¹ How to classify, detect, and manage univariate and multivariate outliers, with emphasis on

² pre-registration

³ Christophe Leys[1], Marie Delacre[1], Youri L. Mora[1], Daniël Lakens[2], & Christophe Ley[3]

⁴ [1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

⁵ [2] Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven,

⁶ the Netherlands

⁷ [3] Universiteit Gent, Department of Applied Mathematics, Computer Science and Statistics,

⁸ Gent, Belgium

⁹ Author Note

¹³ Correspondence concerning this article should be addressed to Christophe Leys,

¹⁴ CP191, avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: christophe.leys@ulb.ac.be

15   Abstract

16   Researchers often lack knowledge about how to deal with outliers when analyzing their

17 data. Even more frequently, researchers do not pre-specify how they plan to manage

18 outliers. In this paper we aim to improve research practices by outlining what you need to

19 know about outliers. We start by providing a functional definition of outliers. We then lay

20 down an appropriate nomenclature/classification of outliers. This nomenclature is used to

21 understand what kinds of outliers can be encountered and serves as a guideline to make

22 appropriate decisions regarding the conservation, deletion, or recoding of outliers. These

23 decisions might impact the validity of statistical inferences as well as the reproducibility of

24 our experiments. To be able to make informed decisions about outliers you first need

25 proper detection tools. We remind readers why the most common outlier detection

26 methods are problematic and recommend the use of the Median Absolute Deviation to

27 detect univariate outliers, and of the Mahalanobis-MCD distance to detect multivariate

28 outliers. An R package was created that can be used to easily perform these detection

29 tests. Finally, we promote the use of pre-registration to avoid flexibility in data analysis

30 when handling outliers.

31   *Keywords:* outliers; preregistration; robust detection; malahanobis distance; median

32 absolute deviation; minimum covariance determinant

33   Word count:

How to classify, detect, and manage univariate and multivariate outliers, with emphasis on pre-registration

"...Most psychological and other social science researchers have not confronted the problem of what to do with outliers – but they should." (Abelson, 1995, p. 69). The past few years have seen an increasing concern about flexibility in data analysis (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). When confronted with a dataset, researchers have to make decisions about how they will analyze their data. This flexibility in the data analysis has come to be referred to as "researcher's degrees of freedom" (Simmons et al., 2011). Even before a statistical test is performed to examine a hypothesis, data needs to be checked for errors, anomalies, and test assumptions. This inevitably implies choices at many levels (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), including decisions about how to manage outliers (Leys, Klein, Dominicy, & Ley, 2018; Simmons et al., 2011). Different choices lead to different datasets, which could possibly lead to different analytic results (Steegen et al., 2016). When the choices about how to detect and manage outliers are based on the outcomes of the statistical analysis (i.e., when choices are based on whether or not tests yield a statistically significant result), the false positive rate can be inflated, which in turn might affect reproducibility. It is therefore important that researchers decide on how they will manage outliers before they collect the data and commit to this pre-specified plan.

Outliers are data points that are extremely distant from most of the other data points (see below for a more formal definition). Therefore, they usually exert a problematic influence on substantive interpretations of the relationship between variables. In two previous papers (Leys et al., 2018; Leys, Ley, Klein, Bernard, & Licata, 2013), the authors conducted two surveys of the psychological literature that revealed a serious lack of concern for (and even a clear mishandling of) outliers. Despite the importance of dealing adequately with outliers, practical guidelines that explain the best way to manage outliers

60 are scarce in the literature. The goal of this article is to fill this lack of an accessible

61 overview of best practices. We will discuss powerful new tools to detect outliers and discuss

62 the emerging practice to preregister analysis plans (Veer & Giner-Sorolla, 2016). Finally,

63 we will highlight how outliers can be of substantive interest, and how carefully examining

64 outliers may lead to novel theoretical insights that can generate hypotheses for future

65 studies. Therefore, this paper's aims are fourfold: (1) defining outliers; (2) discussing how

66 outliers could impact the data; (3) reminding what we consider the most appropriate way

67 to detect outliers and (4) proposing guidelines to manage outliers, with an emphasis on

68 pre-registration.

What is an Outlier?

70 Aguinis, Gottfredson, and Joo (2013) report results of a literature review of 46

71 methodological sources addressing the topic of outliers, as well as 232 organizational science

72 journal articles mentioning issues about outliers. They collected 14 definitions of outliers,

73 39 outliers detection techniques and 20 different ways to manage detected outliers. It is

74 clear from their work that merely defining an outlier is already quite a challenge. The 14

75 definitions differed in the sense that (a) in some definitions, outliers are all values that are

76 unusually far from the central tendency, whereas in other definitions, in addition to being

77 far from the central tendency, outliers also have to either disturb the results or yield some

78 valuable or unexpected insights; (b) in some definitions, outliers are not contingent on any

79 data analysis method whereas in other definitions, outliers are values that disturb the

80 results of a specific analysis method (e.g., cluster analysis, time series, or meta-analysis).

81 Two of these 14 definitions of outliers seemed especially well suited for practical

82 purposes. The first is attractive for its simplicity: *"Data values that are unusually large or*

83 *small compared to the other values of the same construct"* (Aguinis et al., 2013, Table 1,

84 p.275). However, this definition only applies to single constructs, but researchers should

85 also consider multivariate outliers (i.e., outliers because of a surprising pattern across

⁸⁶ several variables). Therefore, we will rely on a slightly more complicated but more

⁸⁷ encompassing definition of outliers: "Data points with large residual values". This

⁸⁸ definition calls for an understanding of the concept of "residual value", which is the

⁸⁹ discrepancy between the observed value and the value predicted by the statistical model.

⁹⁰ This definition does not call for any specific statistical method and does not restrict the

⁹¹ number of dimensions from which the outlier can depart.

⁹² **Error Outliers, Interesting Outliers, and Random Outliers**

⁹³   Aguinis et al. (2013) distinguish three types of mutually exclusive outliers: *error*

⁹⁴ outliers, *interesting* outliers and *influential* outliers. We will introduce two modifications to

⁹⁵ their nomenclature.

⁹⁶   The first modification concerns removing the category of *influential* outliers.

⁹⁷ Influential outliers are defined by Aguinis et al. (2013) as outliers that prominently

⁹⁸ influence either the fit of the model (model fit outliers) or the estimation of parameters

⁹⁹ (prediction outliers)[1]. In our view, according to this definition, all types of outliers could be

¹⁰⁰ influential or not (for additional extensive reviews, see Cohen, Cohen, West, & Aiken, 2003;

¹⁰¹ McClelland, 2000). Moreover, since the influential criterion will not impact how outliers are

¹⁰² managed, we will remove this category from our nomenclature. The second modification

¹⁰³ concerns the addition of a new category that we will name *random* outliers (see Table 1).

