How (Not) to Exclude Outliers: Within-Conditions Exclusions Lead to Dramatic Increases in False-Positive Rates

WORKING PAPER – PLEASE DO NOT CITE OR DISTRIBUTE WITHOUT THE AUTHOR'S APPROVAL

This Version : November 2020

Quentin André

Quentin André (quentin.andre@colorado.edu) is an assistant professor of marketing at the Leeds School of Business, University of Colorado Boulder. I thank Jiyin Cao, Dejun Tony Kong and Adam Galinsky for making the data of their paper publicly available. I am grateful to Bart de Langhe, Bram van den Bergh, Uri Simonsohn, Joe Simmons, Leif Nelson, and Zoé Ziani-Franclet for their helpful advice, comments, and references. The OSF repository of the project contains the code and files needed to reproduce all the figures and analyses reported in this manuscript, as well as additional figures and analysis.

How (Not) to Exclude Outliers: Within-Conditions Exclusions Lead to Dramatic Increases in False-Positive Rates

ABSTRACT

When researchers choose to identify and exclude outliers from their data, should they do so across all the data, or within experimental conditions? A survey of recent papers published in the Journal of Experimental Psychology: General shows that both methods are widely used, and common data visualization techniques suggest that outliers should be excluded at the condition-level. However, I highlight in the present paper that removing outliers by condition runs against the logic of hypothesis testing, and that this practice leads to unacceptable increases in false-positive rates. I demonstrate that this conclusion holds true across a variety of statistical tests, exclusion criterion and cutoffs, sample sizes, and data types, and show in simulated experiments that Type I error rates can be as high as 29%. I then replicate this result in the context of a recent paper excluding outliers per condition (Cao, Kong, and Galinsky, 2020). Using the authors' original data, I show that excluding outliers at the condition level can bring the likelihood of a false-positive result up to 47%, and demonstrate that the exclusion strategy reported by the authors is associated with a 56% Type I error rate. I conclude with a list of alternatives to within-condition exclusions.

INTRODUCTION

Data about human behavior is noisy. Participants misread instructions, get distracted during the task, experience computer errors, or simply do not take a study seriously. To reduce noise and increase statistical power, it is common practice to identify such "nasty data" (McClelland, 2014) in people's response to a task. A common example of such "aberrant responses" are data points that are "too extreme" to reflect to genuine responses.

A well-defined threshold sometimes exists to distinguish between valid responses and extreme responses. For reaction-time to a visual stimuli (e.g., in a Stroop task), it is generally accepted that responses faster than 200ms indicate a human or software error (e.g., Ng & Chan, 2012). For a muscle reaction to an auditory stimuli, the shorter threshold of 100ms is generally considered (Pain & Hibbs, 2007). In most circumstances however, no such threshold is available, and researchers instead focus on the identification of "outliers": Data points that are "inconsistent" or "too far removed" from the remainder of the data (Barnett & Lewis, 1994).

How far is "too far"? Over the years, multiple methods have been offered to establish a threshold between regular responses and outliers, and recent papers have summarized the different techniques available to researchers (Aguinis et al., 2013; Leys et al., 2019). In particular, three metrics are commonly used in papers to detect univariate outliers: The z-score (the response's deviation from the mean, expressed in units of standard deviation), the Median Absolute Distance (MAD; the response's deviation from the median; Leys et al., 2013), and the Inter-Quartile Range (IQR) distance (the response's distance from the upper or lower quartile of the distribution).

The latter method is commonly encountered in the context of boxplots. Since Tukey (1977), boxplots have been widely used by researchers to visualize and report the distribution of

their data. A boxplot summarizes a distribution by displaying a "box" (representing the 25th percentile, the median and the 75th percentile of the data) and two "whiskers" (each representing a 1.5 IQR band extending away from the box). Any data point that falls outside of the "whiskers" is flagged as an outlier, with some statistical software (e.g., SPSS) further distinguishing between outliers and "extreme outliers" (further than 3 IQR from the box). Figure 1 displays an example in the context of an experiment with two conditions: The boxplot identifies no outliers in the "Control" condition, and one outlier in the "Treatment" condition[1].

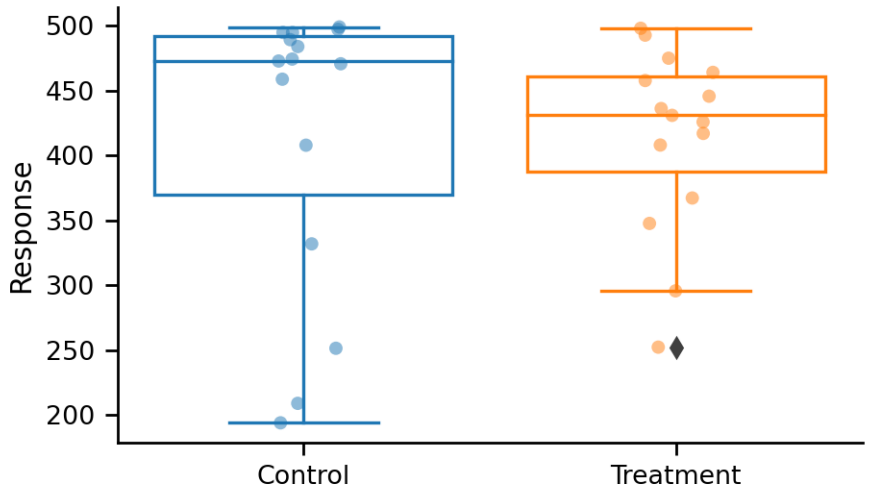


Figure 1.

