	1	0.86	.003	0.18	0.08	B(19)
	G_+	1	0.2%	1	0.75	.162	0.27	−0.00	A
RAK	O_+	58	3.5%	0	0.42	< .001	0.51	−0.10	B(23)
	G_+	71	4.3%	0	0.44	< .001	0.78	−0.06	
TRA	O_+	37	8.7%	6	0.72	< .001	0.12	−0.07	B(4)
	G_+	29	6.8%	0	0.26	< .001	0.99	−0.51	B(8), C(0)
COP	O_+	9	1.1%	0	0.54	< .001	0.43	−0.08	B(14)
	G_+	42	5.1%	0	0.49	< .001	0.91	0.16	B(28), C(0)
ACL	O_+	10	2.3%	0	0.06	.023	0.63	−0.07	
	G_+	15	3.5%	1	0.32	.137	0.69	−0.09	A
WIL	O_+	13	2.6%	0	0.39	< .001	0.53	−0.22	B(17)
	G_+	34	6.9%	0	0.41	< .001	0.86	−0.00	B(27), C(0)
SEN	O_+	2	0.5%	0	0.62	.037	0.18	−0.09	
	G_+	10	2.3%	0	0.55	.058	0.67	0.03	A

Note. L: number of suspected observations identified by Tukey's fences; $L\%$: percentage of suspected observations; K: number of outliers identified by the ESD procedure; λ: Box-Cox power transformation coefficient; S-W p-value: the p-value of the Shapiro-Wilk test. If $p \geq .05$ the Box-Cox power transformation to a normal distribution was considered successful; Before: the skewness of the outlier score without the suspected observations before the Box-Cox power transformation; After: the skewness of the outlier score without the suspected observations after the Box-Cox power transformation. Comments: A = Box-Cox power transformation successful; B(R) = Box-Cox power transformation unsuccessful due to short range of outlier scores, where R is the number of different values of the $N - L$ outlier scores; C(u) = Box-Cox power transformation unsuccessful due to dominant outlier score u; D = Box-Cox power transformation unsuccessful due to platykurtic distribution of the outlier scores.

The percentage of suspected observations (based on Tukey's fences) ranged from 0% (CRY, O_+) to 8.75% (TRA, O_+). This percentage is positively related to the number of outlier scores with value equal to zero (not tabulated). The O_+ scores and G_+ scores generally indicated different observations as suspected. Only for data set BAL, 13 of the 15 suspected observations according to O_+ were also suspected according to G_+; and for data set TRA, 17 of the 37 suspected observations according to O_+ were also suspected according to G_+. This was expected because these data sets had strong positive correlations between O_+ and G_+ ($r = .77$ and $r = .71$, respectively).

The distributions of the outlier scores were skewed to the right except for data set CRY (O_+; almost uniform), data set BAL (O_+; symmetric and leptokurtic), and data set VER (G_+; normal). Except for data sets ACL and SEN in which the O_+ scores were also non-integer valued, in the other data sets the outlier scores were nonnegative integers. Non-integer scores may occur when the ranks of item categories are tied (for an example, see Table 1, item 4). In general, applying the Box-Cox power transformation to the outlier scores without suspected observations decreased the skewness of the distribution (Table 3). The λ value used in the Box-Cox power transformation (Table 3) ranged from $\lambda = 0.06$ (ACL, O_+; almost a logistic transformation) to $\lambda = 2.02$ (BAL, O_+; quadratic transformation). Most λ values were close to $\frac{1}{2}$ or $\frac{1}{3}$, which corresponds to taking the square root or the cubic root of the outlier scores, respectively. For outlier score G_+ of data set VER, λ was close to 1 (i.e., $\lambda = 0.93$), which indicates that no transformation was needed.