¹⁰⁴   *Error* outliers are non-legitimate observations that *"lie at a distance from other data*

¹⁰⁵ *points because they are results of inaccuracies"* (Aguinis et al., 2013, p. 282). This includes

¹⁰⁶ measurement errors and encoding errors. For example, a "77" value on a Likert scale

¹⁰⁷ ranging from 1 to 7 is an error outlier, caused by accidentally hitting the "7" twice while

¹⁰⁸ manually entering the data.

---

[1] The Model fit outliers appear for instance when using statistical methods based on the maximum
likelihood (and variants) method. Prediction outliers appear when using the more common least squares
method (such as in linear regression).

¹⁰⁹ *Interesting* outliers are not clearly errors but could be influenced by potentially

¹¹⁰ interesting moderators [2]. These moderators may or may not be of theoretical interest and

¹¹¹ could even remain unidentified. For this reason, it would be more adequate to speak of

¹¹² *potentially interesting* outliers. In a previous paper, Leys et al. (2018) highlight a situation

¹¹³ where outliers can be considered as heuristic tools, allowing researchers to gain insights

¹¹⁴ regarding the processes under examination (see McGuire, 1997): *"Consider a person who*

¹¹⁵ *would exhibit a very high level of in-group identification but a very low level of prejudice*

¹¹⁶ *towards a specific out-group. This would count as an outlier under the theory that group*

¹¹⁷ *identification leads to prejudice towards relevant out-groups. Detecting this person and*

¹¹⁸ *seeking to determine why this is the case may help uncover possible moderators of the*

¹¹⁹ *somewhat simplistic assumption that identification leads to prejudice"* (Leys et al., 2018, p.

¹²⁰ 151). For example, this individual might have inclusive representations of their in-group.

¹²¹ Examining outliers might inspire the hypothesis that one's social representation of the

¹²² values of the in-group may be an important mediator (or moderator) of the relationship

¹²³ between identification and prejudice.

¹²⁴ *Random* outliers are values that just randomly appear out of pure (un)luck. Imagine

¹²⁵ a perfectly well-balanced coin that yields 100 times heads on 100 throws. Random outliers

¹²⁶ are per definition very unlikely, but still possible.

¹²⁷ Table 1. *Adjusted nomenclature of outliers*

Error	*e.g., coding error*
Interesting	*e.g., moderator underlying a potentially interesting psychological process*
Random	*e.g., a very large value of a given distribution*

---

[2] Note that both error and interesting outliers are influenced by moderators. The moderator of the *error* outlier is identified as being of no theoretical interest and concerns an error (e.g. , coding error). The *interesting* outlier is driven by a moderator that is identified or not and that might potentially be of theoretical interest

Univariate and Multivariate Outliers

Another relevant distinction is the difference between univariate and multivariate outliers. Sultan Kösen is the tallest man currently alive (8ft, 2.8 in/251cm). Because he displays a particularly high value on a single dimension (his height) he can be considered a univariate outlier [3]

Now, let us imagine a cohort of human beings. An observation of a 5 ft 2 in (157 cm) tall person will not be surprising since it is quite a typical height. An observation of 64 lbs (29 kg) will not be surprising either, since many children have this weight. However, weighting 64 lbs *and* being 5 ft 2 in tall is surprising. This example is Lizzie Velasquez, born with a Marfanoid–progeroid–lipodystrophy syndrome that prevents her from gaining weight or accumulating body fat. Values that become surprising when several dimensions are taken into account are called *multivariate* outliers. Multivariate outliers are very important to detect, for example before performing structural equation modeling (SEM), where multivariate outliers can easily jeopardize fit indices (Kline, 2015).

An interesting way to emphasize the stakes of multivariate outliers is to describe the principle of a regression coefficient (i.e., the slope of the regression line) in a regression between to variable Y (set as *dependent variable*) and X (set as *independent variable*). Firstly, remember that the dot of coordinates $(\bar{X}, \bar{Y})$, named G-point (for Gravity-point; see the crossing of the two grey lines in Figure 1), necessarily belongs to the regression line. Next, the slope of this regression line can be computed by taking the sum of individual slopes of each line linking each data of the scatter dot and the G-point (see the arrows in Figure 1),multiplied by individual weight (ω_i).

---

[3] Although he obviously belongs to the human population, and as such is not an error outlier, it was worth detecting this departure from normality. Indeed, his unusual height is caused by an abnormal pituitary gland that never stopped secreting growth hormone. He stopped growing after a surgical treatment. This is a simple example of a univariate outlier that is not attributed to any inaccuracy but that is related to an
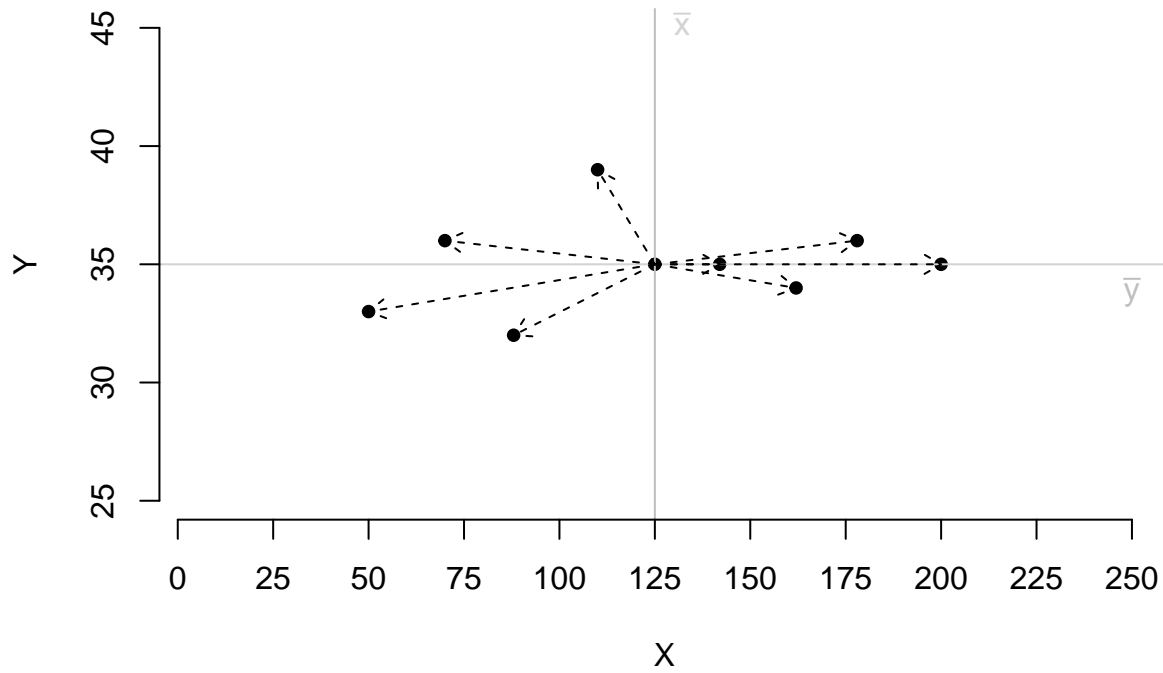
Figure 1. Illustration of individual slopes of lines linking all data of the scatter dot and the G-point

