Outlier detection in the medical Questionnaire Rising and Sitting Down (QR&S)
Zijlstra, W.P.; van der Ark, L.A.; Sijtsma, K.

Published in:
New Trends in Psychometrics

Publication date:
2008

Link to publication

Citation for published version (APA):
Outlier Detection in the Medical Questionnaire Rising and Sitting Down (QR&S)

Wobbe P. Zijlstra,1 L. Andries van der Ark, and Klaas Sijtsma
Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands
(1) w.p.zijlstra@uvt.nl

Abstract

Outlier detection in item scores from questionnaires for the measurement of medical concepts has to deal with highly discrete data. In this study, two outlier scores are used which both indicate the degree of inconsistency of a subject’s item-score vector with the remainder of the data. In two studies, simulated data are used to investigate the error rates and the sensitivity of four statistical tests that are used to decide whether an outlier score is discordant. In the third study, the outlier scores and the discordancy tests are applied to real data obtained by means of the medical Questionnaire Rising and Sitting Down (QR&S)*.

1. Introduction

Identification of outliers is an important step in data analysis. Outliers can be thought of as observations that are inconsistent with the remainder of the data (Barnett and Lewis, 1994, p. 7). Note that this description is rather vague; therefore, we use more precise terms that replace the term outlier. It is assumed that observations in the sample stem either from the population of interest – then they are called regular observations – or from another population – in which case they are called contaminant observations. Observations that are unusual, extreme, or surprising are called suspected observations. A formal discordancy test is used to decide whether the suspected observations should be considered contaminant observations or regular observations. Observations that are tested positively are called discordant observations.

Many questionnaires in medical and health research contain variables (called items) are dichotomously or polytomously scored. Let $X_j$ denote the random variable for the ordered integer score on item $j$ ($j = 1, \ldots, J$), and let $x_j$ be a realization of $X_j$. Items are scored $x_j = 0, \ldots, m$; for dichotomous items $m = 1$ and for polytomous items $m \geq 2$. Usually, $m$ does not exceed 4. Based on so few answer categories, suspected observations cannot be identified by investigating one single item. A viable alternative is to investigate the item-score vectors based on all $J$ items.

Recently Zijlstra, Van der Ark, and Sijtsma (2007) proposed two simple statistics, called outlier scores, which are assigned to each individual’s item score vector, and which can be used to identify suspected observations. These outlier scores reflect the degree of inconsistency with the remainder of the item-score vectors. The first was the item-based outlier score, $O_+$, which is defined as the frequency of unpopular item scores in an individual’s vector of $J$ item scores. The second was the item-pair based outlier score, $G_+$, which is defined as the number of weighted Guttman errors (Molenaar, 1991). The outlier-score distributions were inspected for discordant observations by means of two discordancy tests, Tukey’s

*Acknowledgement: The authors are grateful to Leo D. Roorda for making available the data from the QR&S.
Table 1 Possible Outcomes of a Discordancy Test With the Number of Contaminants ($N_C$) and the Number of Regular Observations ($N_R$).

<table>
<thead>
<tr>
<th>True situation</th>
<th>Discordancy test result</th>
<th>Contaminant</th>
<th>Regular</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discordan</td>
<td>valid positive</td>
<td>false negative</td>
</tr>
<tr>
<td></td>
<td>Not discordan</td>
<td>false positive</td>
<td>valid negative</td>
</tr>
</tbody>
</table>

(1977) fences (also known as the boxplot) and Rosner's (1983) generalized extreme studentized deviate (ESD) procedure after normality transformation of the outlier-score distribution (denoted ESD-T). Also, the influence of the discordant observations on several statistics was investigated. Tukey's fences identified between 0% and 8.7% discordant observations, but the ESD-T hardly identified any discordant observations at all. Because it was suspected that the ESD procedure has lower Type I error rate than Tukey's fences and the transformation to normality lowers the Type I error rate, and because both factors cause fewer discordant outlier scores to be detected, the present study considers the Type I error rate and the sensitivity of discordancy tests.