Seventeen out of 20 Box-Cox power transformations resulted in a rejection of the hypothesis that the transformed data follow a normal distribution (based on the Shapiro-Wilk test with $\alpha = .05$). Figure 1 (top row) shows an example of a successful Box-Cox power transformation (i.e., G_+ for data set ACL). When the Box-Cox power transformation failed to produce a normal distribution, this could be attributed to one of the following reasons (or a combination of these reasons) (Table 3, last column):

1. *Short range of outlier scores.* The Box-Cox power transformation of outlier scores is unlikely to be useful when the range of the outlier scores is small (Hoaglin et al., 1983, pp. 124–125). For outlier scores O_+ and G_+ a short range means that few different values of the outlier scores were observed. Data sets containing few items and/or items having few answer categories cause the range of the outlier scores to be short. Removing the suspected observations from the data reduced the range even more. This was an important cause for failure of the Box-Cox power transformation of O_+ scores in data sets VER, BAL, CRY, IND, RAK, TRA, COP, and WIL, and of G_+ scores in data sets TRA, COP, and WIL. For example, for data set TRA outlier score O_+ had only four different values (Table 3). Figure 1 (second, third, fourth, and fifth row) shows the Box-Cox power transformation of distributions with a short scale range.
2. *Dominant outlier score value*. An outlier score value is dominant when it is observed more often than other outlier score values or more often than expected. The O_+ scores of data sets BAL and CRY, and the G_+ scores of data sets BAL, CRY, TRA, COP, and WIL had one dominant value which caused the Box-Cox power transformation to be unsuccessful. Changing

the relative distances between the scores did not affect the dominance of a particular value. Figure 1 (second, third, and fourth row) shows the Box-Cox power transformation of a distribution with dominant value $G_+ = 0$ (data set TRA) and a distribution with a dominant O_+ value in the middle of scale (data set BAL) and at the left of the scale (data set CRY).

3. *Platykurtic distribution*. The distribution of the O_+ scores of data set CRY was almost uniform (kurtosis = 1.9) (Figure 1, fourth row). Transformation of a uniform distribution cannot result in a normal one.

Alternatively, none of the explanations above applied to failure of the Box-Cox power transformation of O_+ in data sets ACL and SEN or to the transformation of G_+ in data sets VER and RAK (Figure 1, fifth row). The transformed distributions of O_+ in data sets VER, IND, RAK, COP, ACL, WIL, and SEN, and of G_+ in data sets VER and RAK were found to be non-normal (Shapiro-Wilk test) but appeared bell-shaped. The number of outliers K was determined regardless of the Shapiro-Wilk test results, and ranged from $K = 0$ (14 times) to $K = 11$ (Table 3, fifth column).

Influence of outliers. Table 4 shows the separate effects of deleting L outliers identified by means of Tukey's fences and K outliers identified by means of the ESD procedure on the following statistics: Cronbach's alpha, the item-rest correlation of item j, coefficient H, and coefficient H_j. Item j is the item out of J items in the test or questionnaire which has its H_j value closest to .3; this is an important lower bound for selecting items (Sijtsma & Molenaar, 2002, pp. 60–61). Notation "−−" denotes a significant decrease and "++" denotes a significant increase of the statistic of interest.

In general, deleting outliers based on O_+ resulted in a decrease of the statistics, whereas deleting outliers based on G_+ resulted in an increase. The explanation for the first result is that almost all outliers identified by O_+ were in the tails of the X_+ distribution, and that their removal resulted in a truncated distribution of X_+. This caused the statistics to have lower values. The explanation for the second result is that the statistics are based on covariances, which increase when the data contain fewer Guttman errors (Sijtsma & Molenaar, 2002, pp. 55–58). This produced lower covariances and thus lower values of the statistics.

For data set TRA the effects of removing the L outliers based on O_+ were strongest. The decrease of the values of all statistics was large after omission of the L outliers. This effect could be explained as follows. All O_+ values greater than 3 were identified as outliers using Tukey's fences, and given the strong negative linear correlation between O_+ and X_+ ($r = -.81$), this implied that only cases having either one of the four highest test scores (X_+ = 7, 8, 9, and 10) were included in the data. This was a homogeneous group and, as a

TABLE 4
Values of Four Statistics from Psychometrics, and the Influence on These Statistics of Omitting L or K Outliers From Ten Real-Data Sets on the Basis of Outlier Scores O_+ and G_+