150   Individual slopes are computed as follows:

$$slope_i = \frac{Y_i - \bar{Y}}{X_i - \bar{X}} \tag{1}$$

151   Individual weights are computed by taking the distance between the X coordinate of
152 a given observation and $\bar{X}$ and dividing that distance by the sum of all distances:

$$\omega_i = \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2)} \tag{2}$$

153   As a consequence, the slope of the regression line can be computed as follows:

interesting moderator (the dysfunctional pituitary gland) that could account for the unusual observation.

$$b_i = \sum \omega_i \left(\frac{Y_i - \bar{Y}}{X_i - \bar{X}}\right) = \sum \frac{(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2}\left(\frac{Y_i - \bar{Y}}{X_i - \bar{X}}\right) \tag{3}$$

154   Given this equation, an individual having an extremely large or low coordinate on the

155 Y axis will unequally influence the regression slope depending on the distance between the

156 X_i coordinate of this individual and $\bar{X}$. As an illustration, Figure 2 shows 4 scatter dots.

157 In plot a, the coordinate of 3 points on the Y axis exactly equals $\bar{Y}$ (see points A, B and C

158 in plot a). In plots b, c and d, the coordinate of one of these 3 points is modified in order

159 that the point is moved away from $\bar{Y}$. If an observation is extremely high on the Y axis but

160 its coordinate on the X axis exactly equals $\bar{X}$ (i.e. $X_i = \bar{X}$), there is no consequence on the

161 slope of the regression line (because $\omega_i = 0$; see plot b). On the contrary, if an observation

162 is extremely high on both the Y axis and the X axis, the influence on the regression slope

163 can be tremendous and the further the coordinate on the X axis from $\bar{X}$, the higher the

164 impact (because ω_i increases; see plots c and d).

165   The detection of multivariate outliers relies on different methods than the detection

166 of univariate outliers. Univariate outliers have to be detected as values too far from a

167 robust central tendency indicator, while multivariate outliers have to be detected as values

168 too far from a robust ellipse (or a more complex multidimensional cloud when there are

169 more than two dimensions) that includes most observations (Cousineau & Chartier, 2010).

170 We will present recommended approaches for univariate and multivariate outlier detection

171 later in this article, but we will first discuss why checking outliers is important, how they

172 can be detected, and how they should be managed when detected.

Why Are Outliers Important?

174   An extreme value is either a legitimate or an illegitimate value of the distribution.

175 Let us come back on the perfectly well-balanced coin that yields 100 times "heads" in 100

176 throws. Deciding to discard such an observation from a planned analysis would be a
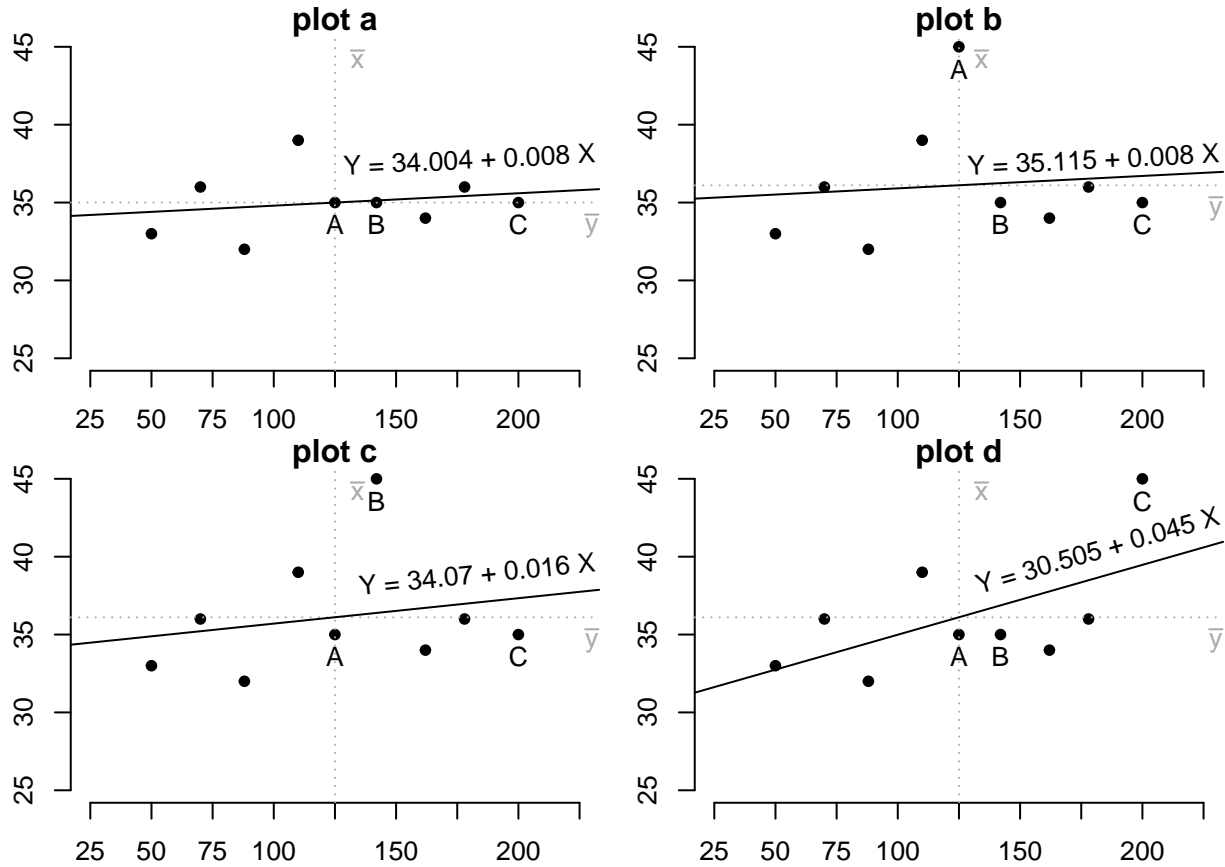
Figure 2. Impact of a an individual having extremely large or low coordinate on the Y axis, on the regression slope, as a function of its coordinate on the X axis