In this study, four discordancy tests were applied to the outlier-score distributions of $O_+$ and $G_+$. The four discordancy tests were: (1) Tukey's fences; (2) the adjusted boxplot, which is the Tukey's fences with an adjustment for skewness; (3) the ESD; and (4) the ESD-T. A discordancy test classifies an observation as being discordant (positive) or not discordant (negative); this classification is correct (valid) or incorrect (false). Table 1 shows the four possibilities. A valid positive is a contaminant that is identified as discordant and a valid negative is a regular observation that is not identified as discordant. A misclassification is either a false positive or a false negative.

The performance of a discordancy test can be evaluated by means of two quantities. The sensitivity is the probability of identifying valid positives, and the specificity is the probability of identifying valid negatives. The sensitivity is computed by dividing the number of valid positives by the number of contaminants ($N_C$) (Table 1). It is the power of a discordancy test. The specificity is computed by dividing the number of valid negatives by the number of regulars ($N_R$). In this study, the error rate is reported, which is $(1 - \text{specificity})$, and which is computed by dividing the number of false positives by the number of regulars ($N_R$).

Two null hypotheses are relevant for discordancy testing. The first null hypothesis (i.e., $H_{10}$) is that an observation belongs to the population of regular observations. The Type I error associated with $H_{10}$ is the error rate and is denoted by $\alpha_N$. Thus, $\alpha_N = \text{error rate} = (1 - \text{specificity})$. The second null hypothesis ($H_{20}$) is that all $N$ observations in the sample are regular. The Type I error associated with $H_{20}$ is the some-outside rate (Hoaglin, Iglewicz, and Tukey, 1986) and is denoted by $\alpha$. The some-outside rate is the probability of finding at least one false positives in the sample. Under $H_{20}$, the probability that the discordancy test identifies a sample without false positives is $1 - \alpha$. For a sample of size $N$ drawn from a normal distribution, $\alpha$ and $\alpha_N$ are related by $\alpha_N = 1 - (1 - \alpha)^{1/N}$ (Davies and Gather, 1993).

The performance of the four discordancy tests applied to $O_+$ and $G_+$ was investigated in two simulation studies and one real-data study. The first simulation study investigated the tests' error rates ($\alpha_N$) and some-outside rates ($\alpha$) in samples of only regular subjects. The second simulation study also investigated the tests' sensitivity in contaminated samples. In the third study, the discordancy tests were applied to real data.
Table 2  Examples of Item Category Proportions \([P(X_j = x)]\) of Five Dichotomous Items, the Item-Based Outlier Score \((O_j)\) for Each Answer Category, and the \(O_{vj}\) Scores for Item-Score Vector \(x_v = (1, 1, 0, 1, 0)\). The Last Column Shows \(O_{v+}\).

<table>
<thead>
<tr>
<th></th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>(O_{v+})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>(P(X_j = x))</td>
<td>.1</td>
<td>.9</td>
<td>.25</td>
<td>.75</td>
<td>.4</td>
<td>.6</td>
</tr>
<tr>
<td>(O_j)</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(O_{vj})</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

2. Definitions of Outlier Scores and Discordancy Tests

2.1. Outlier scores

Because this study used dichotomous item scores, the \(O_+\) and \(G_+\) outlier scores are explained only for dichotomous items.

*Item-based outlier score.* Outlier score \(O_+\) rests on the idea that responses in the modal (most popular) score categories of items are not suspected, responses in the next, less popular score category are a little suspected, and so on; and that responses in the least popular score category are the most suspected. The score distribution of item \(j\) is denoted by \([P(X_j = 0), P(X_j = 1)]\). Outlier item-score, \(O_j\), equals 0 for the modal category, and 1 for the least popular category. For \(P(X_j = 0) = P(X_j = 1)\), we define \(O_j = .5\). For respondent \(v\), item-based outlier score \(O_{v+}\) is defined as

\[
O_{v+} = \sum_{j=1}^{J} O_{vj},
\]

with \(0 \leq O_{v+} \leq J\). Table 2 shows the frequency distributions for five dichotomous items. Let \(X_v = (X_{v1, \ldots, X_{v5}})\) and let \(x_v\) contain the 5 item scores of respondent \(v\). For items 1, 2, and 3, \(X_j = 1\) is modal and for items 4 and 5, \(X_j = 0\) is modal. Respondent \(v\), with \(x_v = (1, 1, 0, 1, 0)\), has scores in the most popular category for items 1, 2, and 5 and in the least popular category for items 3 and 4. Thus, for this respondent \(O_{v+} = 2\) (Table 2).