Data Set	*Outlier*	O_+ *alpha*	O_+ *IRC(j)*	O_+ H	O_+ H_j	*Outlier*	G_+ *alpha*	G_+ *IRC(j)*	G_+ H	G_+ H_j
VER		.8594	.2132	.2457	.3014		.8594	.2132	.2457	.3014
	$L = 8$	−−	−−	−−	−−	$L = 6$	++	++	++	++
	$K = 0$					$K = 0$				
BAL		.5621	.6393	.0993	.3126		.5621	.6393	.0993	.3126
	$L = 15$	−−	+	−	++	$L = 28$	++	++	++	++
	$K = 11$	−−	+	−	++	$K = 1$	−	++	−	++
CRY		.9237	.5097	.4476	.3866		.9237	.5097	.4476	.3866
	$L = 0$					$L = 2$	++	−	++	+
	$K = 0$					$K = 0$				
IND		.8456	.5391	.1898	.3004		.8456	.5391	.1898	.3004
	$L = 8$	−−	++	−−	−	$L = 1$	0	++	+	++
	$K = 1$	−−	++	−−	++	$K = 1$	0	++	+	++
RAK		.9464	.4274	.5798	.4254		.9464	.4274	.5798	.4254
	$L = 58$	−−	−−	−−	−−	$L = 71$	+	++	++	++
	$K = 0$					$K = 0$				
TRA		.5162	.3740	.2048	.2929		.5162	.3740	.2048	.2929
	$L = 37$	−−	−−	−−	−−	$L = 29$	−−	−−	−	−
	$K = 6$	−−	−−	−−	−−	$K = 0$				
COP		.7120	.4164	.3123	.3069		.7120	.4164	.3123	.3069
	$L = 9$	−−	−−	−−	−−	$L = 42$	++	++	++	++
	$K = 0$					$K = 0$				
ACL		.9497	.5104	.3021	.3002		.9497	.5104	.3021	.3002
	$L = 10$	−−	−−	−−	−−	$L = 15$	+	−	+	+
	$K = 0$					$K = 1$	++	+	++	+
WIL		.7444	.4377	.3584	.3420		.7444	.4377	.3584	.3420
	$L = 13$	−−	+	−−	−	$L = 34$	++	++	++	++
	$K = 0$					$K = 0$				
SEN		.8584	.4575	.3465	.2996		.8584	.4575	.3465	.2996
	$L = 2$	−	−	−	−	$L = 10$	++	+	++	+
	$K = 0$					$K = 0$				

Note. j is the item which has H_j value closest to .3; "−−": omission of outliers leads to significantly lower values than random omission; "−": omission of outliers leads to lower values than random omission but not significant; "++": omission of outliers leads to significantly larger values than random omission; "+": omission of outliers leads to larger values than random omission but not significant; "0": omission of outliers does not lead to difference; IRC(j): item-rest correlation of item j.

result, the correlational structure in the data was lost. Thus, O_+ should not be used as an outlier score for data set TRA.

DISCUSSION

Outlier identification and accommodation is a neglected topic in the analysis of test and questionnaire data collected in psychology, education, sociology, political science, and other fields. In this study, two scores were used to assess the degree to which an observation is inconsistent with the remainder of the data. The first score was the item-based outlier score O_+, which quantifies the number of times a subject has item scores in the less frequently observed answer categories. The second was the item-pair based outlier score G_+, which counts the number of Guttman errors.

Two methods were used to identify inconsistent observations as outliers. The first method was Tukey's fences procedure and the second was the extreme studentized deviate (ESD) procedure. The ESD procedure assumes normality of the distribution of outlier scores. For most data sets, the distributions of O_+ and G_+ were highly skewed to the right. A Box-Cox power transformation to achieve normality was successful in three cases, but failed in 17 cases. Unsuccessful transformation of O_+ to normality (for all data sets) was mostly caused by the short scale range of the outlier scores (8 times). However, in most cases when the transformation of O_+ appeared to be unsuccessful the transformed data looked approximately normal. Unsuccessful transformation of G_+ to normality (7 times) was mostly caused by a dominant outlier score (5 times). Four out of five times the dominant value was zero. In these cases, transforming the data to normality is nearly impossible.

A respondent who has (nearly all) J item scores either equal to 0 or m, has a G_+ value equal to or close to 0, which will not show up as an outlier when G_+ is used. This property of G_+ should be taken into consideration when G_+ is used. Also, an item that does not measure the attribute well can cause many errors, and thus may influence the distribution of G_+. On the other hand, all respondents are influenced by this "bad" item, and this may prevent outliers from appearing.