177   mistake in the sense that, if the coin is perfectly well-balanced, it is a legitimate

178   observation that has no reason to be altered. If, on the contrary, that coin is an allegedly

179   well-balanced coin but in reality a rigged coin with a zero probability of yielding "tails",

180   then keeping the data unaltered would be the incorrect way to deal with the outlier. In the

181   first scenario, altering (e.g., excluding) the observation implies inadequately reducing the

182   variance by removing a value that rightfully belongs to the considered distribution. On the

183   contrary, in the second scenario, keeping the data unaltered implies inadequately enlarging

184   the variance since the observation does not come from the distribution underpinning the

185   experiment. In both cases, a wrong decision may influence the Type I error (alpha error,

186   i.e., the probability that a hypothesis is rejected when it should not have been rejected) or

the Type II error (beta error, i.e., the probability that an incorrect hypothesis is not rejected) of the test. Making the correct decision will not bias the error rates of the test.

Unfortunately, more often than not, one has no way to know which distribution an observation is from, and hence there is no way to being certain whether any value is legitimate or not. Researchers are recommended to follow a two-step procedure to deal with outliers. First, they should aim to detect the possible candidates by using appropriate quantitative (mathematical) tools. As we will see, even the best mathematical tools have an unavoidable subjective component. Second, they should manage outliers, and decide whether to keep, remove, or recode these values, based on qualitative (non-mathematical) information. If the detection or the handling procedure is decided *post hoc* (after looking at the results), then researchers introduce bias in the results.

Detecting Outliers

In two previous papers, Leys et al. (2013) and Leys et al. (2018) reviewed the literature in the field of Psychology and showed that researchers primarily rely on two methods to detect outliers. For univariate outliers, psychologists consider values to be outliers whenever they are more extreme than the mean plus or minus the standard deviation multiplied by a constant, where this constant is usually 3, or 3.29 (Tabachnick & Fidell, 2013). These cut-offs are based on the fact that when the data are normally distributed, 99.7% of the observations fall within 3 standard deviations around the mean, and 99.9% fall within 3.29 standard deviations. In order to detect multivariate outliers, most psychologists compute the Mahalanobis distance (Mahalanobis, 1930; see also Leys et al., 2018 for a mathematical description of the Mahalanobis distance). Both these methods of detecting outliers rely on the mean and the standard deviation, which is not ideal because the mean and standard deviation themselves can be substantially influenced by the outliers they are meant to detect. Outliers pull the mean towards more extreme values (which is especially problematic when sample sizes are small), and because the mean is

213 further away from the majority of data points, the standard deviation increases as well.

214 This circularity in detecting outliers based on statistics that are themselves influenced by

215 outliers can be prevented by the use of robust indicators of outliers.

216   A useful concept when thinking about robust estimators is the *breakdown point*

217 (Donoho & Huber, 1983), defined as the proportion of values set to infinity (and thus

218 outlying) that can be part of the dataset without corrupting the estimator used to classify

219 outliers. For example, the median has a breakdown point of .5, which is the highest

220 possible breakdown point. A breakdown point of .5 means that the median allows 50% of

221 the observations to be set to infinity before the median breaks down. Consider, for the sake

222 of illustration, the following two vectors: X= {2, 3, 4, INF, INF, INF} and Z={2, 3, 4, 5,

223 INF, INF}. The vector X consists of 6 observations of which half are infinite. Its median,

224 computed by averaging 4 and INF, would equal infinity and therefore be meaningless. For

225 the vector Z, where less than half of the observations are infinite, a meaningful median of

226 4.5 can still be calculated. Contrary to the median, both the standard deviation and the

227 mean have a breakdown point of zero: one single observation set to infinity implies an

228 infinite mean and an infinite standard deviation, rendering the method based on standard

229 deviation around the mean useless. The same conclusion applies to the Mahalanobis

230 distance, which also has a breakdown point of 0.5.

231   Since the most common methods psychologists use to detect outliers do not rely on

232 robust indicators, switching to robust indicators is our first recommendation to improve

233 current practices. To detect univariate outliers, we recommend using the method based on

234 the Median Absolute Deviation (MAD), as recommended by Leys et al. (2013). The MAD

235 is calculated based on a range around the median, multiplied by a constant (with a default

236 value of 1.4826). To detect multivariate outliers, we recommend using the method based on

237 the MCD, as advized by Leys et al. (2018). Note that, although any breakdown point

238 ranging from 0 to .5 is possible with the MCD method, simulations by Leys et al. (2018)

239 encourage the use of the MCD with a breakdown point of .25 (i.e. computing the mean and

240 covariance terms using 75% of all data) if there is no reason to suspect that more than 25%

241 of all data are multivariate outlying values. For R users, examples of applications of

242 outliers detection based on the MAD and MCD methods are given at the end of the

243 section. For SPSS users, refer to the seminal papers Leys et al. (2018) to compute the

244 MAD, MCD50 (breakdown point = .5) and MCD75 (breakdown point = .25).

245   In addition to the outlier detection method, a second important choice researchers

246 have to make is the determination of a plausible criterion for when observations are

247 considered too far from the central tendency. There are no universal rules to tell you when

248 to consider a value as "too far" from the others. Researchers need to make this decision for

249 themselves and make an informed choice about the rule they use. For example, the same

250 cutoff values can be used for the median plus minus a constant number of absolute

251 deviation method as is typically used for the mean plus minus a constant number of *SD*

252 method (e.g., median plus minus 3 MAD). As for the Mahalanobis distance, the threshold

253 relies on a chi-square distribution with k degrees of freedom, where k is the number of

254 dimensions (e.g., when considering both the weight and height, $k = 2$). A conservative

255 researcher will then choose a Type I error rate of .001 where a less conservative researcher

256 will choose .05. This can be applied to the MCD method. A criterion has to be chosen for

257 any detection technique that is used. We will provide recommendations in the section

258 "Handling Outliers and Pre-registration" and summarize them in the section "Summary of

259 the main recommendations".