*Item-pair based outlier score.* The item-pair based outlier score, \(G_+\), uses weighted Guttman errors (Molenaar, 1991). Assume that the \(J\) items in a test are ordered according to decreasing popularity and then numbered accordingly. This is the common item ordering (e.g., Table 2). For two items, \(j\) and \(k\), with \(j < k\), this implies \(P(X_j = 1) \geq P(X_k = 1)\). Based on the common item ordering, item-pair scores can represent either Guttman errors or conformal patterns. Given that \(P(X_j = 1) > P(X_k = 1)\), a Guttman error occurs when \(X_{vj} = 0\) and \(X_{vk} = 1\), denoted \((x_{vj}, x_{vk}) = (0, 1)\), and a conformal pattern when \((x_{vj}, x_{vk})\) equals either \((1, 0), (1, 1),\) or \((0, 0)\). A Guttman error results in a score \(G_{vjk} = 1\) and a conformal pattern in \(G_{vjk} = 0\). For respondent \(v\), the item-pair based outlier score \(G_{v+}\) is defined as

\[
G_{v+} = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} G_{vjk},
\]

with \(0 \leq G_+ \leq (J^2 - 1)/4\) if \(J\) is odd, and \(0 \leq G_{v+} \leq J^2/4\) if \(J\) is even. For respondent \(v\) (Table 2), only item pair \((3, 4)\) is a Guttman error and, as a result, \(G_{v+} = 1\). See Molenaar (1991) for the case of polytomous items.
2.2. Discordancy tests

Tukey's fences. Tukey's fences (Tukey, 1977, pp. 43-44), also known as the boxplot method, identifies suspected observations as follows. Let $Q_1$ denote the 25th percentile, $Q_3$ the 75th percentile, and $IQR$ (the interquartile range) the difference between $Q_3$ and $Q_1$; then, the upper fence is located at $Q_3 + 1.5 \times IQR$. The upper fence is used as critical value; that is, outlier scores larger than the upper fence are regarded as suspected. When Tukey's fences is used as a discordancy test, all suspected observations are discordant. Tukey's fences is concerned with $H_1$ and uses a fixed error rate $\alpha_N$. For a standard normal distribution, the (one-sided) error rate for Tukey's upper fence corresponds with $\alpha_N = .0035$ (Hoaglin, Iglewicz, and Tukey, 1986).

Adjusted boxplot. Vanderviere and Hubert (2004) proposed an adjusted boxplot that takes the possible right skewness of the outlier-score distribution (Zijlstra, Van der Ark, and Sijtsma, 2007) into account to control the error rate. The robust skewness measure $medcouple$ (MC) is used for this purpose. Medcouple is defined as follows. Let the generic notation $U$ denote an outlier score with sample median $med(U)$. For all outlier-score pairs $(U_v, U_w)$ from the sample, with $U_v < med(U)$ and $U_w > med(U)$, $MC$ is the median of a kernel function and is defined as

$$MC = med \left( \frac{|U_w - med(U)| - |med(U) - U_v|}{U_w - U_v} \right), \text{ for } U_v < med(U) < U_w. \quad (3)$$

Vanderviere and Hubert (2004) defined the upper fence of the adjusted boxplot as $Q_3 + (1.5 \times IQR) \times A$, with $A = e^{(3.87 \times MC)}$. For example, for a right-skewed distribution with $MC = .2$, the upper fence of the adjusted boxplot is $A = e^{(3.87 \times .2)} = 2.17$ times further above $Q_3$ than Tukey's fences. For $MC = 0$, the adjusted boxplot equals Tukey's fences.