From this visualization, a researcher might conclude that it is acceptable (and perhaps desirable) to identify outliers within conditions. But is this approach correct? In an experiment with multiple conditions, should one identify and remove outliers *across all the data*, or *within each condition*? Recent papers covering the topic of outlier removal (e.g., Aguinis et al., 2013; Leys et al., 2019) have not broached this important question, and an inspection of the most-cited

[1] This "split-by-condition" boxplot is the default in SPSS, further suggesting that it is the recommended approach. To obtain a boxplot across all the data, researchers must instead select the "1-D Boxplot" option.

books and papers on the topic of univariate outliers (e.g., Barnett & Lewis, 1994; Ghosh & Vogt, 2012; Hawkins, 1980; Miller, 1993; Osborne & Overbay, 2004; Ratcliff, 1993) reveals no explicit discussion of this question[2].

It is of course not appropriate to apply *different exclusion rules* to different conditions: One cannot for instance remove all responses that are more than 3 SD away from the mean in the "Control" condition, and all responses that are more than 2 SD away from the mean in the "Treatment" condition. It is transparent that doing so would introduce a systematic difference between the two conditions and threaten the researchers' ability to compare them.

But is it appropriate to apply the *same exclusion rule* (e.g., "any response that is more than 1.5 IQR lower than the 25th percentile") *within conditions taken separately*? For instance, should a researcher, on the basis of the boxplot presented in Figure 1, exclude the "250" response in the "Treatment" condition, but keep the "250", "210," and "200" responses in the "Control" condition? A survey of recent papers published in the Journal of Experimental Psychology: General suggests that it is, indeed, an appropriate decision: Out of 31 papers published in 2019 and 2020 that report univariate outliers exclusions, 9 of them are excluding outliers within the different experimental conditions[3].

In the present article however, I warn that it is in fact not appropriate to identify and exclude outliers within conditions. I highlight that doing so runs against the logic of null-hypothesis significance testing, and present evidence that this practice leads to high false-positive rates, both in simulated and actual data.

---

[2] To the best of my knowledge, only Cousineau and Chartier (2010) and Meyvis and van Osselaer (2018) have offered an explicit discussion of this question. Both papers (incorrectly) suggest that outliers should be searched for, and excluded, within conditions.

[3] A search for all papers including the keyword "outlier" published since 2019 in JEP: General returned 43 papers, 31 of which included a univariate exclusion. The spreadsheet summarizing this search is available on the OSF repository of the paper.

A REFRESHER ON NULL-HYPOTHESIS SIGNIFICANCE TESTING

To determine if a treatment had an effect, researchers commonly engage in null-hypothesis significance testing (NHST): They compare the observed impact of the treatment to what would be expected if the treatment did not have any effect (the *null hypothesis*). This null hypothesis consists of a set of assumptions about the process that generated the data, and forms the basis of the statistical set (Nickerson, 2000). For instance, the null hypothesis of a Student t-test is that the two groups were independently sampled at random from a common normal distribution, and therefore have equal mean.

From these assumptions, statisticians derive the *theoretical distribution* of the test statistics under the null: The distribution of results that the statistical test would return when the treatment does not have any effect. The NHST procedure then compares the result observed in the experiment to this theoretical distribution and returns a p-value: the probability of observing a result at least as extreme as that of their experiment under the null hypothesis. If the p-value is smaller than a pre-determined threshold (typically $\alpha = .05$), it is common practice to conclude that the null hypothesis is not an appropriate description of the observed data, and to "reject the null."

However, the p-value thus obtained is only valid if the structure of the data matches the assumptions of the statistical test. When one (or several) assumptions are violated, the *theoretical* distribution of the test statistics under the null (i.e., the distribution of values that is predicted from the assumptions of the statistical test) will no longer match the *empirical* distribution of the test statistics under the null (i.e., the distribution of values that we will actually observe in the

experiment when the null hypothesis is true). The test then becomes "inexact," and its conclusions may no longer be trusted.

Specifically, if extreme values are more frequent in the theoretical distribution than in the empirical distribution, the test is "too conservative": The threshold to reject the null is too high. On the contrary, if extreme values are less frequent in the theoretical distribution than in the empirical distribution, the test becomes "too liberal": The threshold to reject the null is too low.

EXCLUDING OUTLIERS WITHIN CONDITIONS INVALIDATES NULL-HYPOTHESIS TESTING

While small deviations from the assumptions are typically inconsequential, larger deviations can threaten the conclusions of statistical tests. In particular, the practice of excluding outliers within conditions defies the logic of null-hypothesis significance testing: When researchers choose to exclude outliers within conditions (rather than across the data), they are considering that the conditions are different from each other… and have therefore implicitly rejected the null hypothesis. But if we have already accepted that the null hypothesis is not true, how can we then interpret a procedure that assumes that the null is true?

This paradox is not simply an intellectual curiosity: When outliers are identified and excluded within conditions, the data-generating mechanism of the experiment changes, and the assumptions of statistical tests are automatically violated. To illustrate the consequences of this violation, consider a simple two-cells experiment first: A team of researchers will elicit a single response from 200 participants, randomly assigned to a "Control" condition or a "Treatment" condition. The researchers are unaware of it, but the treatment does not have any effect: The response for all participants is drawn from the same log-normal distribution.

The researchers will compare the responses in the two conditions using a t-test, but they are concerned about the presence of outliers. They therefore decide to use a boxplot, and to exclude any participant that is flagged as an outlier prior to analysis. However, the two researchers disagree in how the boxplot should be used: Researcher A argues that they should identify and exclude outliers *across* all the data, while Researcher W believes that they should identify and exclude outliers *within* each condition. In light of this disagreement, they decide to try both strategies.