260   Finally, it is important to specify that outlier detection is a procedure that is applied

261 only once to a dataset. A common mistake is to detect outliers, manage them (e.g., remove

262 them, or recode them), and then reapply the outlier detection procedure on the new

263 changed dataset.

264   In order to help researchers to detect and visualize outliers based on robust methods,

265 we created an R package (see https://github.com/mdelacre/Routliers). *outliers_mad* and

²⁶⁶ *plot_outliers_mad* functions are created in order to respectively detect and visualise

²⁶⁷ univariate outliers, based on the MAD method. In the same way of thinking, *outliers_mcd*

²⁶⁸ and *plot_outliers_mcd* functions are created in order to respectively detect and visualise

²⁶⁹ multivariate outliers, based on the MCD method. Finally, in a comparative perspective,

²⁷⁰ *outliers_mahalanobis* and *plot_outliers_mahalanobis* are created in order to respectively

²⁷¹ detect and visualise multivariate outliers, based on the classical mahalanobis method. As

²⁷² an illustration, we used data collected on 2077 subjects the day after the terrorist attacks

²⁷³ in Brussels (on the morning of 22 March 2016). We focused on two variables: the sense of

²⁷⁴ coherence (SOC-13 self report questionnaire, Antonovsky, 1987) and anxiety and

²⁷⁵ depression symptoms (HSCL-25, Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1974).

²⁷⁶ Figure 3 shows the output provided by *outliers_mad* applied on the SOC-13 and Table 2

²⁷⁷ shows the plot provided by *plot_outliers_mad* on the same variable.

²⁷⁸   Table 2. *Output provided by the outliers_mad function when trying to detect*

²⁷⁹ *univariate extreme values of sense of coherence (Antonovsky, 1987) on a sample of 2077*

²⁸⁰ *subjects the day after the terrorist attacks in Brussels (on the morning of 22 March 2016)*

²⁸¹ `## $Median`

²⁸² `## [1] 4.615385`

²⁸³ `##`

²⁸⁴ `## $MAD`

²⁸⁵ `## [1] 0.9123692`

²⁸⁶ `##`

²⁸⁷ `## $LL_CI_MAD`

²⁸⁸ `## [1] 1.878277`

²⁸⁹ `##`

²⁹⁰ `## $UL_CI_MAD`

²⁹¹ `## [1] 7.352492`

Detecting values out of the Confidence Interval
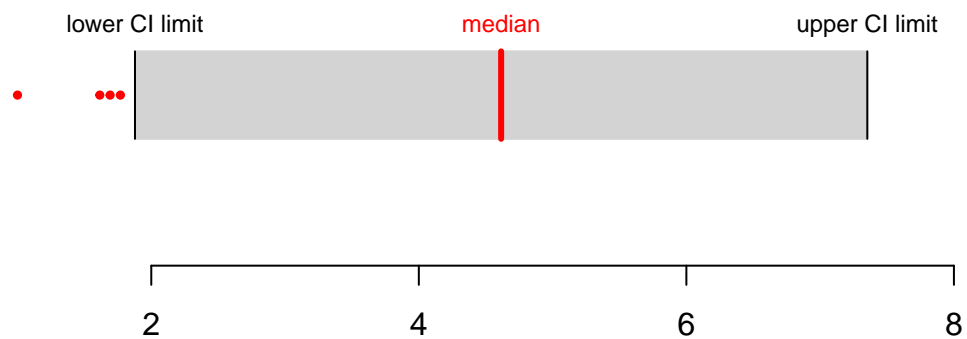**CI = Median ± 3 MAD**

4 outliers are detected

Figure 3. Univariate extreme values of sense of coherence (Antonovsky, 1987) detected by the MAD method on a sample of 2077 subjects the day after the terrorist attacks in Brussels (on the morning of 22 March 2016)

```
292  ##
293  ## $L_outliers
294  ## [1] 1.692308 1.615385 1.769231 1.000000
295  ##
296  ## $U_outliers
297  ## numeric(0)
298  ##
299  ## $outliers
300  ## [1] 1.692308 1.615385 1.769231 1.000000
```

```
301   ##
302   ## $median
303   ## [1] 4.615385
304   ##
305   ## $call
306   ## outliers_mad.default(x = SOC)
307   ##
308   ## $limits
309   ## [1] 1.878277 7.352492
310   ##
311   ## $nb
312   ##   extremely low extremely high          total
313   ##               4               0              4
314   ##
315   ## attr(,"class")
316   ## [1] "outliers_mad"
```

317   Figure 4 shows the plot provided by *plot_outliers_mcd* in order to detect bivariate

318   outliers (in red on the plot) when considering both the SOC-13 and the HSCL-25. The

319   *plot_outliers_mcd* function also returns two regression lines: one computed based on all

320   data and one computed after outliers exclusion. It allows to easily observe if there is a

321   strong impact of outliers on the regression line.

322   Table shows the output provided by *outliers_mcd* on the same variable.

323   Table 3. *Output provided by the outliers_mad function when trying to detect*

324   *univariate extreme values of sense of coherence (Antonovsky, 1987) on a sample of 2077*