ESD. The ESD procedure (e.g., Barnett and Lewis, 1994, pp. 221–222; Rosner, 1983) tests the null hypothesis that regular scores have a normal distribution $N(\mu, \sigma^2)$ against the alternative that the distribution is contaminated by scores from a normal distribution $N(\mu + \Delta, \sigma^2)$, with slippage parameter $\Delta > 0$. For sample mean $\bar{U}$ and sample standard deviation $S_U$, ESD is defined as

$$ESD = \max_{U_v} |U_v - \bar{U}| / S_U. \quad (4)$$

Testing multiple suspected observations was done by means of outward consecutive testing (Barnett and Lewis, 1994, p. 131; Simonoff, 1984). In this study, we chose the highest $(N - 1)/2 U$ values as the suspected observations (based on Simonoff, 1984). The least deviating suspected observation is tested first. The critical value is determined as if there was only one observation tested for discordancy (see, Simonoff, 1984). If it is tested discordant, all other suspected observations are also labelled discordant, and testing is stopped. If this observation is not tested discordant, the next suspected observation is tested, and so on. When several suspected observations have the same outlier-score value, only one test is performed. During testing, $\bar{U}$, $S_U$, and $ESD$ are computed anew in each step, using the unsuspected observations, the suspected observations that appeared not discordant, and the observation to be tested for discordancy. The ESD is concerned with $H_2$, that is, a nominal some-outside rate is assumed, which in this study equals $\alpha = .05$. ESD-T. To fix the possible right skewness of the outlier-score distribution (Zijlstra, Van der Ark, and Sijtsma, 2007), this distribution may be transformed to a normal distribution using the Box-Cox power transformation (Box and Cox, 1964).
Table 3 Descriptives of the Two Patient Groups (Amputation and Osteoarthritis); Sample Size (N), Percentage Female (%fem), Mean Age, and Mean Test Score (X+). Standard Deviations Are Given in Parentheses.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>%fem</th>
<th>Age (SD)</th>
<th>X+ (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM</td>
<td>230</td>
<td>29.6</td>
<td>58.0 (16.5)</td>
<td>16.9 (11.3)</td>
</tr>
<tr>
<td>OA</td>
<td>295</td>
<td>69.2</td>
<td>69.3 (10.1)</td>
<td>22.8 (10.9)</td>
</tr>
</tbody>
</table>

This transformation yields the ESD-T. Let \( \lambda \) be a parameter defining a particular transformation, and \( Y_v(\lambda) \) the transformed outlier score, then for \( U_v > 0 \) the Box-Cox power transformation is defined as

\[
Y_v(\lambda) = \frac{U_v^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0, \text{ and } Y_v(\lambda) = \ln(U_v) \text{ if } \lambda = 0.
\]  

(5)

The estimation of \( \lambda \) is based on the observations smaller than the upper fence of Tukey’s fences (Zijlstra, Van der Ark, and Sijtsma, 2007). Like the ESD, the ESD-T assumes a nominal some-outside rate of \( \alpha = .05 \).

3. Method

The medical survey Questionnaire Rising and Sitting Down (QR&S; Roorda, Molenaar, Lankhorst, and Bouter, 2005) was used in all three studies. It uses 39 dichotomous items to measure activity limitations in rising and sitting down. Each item consists of three parts: (1) an activity limitation with respect to rising or sitting down concerning (2) a particular aspect of this limitation (velocity, difficulty, use of arm[rest]s or other adaptations), which happens with (3) a specific object (high or low chair, toilet, bed, or car seat). Patients indicated whether (score 1) or not (score 0) an item applied to the them. Sitting down was easier than rising, the order for rising from easiest to most difficult was from high chair, toilet, bed, low chair, to car seat, and the order for sitting down from easiest to most difficult was from high chair, bed, toilet, low chair, to car seat (Roorda, Molenaar, Lankhorst, and Bouter, 2005).

The data used in this study are based on two patient groups. The descriptives are given in Table 3. The test score, \( X_+ \), is defined as the sum of the \( J \) item scores. The first group were patients with an amputation (AM) to the legs (\( N = 230 \)) and consisted of 68 women (29.6%) and 162 men. The average age was 58.0 years; women and men had the same mean age [Welch’s \( t(127.6) = 1.12, p > .25 \)]. The second group were patients with osteoarthritis (OA) to the hip or the knees (\( N = 295 \)) and consisted of 204 women (69.2%) and 91 men. The average age was 69.3 years, and women were on average almost 5 years older than men [Welch’s \( t(138.8) = 3.56, p < .001 \)]. The average test score for the AM group (\( X_+ = 16.9 \)) was much lower than for the OA group (\( X_+ = 22.8 \)) [Welch’s \( t(483.6) = 6.98, p < .001 \)], which means that the OA group had more limitations in rising and sitting down than the AM group.