Tukey's fences procedure identified 0.3% to 8.7% of the observations as outliers. The only exception was data set CRY, in which no outliers were identified by means of O_+. The ESD procedure identified outliers in four out of ten data sets but none in the other six data sets. When the Box-Cox power transformation was unsuccessful, the quality of the ESD procedure could not be guaranteed (i.e., we do not know whether the ESD procedure is robust to non-normality). When the Box-Cox power transformation is successful the ESD procedure can be considered. However, the transformation could cause extreme observations to be not extreme anymore when λ is small, and vice versa, cause normal observations to be extreme when λ is large. Also, some criticism has been exercised on using Tukey's fences for detecting outliers when the distribution is extremely skewed. Because Tukey's fences are based on measures of location and scale of a distribution, but not on measures of skewness, Tukey's fences may identify too many outliers when the data are skewed (Vanderviere & Hubert, 2004). Alternatively, Vanderviere and Hubert proposed the use of an adjusted boxplot. Since in our study real-data sets were used, it is unknown how many outliers were present, let alone if any outliers were present at all. Simulation studies should be performed to answer the question how well outliers are detected by the outlier scores and the testing methods defined here.

Removal of outliers detected with item-based O_+ outlier scores resulted in a decrease of the value of these statistics and removal of outliers detected with item-pair based G_+ outlier scores resulted in an increase of the value of the statistics. In most cases, the detected outliers were influential on the statistics from psychometrics. This is taken as an indication that detection of outliers was successful. Removing outliers should lead to values of statistics closer to the population value. Thus, an outlier score such as G_+ which tends to increase Cronbach's alpha and other statistics is not automatically a good method unless, after removal of outliers, it produces closer approximations to the population value. The two outlier scores have different effects on statistics from psychometrics, they have different relationships with the test score, and for most real-data sets they have a weak relationship with each other. This suggests that they quantify different concepts and may be used complementary.

Identified outliers may contain valuable information and should be investigated carefully. If a reasonable theoretical explanation is available for an observation to be an outlier and if it may be concluded that the observation is not representative for the population under study, it may be deleted from the analysis. However, if such an explanation is absent, one should consider the possibility that the model is wrong. To overcome the influence of outliers if deleting them is not an option, a proper procedure is to accommodate the outliers by using robust estimation procedures, or transforming the data.

Future research may concentrate on other outlier scores. One may think of identification of item-score patterns typical of response styles, such as the tendency to primarily give neutral, extreme, or affirmative answers to rating-scale items. Usually, item-score vectors based on one of these mechanisms give evidence of not responding according to instruction. Their presence calls for closer inspection of the statistical results.

Another topic for future research is accommodation of categorical influential data. The results presented here are only a first step in this direction, but more definitive results may be obtained from a systematic investigation using simulated data. Such data could contain outliers simulated according to definitions on which outlier indices are based, and the power of such indices for identifying these cases may be investigated. Also, some more insight could be gained into the way in which relevant outcome variables are influenced by outliers. It is a hopeful sign that the analysis of ten real-data sets already gave some indications of the usefulness of two outlier indices proposed, and also suggested a methodology for identifying influential cases and how to accommodate them.

The third topic for future research is the choice of meaningful outcome variables. Here, we have chosen some well known and much used statistics in psychometric data analysis, but modern test and questionnaire analysis would be served well by including outcome statistics such as estimated latent person and item parameters from item response models, their standard errors, test information functions, and diagnostic model tests, both for testing models under the null hypothesis and for model selection (such as the AIC and the BIC). Together, we believe that the suggestions made here set up a complete research program. This study is a modest albeit useful start of this program.

REFERENCES

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New York: Springer.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data.* New York: Wiley.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1985). *Revisie Amsterdamse Kinderintelligentie Test (RAKIT)* [Revision of the Amsterdam Child Intelligence Test.] Lisse, The Netherlands: Swets & Zeitlinger.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B, 26*, 211–252.

Bradlow, E. T., & Weiss, R. E. (2001). Outlier measures and norming methods for computerized adaptive tests. *Journal of Educational and Behavioral Statistics, 26*, 83–102.

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

Cavalini, P. M. (1992). *It's an ill wind that brings no good: Studies on odour annoyance and the dispersion of odour concentrations from industries*. Unpublished Ph.D. thesis, University of Groningen, The Netherlands.