325   *subjects the day after the terrorist attacks in Brussels (on the morning of 22 March 2016)*
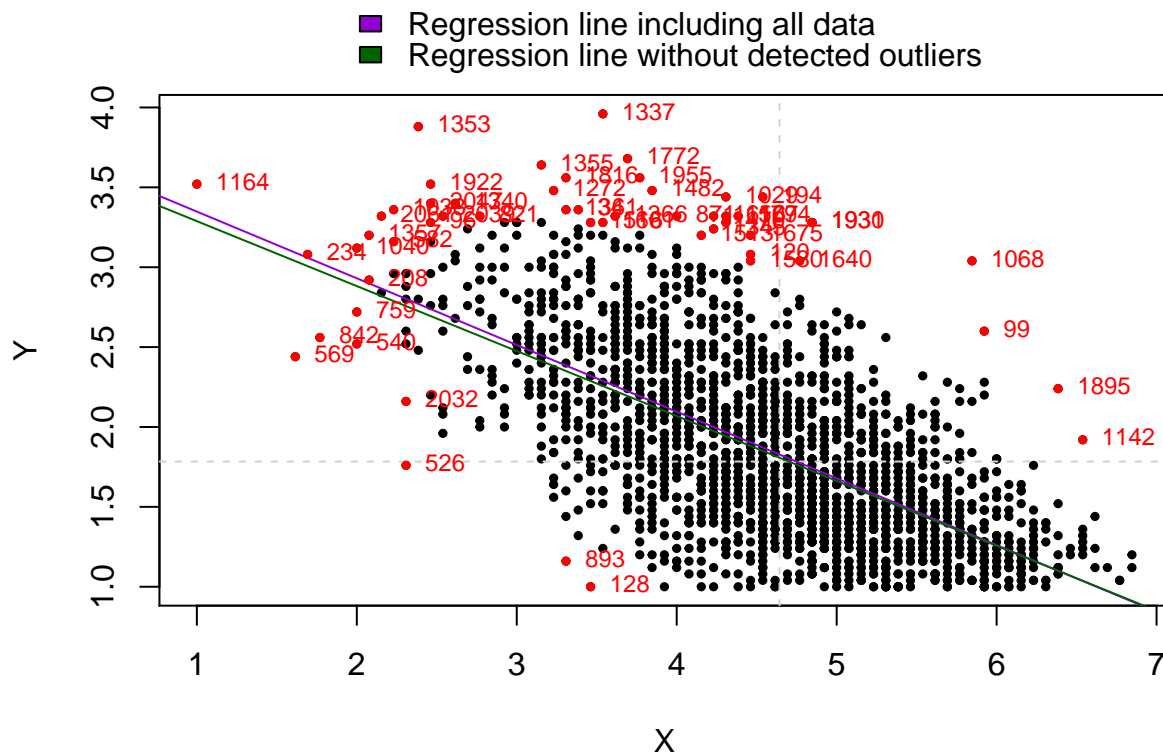
```
326   ## $Median
```

Figure 4. Bivariate extreme values when considering the combination of sense of coherence (Antonovsky, 1987) and anxiety and depression symptoms (Derogatis et al., 1974) detected by the MCD method on a sample of 2077 subjects the day after the terrorist attacks in Brussels (on the morning of 22 March 2016)

```
## [1] 4.615385

##

## $MAD

## [1] 0.9123692

##

## $LL_CI_MAD

## [1] 1.878277

##
```

```
## $UL_CI_MAD
## [1] 7.352492
##
## $L_outliers
## [1] 1.692308 1.615385 1.769231 1.000000
##
## $U_outliers
## numeric(0)
##
## $outliers
## [1] 1.692308 1.615385 1.769231 1.000000
##
## $median
## [1] 4.615385
##
## $call
## outliers_mad.default(x = SOC)
##
## $limits
## [1] 1.878277 7.352492
##
## $nb
##   extremely low extremely high          total
##               4              0              4
##
## attr(,"class")
## [1] "outliers_mad"
```

Handling Outliers

363   After detecting the outliers, it is important to discriminate between *error* outliers

364 and other types of outliers. Error outliers should be corrected whenever possible. For

365 example, when a mistake occurs while entering questionnaire data, it is still possible to go

366 back to the raw data to find the correct value. When it is not possible to retrieve the

367 correct value, outliers should be deleted. To manage other types of outliers (i.e., interesting

368 and random), researchers have to choose among 3 strategies, which we summarize based on

369 the work by Aguinis et al. (2013) as 1) keeping the outliers, 2) removing the outliers, or 3)

370 recoding the outliers.

371   *Keeping* outliers (Strategy 1) is a good decision if most of these outliers rightfully

372 belong to the distribution of interest (e.g. provided that we have a normal distribution,

373 they are simply the 0.27% of values expected to be further away from the mean than three

374 standard deviations). However, keeping outliers in the dataset can be problematic for

375 several reasons if these outliers do in fact belong to an alternative distribution. First, a test

376 could become significant because of the presence of outliers and therefore, the results of the

377 study can depend on a single or few data points, which questions the robustness of the

378 findings. Second, the presence of outliers can jeopardize the assumptions of the parametric

379 tests (mainly normality and equality of variances), especially in small sample datasets.

380 This would require a switch from parametric tests to alternative robust tests, such as tests

381 based on the median or ranks (Sheskin, 2003), or bootstrapping methods (Efron &

382 Tibshirani, 1994), while such approaches might not be needed when outliers that do not

383 belong to the underlying distribution are removed.

384   Note also that some analyses do not have that many alternatives. For example,

385 mixed ANOVA, or factorial ANOVA are very difficult to conduct with nonparametric

386 alternatives, and when alternatives exist, they are not necessarily immune to

387 heteroscedasticity. However, if outliers are a rightful value of the distribution of interest,

388 then removing this value is not appropriated and will also corrupt the conclusions.

389   *Removing* outliers (Strategy 2) is efficient if outliers corrupt the estimation of the

390 distribution parameters, but it can also be problematic. First, as stated before, removing

391 outliers that rightfully belong to the distribution of interest artificially decreases the error

392 estimation. In this line of thinking, Bakker and Wicherts (2014) recommend keeping

393 outliers by default since their presence do not seem to compromise much the statistical

394 conclusions and since alternative tests exist (they suggest using the Yuen-Welch's test to

395 compare means). However, their conclusions only concern outliers that imply a violation of

396 normality but not of homoscedasticity. Moreover, the Yuen-Welch's test uses the trimmed

397 mean as an indicator of the central tendency, which disregards 20% (a common subjective

398 cut-off) of the extreme values (and therefore do not take outliers into account).

399   Second, removing outliers leads to the loss of a large amount of observations,

400 especially in datasets with many variables, when all univariate outliers are removed for

401 each variable. When researchers decide to remove outliers, they should clearly report how

402 outliers were identified (preferably including the code that was used to identify the

403 outliers), and when the way to manage outliers was not pre-registered, report the results

404 with and without outliers.

405   *Recoding* outliers (Strategy 3) avoids the loss of a large amount of data. However,

406 recoding data should rely on reasonable and convincing arguments. A common approach to

407 recoding outliers is Winsorization (Tukey & McLaughlin, 1963) and, where all outliers are

408 transformed to a value at a certain percentile of the data. The observed value of all data

409 below a given percentile observation k (generally $k = 5$) is recoded into the value of the kth

410 percentile observation (and similarly, all data above a given percentile observation, i.e.,

411 (100 - k), is recoded to the value of the (100 - k)th percentile). An alternative approach is

412 to transform all data by applying a mathematical function to all observed data points (e.g.,

413 to take the log or arcsin) in order to reduce the variance and skewness of the data points

(Howell, 1997). We specify that, in our conception, such recoding solutions are only used for pragmatic reasons (i.e., avoiding the loss of too many data) but not for statistical reasons. When possible, it is always best to avoid such seemingly ad hoc transformations in order to cope with data loss. In other words: (1) we suggest to collect enough data so that removing outliers is possible without compromising the statistical power; (2) if outliers are believed to be random, then it is acceptable to leave them as they are; (3) if, for pragmatic reasons, researchers are forced to keep outliers that they detected as outliers influenced by moderators, the Winsorization or such transformations are acceptable in order to avoid a loss of power.