3.1. Study 1

Study 1 investigated the error rates \( (\alpha_N) \) and the some-outside rates \( (\alpha) \) of the four discordancy tests applied to \( O_+ \) and \( G_+ \) in simulated samples of regular observations only. Regular observations were generated as follows. Let \( \theta \) denote the latent trait with \( \theta \sim N(0, 1) \), and \( b_j \) and \( a_j \) the location and discrimination parameters of item \( j \), respectively. The two-parameter logistic model (2PLM) uses the latent trait and these item parameters to describe item scores, and is defined as
\[ P(X_j = 1|\theta) = \frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]}. \] (6)

The Rasch model is a special case in which \( a_j = 1 \) for all items. The software package MULTILOG (Thissen, Chen, and Bock, 2003) was used to compute a deviance test, which showed that the 2PLM fitted significantly better to the QR&S data than the Rasch model: for AM, \( \chi^2(38) = 132.2, p < .0001 \); and for OA, \( \chi^2(38) = 172.2, p < .0001 \). Next, for both patient groups, MULTILOG was used to estimate the items' \( b \) and \( a \) parameters (mean and standard deviation in Table 3). Regular item-score vectors were generated by means of Equation 6, in which the estimated parameters had been inserted. Outlier scores \( O_+ \) and \( G_+ \) were computed for the resulting regular item-score vectors. For the AM group, the simulations were based on \( N = 230 \) and for the OA group the simulations were based on \( N = 295 \) (same sample size as in the real data). For each group, 10,000 samples were drawn, and the error rates and the some-outside rates of the four discordancy tests were computed.

3.2. Study 2

Study 2 investigated the sensitivity of the four discordancy tests applied to \( O_+ \) and \( G_+ \) in contaminated samples. Contaminant observations were generated as follows. Let \( F \) be the empirical distribution of the regular outlier scores, generated as in Study 1, with standard deviation \( S_F \). Furthermore, let \( u \) be the realization of outlier score \( U \) and let \( H \) be the distribution of the contaminant outlier scores. The contaminant outlier scores were generated from \( H(u) = F(u + \Delta) \), with \( \Delta = 4 \times S_F \). The distribution \( F \) and \( H \) have an identical shape but differ by a location slippage equal to \( 4 \times S_F \), which was large enough for contaminants to show up as extreme outlier scores.

A sample consisted of \( N_R \) regular outlier scores from distribution \( F \) and \( N_C \) contaminant outlier scores from distribution \( H \) (\( N = N_R + N_C \)). Values \( N_C = 5, 10, \) and 25 were considered. Given sample sizes \( N = 230 \) (AM group) and \( N = 295 \) (OA group), contamination was 2.17%, 4.34%, and 10.87% (AM group), and 1.69%, 3.39%, and 8.47% (OA group). For each of the six cells, 10,000 samples were drawn. The four discordancy tests were applied to the contaminated samples, and the error rate, the some-outside rate, and the sensitivity were computed.

3.3. Study 3

The four discordancy tests applied to the two outlier scores were used to analyze the real QR&S data of the two patient groups. The distribution of the outlier scores, the number of discordant subjects identified by the discordancy tests, and the item-score patterns of the discordant subjects were investigated.

4. Results

4.1. Study 1

For Tukey's fences and the adjusted boxplot, the nominal error rate was \( \alpha_N = .0035 \), and for the ESD and the ESD-T the nominal some-outside rate was \( \alpha = .05 \). Tukey's fences identified too many regular observations as discordant (Table 4; \( \alpha_N = .0146 \) for \( O_+ \) in the OA group), whereas the adjusted boxplot controlled the error rate to a great extent (i.e., the observed error rates were close to the nominal rate). For \( G_+ \), on average the adjusted boxplot produced larger error rates than Tukey's fences. This indicates that for some distributions the medcouple was negative when positive values were expected. Furthermore, for \( G_+ \) both Tukey's fences and the adjusted boxplot identified at least one false positive in approximately 50% of the samples (see \( \alpha \) in Table 4). The ESD and the ESD-T controlled the
some-outside rate by adjusting the error rate. For $O_+$, this adjustment resulted in no or almost no false positives (Table 4; $\alpha_N < .0001$ and $\alpha < .0010$), rendering the ESD and the ESD-T too conservative. For $G_+$, the ESD and the ESD-T had some-outside rates much lower than Tukey’s fences and the adjusted boxplot, with the ESD-T closer to the nominal $\alpha = .05$.