The histograms in Figure 2 shows the results that each researcher would obtain if they repeated the experiment a large number of times. The dashed line on each panel displays the *theoretical* null distribution of the t-test: The results that would be expected when the assumptions of the t-test are met (i.e., the two samples are independently sampled at random from the same distribution). Since the null hypothesis is correct in this case, we should expect the results of the experiments to closely match this distribution.

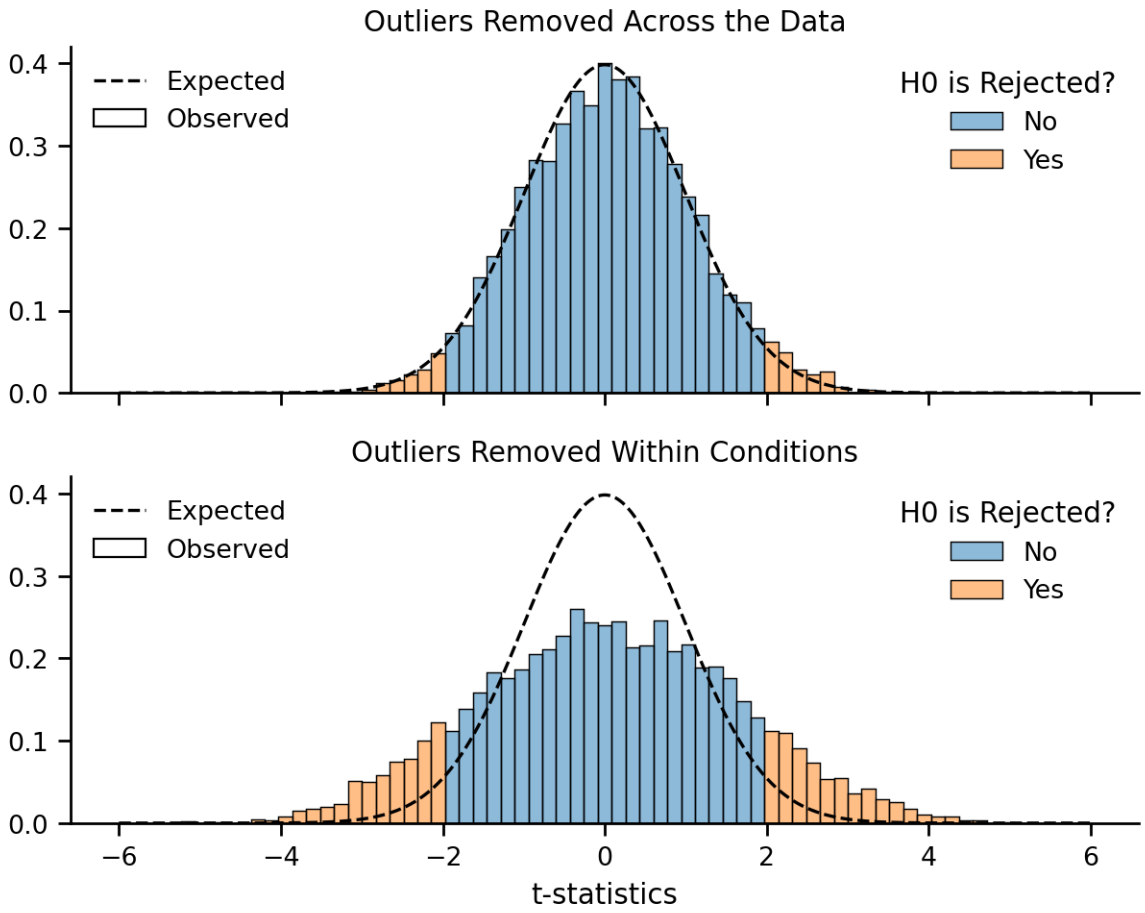


Figure 2

The histogram in the top panel shows the results that Researcher A, who is excluding outliers across the data, would obtain. We see that these results closely match the theoretical null distribution: Extreme differences between conditions are rare, such that the null hypothesis is, as expected, only rejected 5% of the time. This confirms that excluding outliers across the data does not violate the assumptions of the statistical test, and therefore maintains the Type I error at a nominal level.

In contrast, we see in the bottom panel that the differences observed by Researcher W are larger than what the theoretical distribution would predict. Differences that the theoretical null distribution would consider extremely unlikely are, in fact, relatively common when the outliers

are excluded within conditions. This translates into a Type I error rate that is grossly inflated: Researcher W would incorrectly reject the null 22% of the time.

This result has an intuitive explanation. A key assumption of the null hypothesis of the t-test (and of almost all NHST procedures) is that the samples are drawn from a common distribution. This assumption is violated once outliers are excluded within conditions: Each of the samples was submitted to a different data transformation that *amplified* any pre-existing difference between them.