It is crucial that researchers understand that handling outliers is a non-mathematical decision. Mathematics can help to set a rule and examine its behavior, but the decision of whether or how to remove, keep or recode outliers is non-mathematical. As such, it is up to researchers to make an educated guess for a criterion and technique and justify this choice. We developed the nomenclature of outliers provided earlier to help researchers to make such decisions. Error outliers need to be removed when detected as such, as they are not valid observations of the investigated population. Both interesting and random outliers can either be kept, recoded, or excluded. Ideally, interesting outliers should be removed and studied in future studies, and random outliers should be kept. Unfortunately, raw data generally do not allow researchers to easily differentiate interesting and random outliers from each other. In practice, we will therefore treat both of them similarly.

Because multiple justifiable choices are available to researchers, the question of how to manage outliers is a source of flexibility in the data analysis. To prevent the inflation of Type 1 errors, it is essential to specify how to manage outliers following *a priori* criteria, before looking at the data. For this reason, researchers have stressed the importance of specifying how outliers will be dealt with "specifically, precisely, and exhaustively" in a pre-registration document (Wicherts et al., 2016). We would like to add that the least ambiguous description of how outliers are managed takes the form of the computer code

that is run on the data to detect (and possibly recode) outliers. If no decision rules were pre-registered, and several justifications are possible, it might be advisable to report a sensitivity analysis across a range of justifiable choices to show the impact of different decisions about managing outliers on the main results that are reported (see, for example, Saltelli, Chan, & Scott, 2000). If researchers conclude that interesting outliers are present, this observation should be discussed, and further studies examining the reasons for these outliers could be proposed, as they offer insight in the phenomenon of interest and could potentially improve theoretical models.

Pre-registering Outlier Management

More and more researchers (Klein et al., 2018; Nosek, Ebersole, DeHaven, & Mellor, 2018; Veer & Giner-Sorolla, 2016) stress the need to pre-register any material prior to data collection. Indeed, as discussed above, *post hoc* decisions can cast a shadow on the results in several ways, whereas pre-registration avoids an unnecessary deviation of the Type I error from the nominal alpha level. We invite researchers to pre-register: 1) the method they will use to detect outliers, including the criterion (i.e., the cutoff). 2) the decision how to manage outliers.

Several online platforms allow one to preregister a study. The Association for Psychological Science (APS, 2018) non-exhaustively listed the Open Science Framework (OSF), ClinicalTrials.gov, AEA Registry, EGAP, the WHO Registry Network, and AsPredicted.

However, we are convinced that some ways to manage outliers may not be predicted but still be perfectly valid. To face situations not envisaged in the pre-registration or to deal with instances where sticking to pre-registration seems erroneous, we propose three other options: 1) Asking judges blind to the research hypotheses to make a decision on whether or not outliers that do not correspond to the *a priori* decision criteria should be

included. This should be done prior to further analysis, which means that detecting outliers

should be among the first steps when analyzing data. 2) Sticking to the pre-registered

decision regardless of any other argument, since keeping an a priori decision might be more

credible than selecting what seems the best option *post hoc.* 3) Pre-registering a coping

strategy for such unexpected outliers. For example, researchers could decide *a priori* that

all detected outliers that do not fall in a predicted category shall be kept (or removed)

regardless of any *post hoc* reasoning. Lastly, we strongly encourage researchers to report

information about outliers, including the number of outliers that were removed, and the

values of the removed outliers. Best practice would be to share the raw data, possibly a

data plot, as well as the code that was used to detect (and possibly recode) outliers.

Perspectives

Although we provided some guidelines to manage outliers, there are interesting

questions that could be addressed in meta-scientific research. Given the current

technological advances in the area of big data analysis, machine learning or data collection

methods, psychologists have more and more opportunities to work on large datasets

(Chang, McAleer, & Wong, 2018; Yarkoni & Westfall, 2017). In such context, an

interesting research question is whether outliers in a database appear randomly, or whether

outliers seem to follow a pattern that could be detected in such large datasets. This could

be used to identify the nature of the outliers that researchers detect and provide some

suggestions for how to manage them. Four situations can be foreseen: (1) outliers are

randomly distributed and quite rare; (2) outliers are randomly distributed and numerous;

(3) outliers follow a pattern but are quite rare; (4) outliers follow a pattern and are

numerous. The case (1) suggests that outliers belong to the distribution of interest (if the

number of outliers is consistent with what should be expected in the distribution), and, as

such, should be kept. The case (2) would be difficult to interpret. It would suggest that a

large amount of values is randomly influenced by an unknown moderator (or several) able

to exert its influence on any variable. We could be tempted to keep them for pragmatic reasons (i.e., to avoid the loss of a large number of data) but should then address the problem in discussion. In (3) and (4), a pattern emerges, which might suggest the presence of a moderator (of theoretical interest or not). Whenever a pattern emerges (e.g., when the answers of a given participant are consistently outlying from one variable to another), we recommend removing outliers and, eventually, trying to understand the nature of the moderator in future studies.

To go one step further in this line of thinking, some outliers could appear randomly whereas others could follow a pattern. For example, one could suspect that outlying values close to the cutoff belong more likely to the distribution of interest than outliers far from the cutoff (since the further they are the more likely they belong to an alternative distribution). Therefore, outliers close to the cutoff could be randomly distributed in the database, whereas further outliers could follow a pattern. This idea is theoretically relevant, but implies serious hurdles to be overcome, such as devising rules to split outliers in two subsets of interest (one with a pattern the other randomly distributed) without generating false detection.

In conclusion, a useful tool could be a mathematical algorithm that evaluates the detected outliers in a database in order to detect patterns. This tool could also determine whether one subset of outliers follows a pattern whereas other subsets are randomly distributed. It could guide researchers' decisions on how to cope with these types of outliers. However, we currently do not have such a tool and we will leave this topic for further studies.

Summary of the main recommendations

1) Correct or delete obvious erroneous values.