4.2. Study 2

The error rate and the some-outside rate were lower for contaminated data, and decreased with increasing contamination (not tabulated). Tukey’s fences had the highest sensitivity and the ESD-T the lowest sensitivity; and the adjusted boxplot and the ESD had a sensitivity in between (Table 5). Furthermore, the sensitivity decreased as the number of contaminants ($N_C$) increased. This decrease was smaller for Tukey’s fences than for the other three discordancy tests. Thus, Tukey’s fences may be more robust to contamination with respect to the sensitivity. The ESD and the ESD-T had low sensitivity when $N_C = 25$.

Except for the adjusted boxplot and the ESD-T for $O_+$ in the OA group, only minor differences in sensitivity of the discordancy tests were found between $O_+$ and $G_+$ and between the AM and OA groups (Table 5). The adjusted boxplot and the ESD-T had much lower sensitivity for $O_+$ in the OA group, perhaps because the distribution of $O_+$ was more skewed in the OA group than the AM group (Figure 1). As a result, the adjustment for skewness was larger for the OA group causing the critical value to increase and the sensitivity to decrease.

4.3. Study 3

For $J = 39$, $0 \leq O_+ \leq 39$ and $0 \leq G_+ \leq 380$. The distribution of $O_+$ appeared more skewed to the right for OA patients than for AM patients (Figure 1). For both groups, almost no outlier scores $O_+ < 5$ were found. The distribution of $G_+$ appeared skewed to the right. It had relatively many observations with $G_+ = 0$ of which many corresponded with $X_+ = 0$ or $X_+ = 39$ (AM group: 24 subjects; and OA group: 18 subjects). For these item-score vectors, no Guttman errors can be observed. It may be noted that the observed shapes of the distributions of $O_+$ and $G_+$ only pertain to this particular case; for other groups, or when other questionnaires are used, the shapes may be different.

For both the AM group and the OA group, the ESD and the ESD-T applied to $O_+$ and $G_+$ identified no discordant observations (Table 6). For the AM group, for $O_+$ Tukey’s fences and the adjusted boxplot obtained the same upper fence because the medcouple was zero. Both tests identified one (by definition the same)
discordant observation by means of $O_+$. For the AM-group data, only Tukey's fences applied to $G_+$ identified discordant observations (six in total including the one also identified by means of $O_+$). The subject identified discordant by means of both $O_+$ and $G_+$ scores had many activity limitations ($X_+ = 28$) but none with low chairs (eight items). Since for most people low chairs were most problematic, this subject can be considered a contaminant.

For the OA group, Tukey's fences identified eleven discordant observations based on $O_+$ and three based on $G_+$, and the adjusted boxplot identified two discordant observations based on $G_+$ (Table 6). None of the discordant observations were identified by both $O_+$ and $G_+$, thus in total $11 + 3 = 14$ different discordant observations were identified. In the OA group, nine of the eleven discordant subjects identified using $O_+$ reported practically no limitations (seven subjects had test scores of $X_+ = 0$ and two subjects had $X_+ = 1$).

In general, for both patient groups the discordant subjects identified using outlier score $G_+$ had item-score vectors that were not consistent with the common item ordering. For some discordant observations, rising was less problematic than sitting down, which is unusual, and others had few activity limitations with car seats and low seats and more activity limitations with high chairs and toilets.

A reviewer suggested to compare our outlier scores with leverage scores (Hoaglin and Welsch, 1978). Following the rule of thumb given by Hoaglin and Welsch (1978), we found for AM three observations to be discordant and for OA four observations. For both groups, one observation was also found to be discordant using $G_+$.