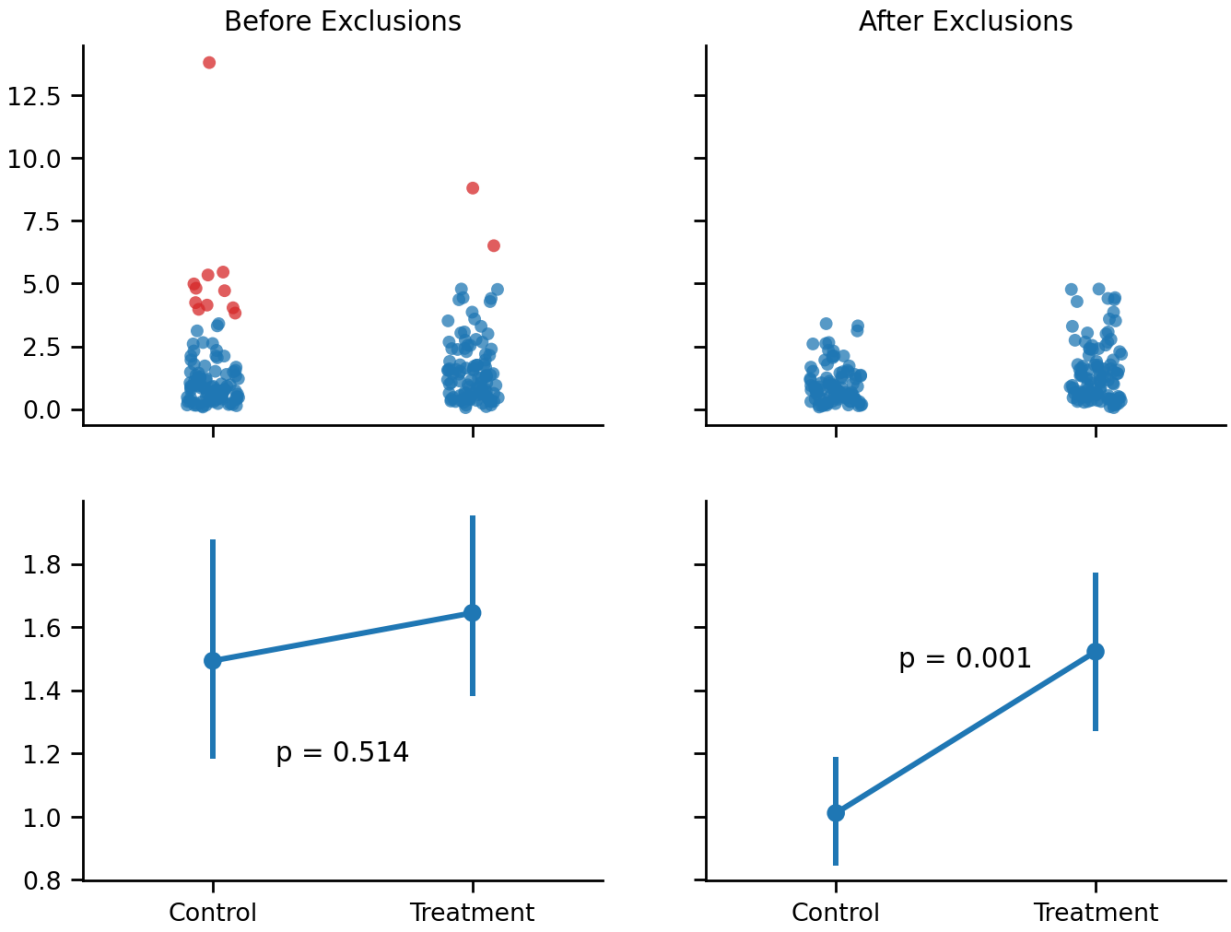


Figure 3.

Figure 3 illustrates this amplification of small differences. The two groups are drawn from the same distribution but, by chance, the mean of the "Control" condition is slightly lower than

the mean of the "Treatment" condition. Because of this minute difference, the same high values that are considered outliers (red dots) in the "Control" condition are not considered as such in the "Treatment" condition. As a consequence, the difference between the two conditions will become larger after exclusions. Since the t-test that compares the two conditions does not "know" that this procedure was applied to the data, it underestimates the magnitude of the differences that can be observed under the null, and will reject the null more often than it should. We see that a difference that was originally considered consistent with the null ($p = .514$) becomes a highly significant result ($p = .001$) once outliers are excluded within conditions.

QUANTIFYING THE PROBLEM IN SIMULATED DATA

This result is not specific to the t-test, or to this particular experiment: When outlier exclusions are not blind to experimental conditions, any statistical procedure that does not account for this exclusion procedure will yield invalid conclusions. In support of this claim, I report in this section the results of simulated experiments showing that the inflation of false-positive rates is observed across a variety of statistical tests, data types, sample sizes, and exclusion criteria.

I considered 243 (3^5) different experimental setups, obtained by orthogonally crossing three possible distribution of responses (a normal distribution, a normal distribution with outliers[4], and a log-normal distribution), three possible samples sizes (50, 100 or 250 observations per condition), three possible methods (z-score, IQR, and Median Absolute Difference) and three possible cutoffs (1.5, 2 or 3 times the z-score/IQR distance/Median

[4] This distribution simulates the presence of large outliers by sampling from a standard normal $\mathcal{N}(0, 1)$ with 95% probability or from $\mathcal{N}(5, 1)$ with 5% probability.

Absolute Difference) for excluding outliers, and three different statistical tests: A parametric test of differences in means (Welsch's t-test), a non-parametric test of differences in central tendencies (Mann-Whitney's U), and a non-parametric test of differences in distribution shapes (the Kolmogorov-Smirnov test)[5].

To obtain a smooth distribution of the potential outcomes, I generated 10,000 simulated experiments in each of those 243 different setups, for a total of 2,430,000 simulated experiments. In each experiment, I draw two samples at random from the same population (such that the null hypothesis is true), and observe the p-value of the differences between the two samples under three different outlier exclusion strategies: 1. No exclusions, 2. Exclusions across the data, 3. Exclusions within each condition. For conciseness, I only present the results by exclusion rules and cutoffs in Figure 4 below. The full breakdown of results (by sample size, data type, statistical test, exclusion rules, and exclusion cutoffs) is reported on the OSF repository of the paper.