Chambers, J. M., & Hastie, T. J. (1992). Statistical Models in S. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical Models* (pp. 309–376). Pacific Grove, CA: Wadsworth & Brooks/Cole.

Chambers, R., Hentges, A., & Zhoa, X. (2004). Robust automatic methods for outlier and error detection. *Journal of the Royal Statistical Society, Series A, 167*, 323–339.

Cook, R. D., & Critchley, F. (2000). Identifying regression outliers in mixtures graphically. *Journal of the American Statistical Association, 95*, 781–794.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

De Koning, E., Sijtsma, K., & Hamers, J. H. M. (2003). Construction and validation of a test for inductive reasoning. *European Journal for Psychological Assessment, 19*, 24–39.

Gough, H. G., & Heilbrun Jr., A. B. (1980). *The Adjective Check List, manual 1980 edition.* Palo Alto, CA: Consulting Psychologists Press.

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.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series A, 54*, 761–771.

Hadi, A. S., & Simonoff, J. S. (1993). Procedures of identification of multiple outliers in linear models. *Journal of the American Statistical Association, 88*, 1264–1272.

Henderson, A. R. (2006). Testing experimental data for univariate normality. *Clinica Chimica Acta, 366*, 112–129.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1983). *Understanding robust and exploratory data analysis.* New York: Wiley.

Hodge, V. J., & Austin, J. (2004). A survey of outlier methodologies. *Artificial Intelligence Review, 22*, 85–126.

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers.* Milwaukee, WI: ASQC Quality Press.

Jain, R. B. (1981). Detecting outliers: Power and some other considerations. *Communications in Statistics, Theory and Methods, 10,* 2299–2314.

Kotze, T. J. W., & Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Applied Statistics, 33*, 215–223.

Lee, A. H., & Yick, J. S. (1999). A perturbation approach to outlier detection in two-way contingency tables. *Australian and New Zealand Journal of Statistics, 41*, 305–314.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin, 45*, 507–530.

Meijer, R. R. (2002). Outlier detection in high-stakes certification testing. *Journal of Educational Measurement, 39*, 219–233.

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

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14,* 283–298.

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

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12*(37), 97–117.

NIST/SEMATECH (2006). *e-Handbook of Statistical Methods*. Retrieved September 21, 2006, from http://www.itl.nist.gov/div898/handbook/eda/section3/eda336.htm.

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.* New York: Wiley.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52*, 591–611.

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

Simonoff, J. S. (1984). A comparison of robust methods and detection of outliers techniques when estimating a location parameter. *Communications in Statistics, Theory and Methods, 13*, 813–842.

Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency tables via backward-stepping. *Technometrics, 30*, 339–345.

SPSS (2005). *SPSS base 14.0 user's guide*. Chicago: SPSS.

Tietjen, G. L., & Moore, R. H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics, 14*, 583–597.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van den Berg, P. T. (1992). *Persoonlijkheid en werkbeleving: De validiteit van persoonlijkheidsvragenlijsten, in het bijzonder die van een spanningsbehoeftelijst* [Personality and work experience: The validity of personality questionnaires, and in particular the validity of a sensation-seeking questionnaire.] Unpublished Ph.D. thesis, Vrije Universiteit, Amsterdam.

Van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory* (pp. 1–28). New York: Springer.

Van der Veen, G. (1992). *Principes in praktijk. CNV-leden over collectieve acties* [Principles in practice: Labor union members on means of political pressure.] Kampen, The Netherlands: Kok.

Vanderviere, E. & Hubert, M. (2004). An adjusted boxplot for skewed distributions. In J. Antoch (Ed.), *COMPSTAT2004 Symposium: proceedings in computational statistics* (pp. 1933–1940). Heidelberg, Germany: Physica-Verlag.

Van Maanen, L., Been, P. H., & Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In E. E. Ch.I. Roskam (Ed.), *Mathematical psychology in progress* (pp. 267–287). Berlin, Germany: Springer.

Verweij, A. C., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23,* 241–264.

Vingerhoets, A. J. J. M., & Cornelius, R. R. (Eds.) (2001). *Adult crying: A biopsychosocial approach.* Hove, UK: Brunner-Routledge.

Yick, J. S., & Lee, A. H. (1998). Unmasking outliers in two-way contingency tables. *Computational Statistics and Data Analysis, 29*, 69–79.