### Table 5
Sensitivity of the Four Discordancy Tests for the Two Outlier Scores ($O_+$ and $G_+$) for the Two Patient Groups (AM and OA) When the Number of Contaminants is $N_C = 5, 10,$ and $25.$

<table>
<thead>
<tr>
<th></th>
<th>$O_+$</th>
<th></th>
<th>$G_+$</th>
<th></th>
<th>$O_+$</th>
<th></th>
<th>$G_+$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>10</td>
<td>25</td>
<td>5</td>
<td>10</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>AM</td>
<td>Tukey</td>
<td>.82</td>
<td>.79</td>
<td>.61</td>
<td>.80</td>
<td>.77</td>
<td>.62</td>
</tr>
<tr>
<td></td>
<td>AdjBox</td>
<td>.73</td>
<td>.65</td>
<td>.32</td>
<td>.66</td>
<td>.55</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>ESD</td>
<td>.61</td>
<td>.54</td>
<td>.09</td>
<td>.59</td>
<td>.51</td>
<td>.07</td>
</tr>
<tr>
<td></td>
<td>ESD-T</td>
<td>.51</td>
<td>.42</td>
<td>.06</td>
<td>.40</td>
<td>.26</td>
<td>.01</td>
</tr>
<tr>
<td>OA</td>
<td>Tukey</td>
<td>.93</td>
<td>.89</td>
<td>.66</td>
<td>.82</td>
<td>.79</td>
<td>.69</td>
</tr>
<tr>
<td></td>
<td>AdjBox</td>
<td>.27</td>
<td>.19</td>
<td>.06</td>
<td>.73</td>
<td>.65</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>ESD</td>
<td>.53</td>
<td>.44</td>
<td>.17</td>
<td>.59</td>
<td>.53</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>ESD-T</td>
<td>.06</td>
<td>.04</td>
<td>.01</td>
<td>.45</td>
<td>.34</td>
<td>.03</td>
</tr>
</tbody>
</table>

### Table 6
The Number of Subjects Identified as Discordant by the Four Discordancy Tests When Applied to Real QR&S Data.

<table>
<thead>
<tr>
<th></th>
<th>Amputation</th>
<th>Osteoarthritis</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_+$</td>
<td>$G_+$</td>
<td>$O_+$</td>
</tr>
<tr>
<td>Tukey</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>AdjBox</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ESD</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ESD-T</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
5. Conclusion and Discussion

Based on this study, it can be concluded that (1) Tukey's fences identified most discordant observations and had the highest probability of identifying contaminant observations as discordant (i.e., the highest sensitivity); (2) The ESD and the ESD-T gave the highest certainty that a discordant observation is indeed a contaminant, (i.e., few regular observations are identified as discordant); (3) The sensitivity of Tukey's fences is fairly robust to the number of contaminants, whereas the ESD and the ESD-T showed a dramatic decrease in the sensitivity when the number of contaminants was large; (4) Adjusting the distributions of \( O_+ \) and \( G_+ \) for skewness by either using the adjusted boxplot or the ESD-T, may lead to more appropriate error rates or some-outside rates. However, the adjusted procedures may also have much lower sensitivity; and (5) The adjusted boxplot did not always function as expected because negative medcouple values were obtained where positive values were expected. This suggests that the medcouple may not be appropriate for distributions with limited numbers of integer values.

Outlier scores \( O_+ \) and \( G_+ \) quantify different concepts and, as a result, they may identify different observations as discordant. Thus, \( O_+ \) and \( G_+ \) may be used complementary. The authors recommend to use Tukey's fences, which was found to be a liberal discordancy test. Although the probability of identifying more false positives is higher for Tukey's fences than for the other discordancy tests, the probability of identifying inconsistent observations is also higher. Identifying suspected observations is the first step in outlier detection. The second step would be scrutinizing the suspected observation before continuing the analysis. Most importantly, discordant subjects may help to better understand the population under study. For example, the discordant subjects may have originated from a rare
or unknown (sub)population, which should be investigated explicitly in the future. Therefore, it may be desirable to identify at least some discordant observations. One may even argue to investigate the 5% or 10% largest outlier scores without the use of a discordancy test. However, the use of a discordancy test gives an important statistical indication to take the discordant observations serious.

**References**