2) Do not use the mean or variance as indicators but the MAD for univariate outliers,

with a cut-off of 3 (for more information see Leys et al., 2013), or the MCD75 (or the

MCD50 if you suspect the presence of more than 25% of outlying values) for the

multivariate outliers, with a chi-square at p = .001, instead (for more information see

Leys et al., 2013).

3) Decide on outlier handling before seeing the results of the main analyses and

pre-register the study at, for example, the Open Science Framework

(http://opensicenceframework.org/).

4) Decide on outlier handling by justifying your choice of keeping, removing or

correcting outliers based on the soundest arguments, at the best of researchers

knowledge of the field of research.

5) If pre-registration is not possible, report the outcomes both with and without outliers

or on the basis of alternative methods (such as Welsch tests, Yuen-Welsch test, or

nonparametric tests, see for example Bakker & Wicherts, 2014; Leys & Schumann,

2010; Sheskin, 2003)

6) Report transparently about how outliers were handled in the results section.

Conclusion

In this paper, we stressed the importance of outliers in several ways: to detect error

outliers; to gain theoretical insights by identifying new moderators that can cause outlying

values; to improve the robustness of the statistical analyses. We also underlined the

problem resulting from the decision on how to manage outliers based on the results yielded

by each strategy. Lastly, we proposed some recommendations based on the quite recent

opportunity provided by platforms allowing to pre-register researchers' studies. We argued

that, above any other considerations, what matters most in order to maximize the accuracy

and the credibility of a given research is to take all possible decisions on outliers' detection

and coping strategies prior to any data analysis.

References

Abelson, R.-P. (1995). *Statistics as principled argument* (Lawrence Earlbaum Associates.). Hillsdale, NJ.

Aguinis, H., Gottfredson, R.-K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, *16*(2), 270–301. doi:10.1177/1094428112470848

Antonovsky, A. (1987). *Unraveling the mystery of health. How people manage stress and stay well* (Jossey-Bass Publishers.). San Francisco.

Bakker, M., & Wicherts, J.-M. (2014). Outlier removal, sum scores, and the inflation of the type i error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods*, *19*(3), 409–427. doi:10.1037/met0000014

Chang, C.-L., McAleer, M., & Wong, W.-K. (2018). Big data, computational science, economics, finance, marketing, management, and psychology: Connections. *Journal of Risk and Financial Management*, *11*(1), 15. doi:10.3390/jrfm11010015

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple correlation/regression analysis for the behavioral sciences* (Lawrence Earlbaum Associates.). Hillsdale, NJ.

Cousineau, D., & Chartier, S. (2010). Outliers detection and treatment: A review. *International Journal of Psychological Research*, *3*(1), 58–67. doi:10.21500/20112084.844

Derogatis, L.-R., Lipman, R.-S., Rickels, K., Uhlenhuth, E.-H., & Covi, L. (1974). The hopkins symptom checklist (hscl): A self-report symptom inventory. *Behavioral Science*, *19*(1), 1–15.

Donoho, D.-L., & Huber, P.-J. (1983). The notion of breakdown point. In P.-J. Bicker, K. Diksum, & J.-L. Hodges (Eds.), *A festschrift for erich l. Lehmann.* California:

Wadsworth.

Efron, B., & Tibshirani, R.-J. (1994). *An introduction to the bootstrap* (Chapman & Hall.). New York.

Howell, D. (1997). *Statistical methods for psychology* (Duxbury Press.). Boston, Massachusetts.

John, L.-K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524–532. doi:10.1177/0956797611430953

Klein, O., Hardwicke, T.-E., Aust, F., Breuer, J., Danielsson, H., Mohr, A.-H., . . . Frank, M.-C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology, 4*(1), 20. doi:10.1525/collabra.158

Kline, R.-B. (2015). *Principles and practice of structural equation modeling* (Guilford publications.). London.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the mahalanobis distance. *Journal of Experimental Social Psychology, 74*, 150–156. doi:10.1016/j.jesp.2017.09.011

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. doi:10.1016/j.jesp.2013.03.013

Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology, 46*(4), 684–688. doi:10.1016/j.jesp.2010.02.007

Mahalanobis, P.-C. (1930). On tests and meassures of groups divergence, theoretical formulae. *International Journal of the Asiatic Society of Bengal, 26*, 541–588.

McClelland. (2000). *Nasty data* (Handbook of research methods in social psychology.).
Cambridge, MA.

McGuire, W.-J. (1997). Creative hypothesis generating in psychology: Some useful
heuristics. *Annual Review of Psychology, 48*, 1–30. doi:10.1146/annurev.psych.48.1.1

Nosek, B.-A., Ebersole, C.-R., DeHaven, A.-C., & Mellor, D.-T. (2018). The
preregistration revolution. *Proceedings of the National Academy of Sciences,
115*(11), 2600–2606. doi:10.1073/pnas.1708274114

Saltelli, A., Chan, K., & Scott, E.-M. (2000). *Sensitivity analysis (vol. 1)* (Wiley.). New
York.

Sheskin, D.-J. (2003). *Handbook of parametric and nonparametric statistical procedures*
(CRC Press.).

Simmons, J.-P., Nelson, L.-D., & Simonsohn, U. (2011). False positive psychology:
Undisclosed flexibility in data collection and analysis allows presenting anything as
significant. *Psychological Science, 22*(11), 1359–1366.
doi:10.1177/0956797611417632

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency
through a multiverse analysis. *Perspectives on Psychological Science, 11*(5),
702–712. doi:10.1177/1745691616658637

Tabachnick, B.-G., & Fidell, L.-S. (2013). *Using multivariate statistics (6th ed.)*
(Pearson.). Boston.

Tukey, J.-W., & McLaughlin, D.-H. (1963). Less vulnerable confidence and significance
procedures for location based on a single sample: Trimming/winsorization 1.
Sankhyã: The Indian Journal of Statistics, Series A, 25(3), 331–352.

Veer, A.-E. van't, & Giner-Sorolla, R. (2016). Pre-registration in social psychology—a
discussion and suggested template. *Journal of Experimental Social Psychology, 67,*

617   2–12. doi:10.1016/j.jesp.2016.03.004

618 Wicherts, J.-M., Veldkamp, C.-L., Augusteijn, H.-E., Bakker, M., Van Aert, R., & Van

619   Assen, M.-A. (2016). Degrees of freedom in planning, running, analyzing, and

620   reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in*

621   *Psychology*, *7*(1832), 1–12. doi:10.3389/fpsyg.2016.01832

622 Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology:

623   Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6),

624   1100–1122. doi:10.1177/1745691617693393