[5] These three statistical tests cover the majority of the NHST procedures that are applied to continuous univariate data. For instance, the z-test is the asymptotic equivalent to the t-test when N is large, the F-test of an ANOVA is the k-samples analog to the t-test, the Kruskal-Wallis test is the k-samples analog to the Mann-Whitney test…

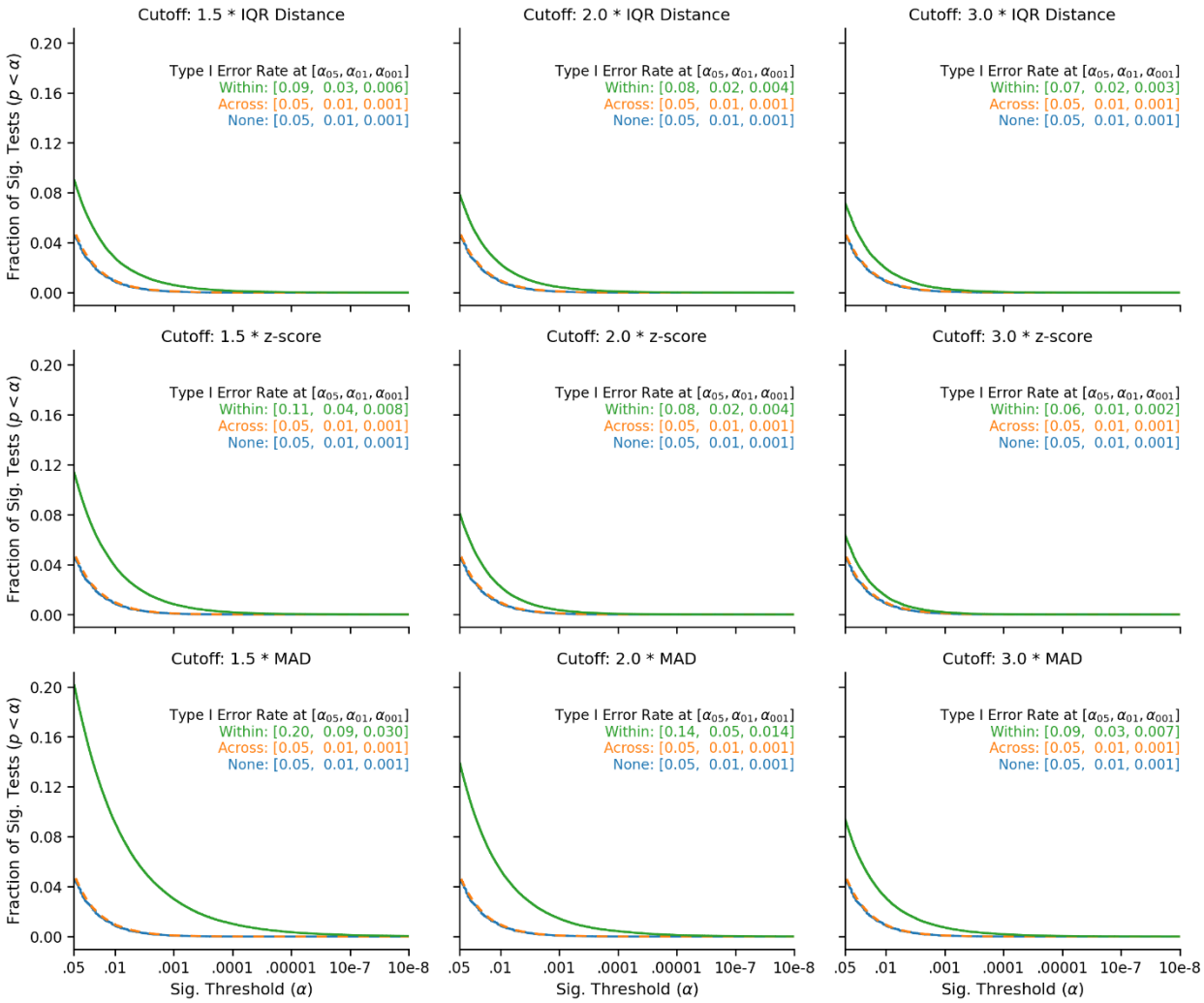


Figure 4.

Figure 4 presents the survival curves of the tests: The fraction of tests that were significant (on the y-axis) at a given significance threshold (on the x-axis), under different outlier exclusion cutoffs (panels) and different outlier exclusion strategies (lines). If the assumptions of the statistical procedure are not violated, we should observe nominal false-positives rate: We should see that 5% of tests are significant at $\alpha = .05$, that 1% are significant at $\alpha = .01$, and that .1% are significant at $\alpha = .001$. We indeed see this pattern when no outliers are excluded (blue) and when the outliers are excluded across the data (orange), which confirms that those practices do not violate the assumptions of the statistical tests.

In contrast, we observe an increase in false-positive rates when applying the exclusion cutoff within conditions (green line). The simulations show that the increase is systematic and serious, and that it varies significantly across exclusion cutoffs: The most favorable case shows a 20% increase in the false-positive rate (from 5% to 6%), and the least favorable case shows a 400% increase (from 5% to 20%). In general, we see that the less stringent the cutoff, the more serious the inflation in false-positive rates: Lower cutoffs increase the number of values excluded within each condition, which further amplifies the original differences between the two samples.

The full breakdown of results (reported on the OSF repository of the paper) reveals significant heterogeneity in the severity of the issue across data types and statistical tests. In particular, the problem appears to be most severe in the presence of parametric tests (i.e., Welsch's t-test) applied to skewed data (i.e., the log-normal distribution), with Type I error rates always higher than 10%, and as high as 29%. It is a concerning result: Outliers are most frequently excluded in the context of over-dispersed data (e.g., reaction times, willingness-to-pay, sum-scores…), and parametric tests are more commonly used than their non-parametric counterparts.

REPLICATING THE PROBLEM IN RECENT DATA

The conclusions presented so far paint a grim picture: Analysis of simulated data suggest that excluding outliers will magnify any minute difference between conditions, and lead to unacceptable Type I error rates. In the next section, I demonstrate that the inflation of false-positive rates is not unique to simulated data, and that the problem is also present (and potentially more severe) in actual data collected from human participants. To do so, I propose a re-analysis of a recent paper: Cao, Kong, and Galinsky (2020). This paper offers an interesting case study for

two reasons: It is one of the most recent paper in a major psychological journal in which outliers were excluded within conditions, and the authors made the raw data of their paper available.

In this paper, the authors report the result of two experiments comparing the negotiation outcomes (measured by Pareto efficiency) of dyads who were randomly assigned to one of three conditions: a “No Eating” condition, a “Separate Eating” condition, and a “Shared Eating” condition. The authors find in both experiments that dyads who were assigned to the “Separate Eating” condition have a lower Pareto efficiency than dyads who were assigned to the “Shared Eating” condition, and conclude that sharing a meal facilitates cooperation. In both experiments, the outliers are removed within conditions: Any dyad with a Pareto efficiency lower than “three times the interquartile range below the lower quartile" of its condition is removed from the data. In addition, it appears that this procedure was recursively applied to the data: After excluding the outliers, the same threshold is applied again within each condition, and newly identified outliers are removed, until no new outliers are found. This procedure, while unusual, has occasionally been recommended to facilitate the identification of outliers in heterogeneous data (Meyvis & Van Osselaer, 2018; Schwertman & de Silva, 2007; Van Selst & Jolicoeur, 1994).

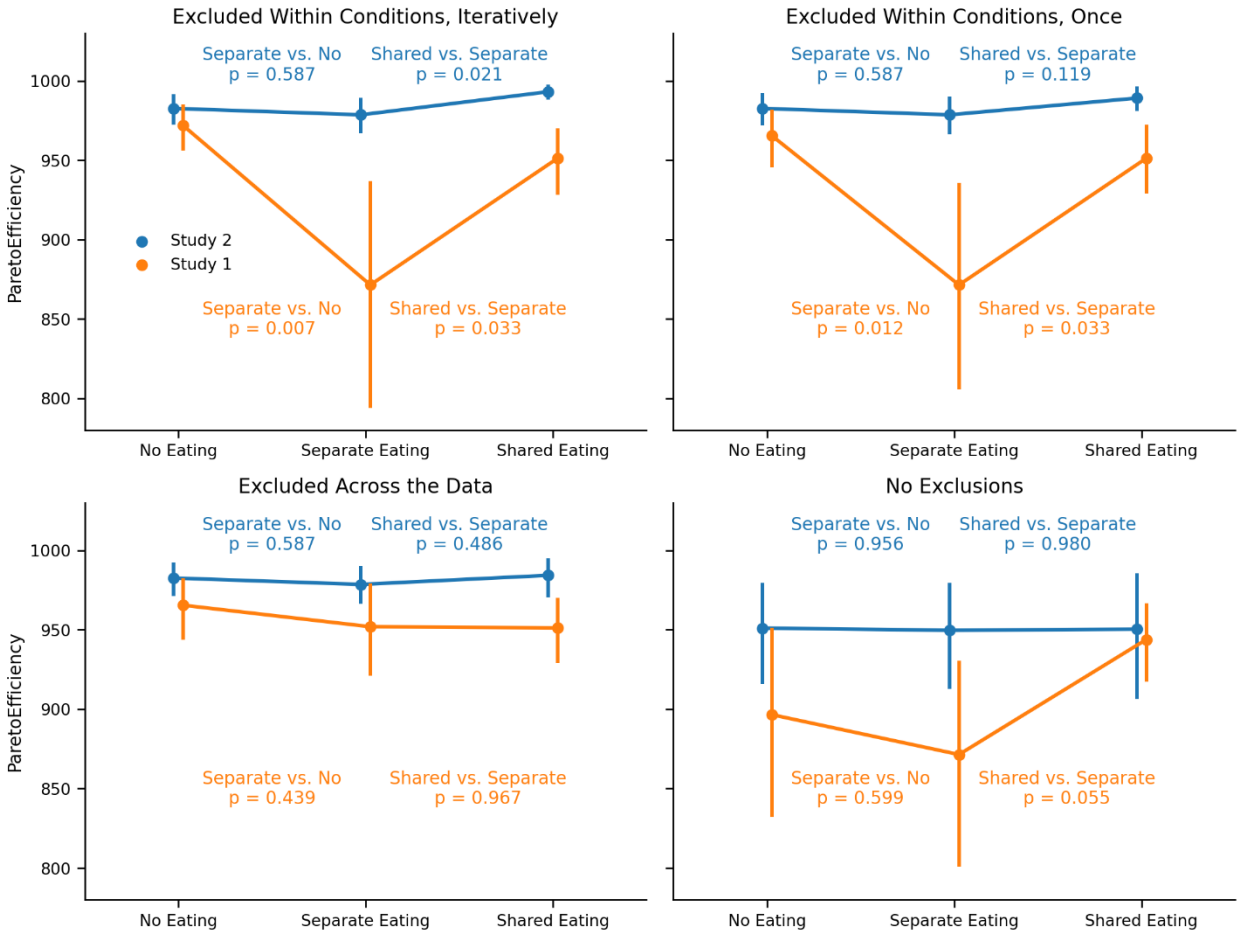
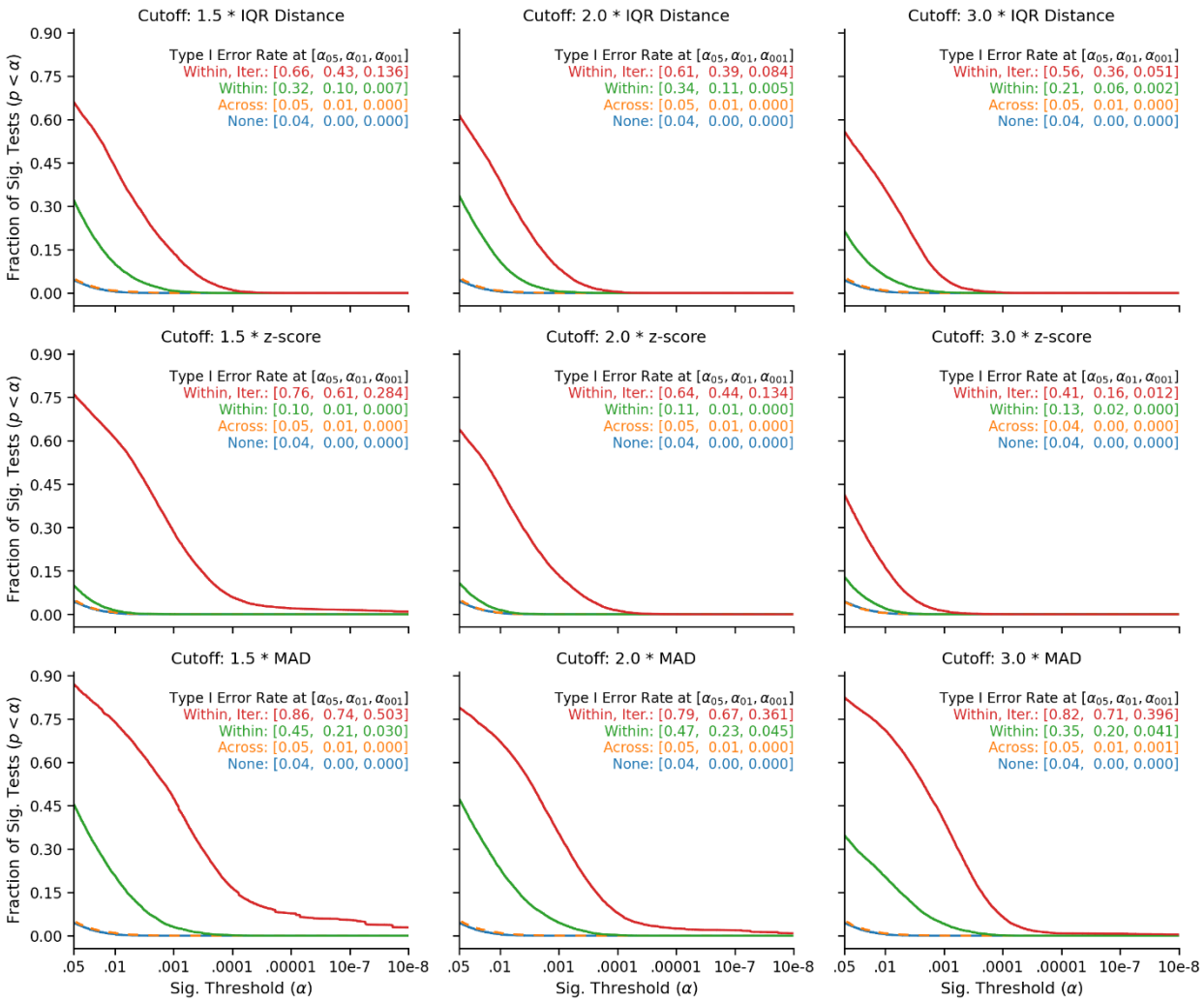


Figure 5/

Having access to the authors' raw data allows me to compare the results obtained in each study under different exclusion strategies (Figure 5). The upper-left panel presents the authors' original results: When outliers are iteratively excluded within conditions, the difference between the "Separate Eating" and the "Shared Eating" condition is large and significant. However, we see that this difference is attenuated when only round of exclusion is performed within conditions (upper-left panel), and shrinks to a small, non-significant amount when outliers are excluded across conditions (bottom-left panel) or when no exclusions are performed (bottom-right panel). This descriptive analysis confirms that excluding outliers within conditions can magnify small differences that are originally present between conditions.

This analysis does not necessarily mean that the authors' result is a false-positive (to reach this conclusion, one would need to know whether the null hypothesis is true). However, one can compute the likelihood of observing a false-positive in the context of the authors' experiment. To do so, I reassign the condition labels in the data of Study 2 at random[6] (so that the differences between conditions are truly zero in expectation), perform a t-test on the two groups, and check how often the Pareto efficiency of dyads in the "Separate Eating" is significantly different from the Pareto efficiency of dyads in the "Shared Eating" condition.



[6] I focus on Study 2 because it was pre-registered, but similar results (reported on the OSF repository of the paper) are observed in Study 1.

Figure 6.

Figure 6 replicates the Type I error inflation previously observed in simulated data (the exclusion criteria used by the authors, removing observations that are at least three times the interquartile range below the lower quartile, appears in the upper right corner). While outlier exclusions are not associated with higher false-positive rates when they are performed across the data (the orange line), the likelihood of a false-positive result increases sharply when outliers are excluded within conditions (the green line): It is always higher than 10%, and can be as high as 47%.

Finally, this figure shows that the increase is even stronger when the outliers are iteratively excluded within conditions (the red line). This result is not surprising: Iterated exclusions are causing the two conditions to diverge even further, and routinely lead to differences that the theoretical null distribution would consider extremely unlikely. In particular, the upper right panel show that the exclusion strategy and cutoff reported in the paper is associated with a false-positive rate of 56%, which translates into a false-positive rate of 28% for the directional hypothesis pre-registered by the authors.

SUMMARY AND RECOMMENDATIONS

A survey of the recent literature, and the common practice of splitting boxplots by conditions, suggest that excluding outliers within conditions is an acceptable strategy. I have demonstrated that this conclusion is erroneous: Excluding outliers by condition amplifies the small differences that are normally expected under the null, which results in inflated Type I error rates. This result is observed for parametric and non-parametric tests, different exclusion criteria, different cutoffs, different sample sizes, and in both simulated and real data. In light of these

results, I conclude that the practice of excluding outliers within conditions is a "questionable research practice" that makes false-positive far too likely (Simmons et al., 2011), and that it should be abandoned by researchers.

What if the Pattern of Responses Does Differ Across Conditions?

It might be tempting to justify the practice of excluding outliers within conditions by the observation that the conditions *look* different: One condition appears to have a higher mean, or a smaller dispersion, than the other(s). As mentioned earlier however, this justification is a paradox: If we assume that the pattern of responses differ across conditions, we have already rejected the null hypothesis, which then begs the interest of using a statistical test to compare them. If the researchers *know* that the values differ across conditions (e.g., when measuring the height of adults vs. children, or how testing how fast people can solve an easy vs. hard math puzzle), then they do not need a statistical test to compare the conditions, and can exclude outliers by group. However, if they want to test for the presence of a difference between the conditions, they cannot exclude outliers by condition and apply a regular statistical test.

Can Researchers Ignore this Problem if they Apply a Stricter Alpha Level, or if they Use Bayesian Statistics?

Using a stricter alpha level would not solve the issue. First, the exact impact of excluding outliers within conditions on false-positive rates is variable and unpredictable: In the simulations and in real data, the increase could be as low as 20% (from 5% to 6%), and as high as 940% (from 5% to 47%), depending on the type of statistical test, the rule for excluding outliers, and the exact structure of the data. As a consequence, it is unclear how large of a correction researchers should apply. Second, the stricter the alpha level, the lower the power of the test (all

other things being equal): Researchers should not adopt a practice that would harm their ability to detect true effects when better alternatives are available.

The default estimation procedures in the Bayesian researcher toolbox (e.g., the Bayesian t-test; Kruschke, 2013) would also not offer a remedy. Indeed, the problem is not specific to NHST, and Bayesian inferences also hinges on the assumption that the data-generating mechanism is correctly identified (Gelman et al., 2013). For this reason, any procedure that does not explicitly model the per-condition exclusion (and therefore does not account for the inflation of differences between conditions) will also yield inaccurate results. In support of this claim, I present additional analysis (reported on the OSF repository of the paper) showing that when a Bayesian t-test is applied to null data, the highest density interval (HDI, the Bayesian counterpart of the confidence interval) contains zero more frequently when exclusions are performed within-condition exclusions (vs. across the data).

How Should Researchers Deal with Outliers Then?

The simplest recommendation would be to exclude outliers across the data, and not within conditions. As shown in the simulations presented in this article, this practice does not cause a Type I error inflation.

A second possibility would be not to exclude the outliers, and to analyze the data using non-parametric tests (e.g., rank-based tests, or resampling-based tests; Erceg-Hurn & Mirosevich, 2008), or heavy-tailed Bayesian models (e.g., West, 1984), that are less sensitive to the presence of extreme values.

Finally, if sample sizes are small, and if power to detect an effect is an important concern, researchers may consider using specific estimators developed for trimmed and winsorized groups (e.g., Kim, 1992; Wilcox, 2011; Wu, 2006; Yuen, 1974). These specific procedures account for

the fact that the data was transformed within conditions, and therefore maintain a nominal Type I error rate. However, they come at some overhead to the researcher: It is important to select the estimator that matches the exclusion strategy that was used (i.e., deviation from mean or median; removing vs. winsorizing), and the design of the experiment (between-subjects, within-subjects, or mixed design).

What If the Experiment Design is More Complex?

The present paper discussed the problem of by-condition exclusions in the context of single-factor experiments. However, the same general principle applies to make complex designs (e.g., factorial, or repeated-measure designs): Any outlier exclusion procedure must be blind to the factor(s) that researchers are interested in testing. The following examples illustrate this rule.

Example 1: A between-subject factorial design in which participants are randomly assigned to solve one easy (vs. hard) math puzzle after engaging in a mindfulness (vs. relaxation) workshop.

It is clear that people will take more time to solve the hard math puzzle than the easy math puzzle. Researchers can therefore choose not to test for the impact of puzzle difficulty, and to exclude outliers within the “hard puzzle” and “easy puzzle” conditions taken separately. However, they cannot exclude outliers within the “mindfulness” and the “relaxation” conditions taken separately without compromising their ability to test for an effect of the workshop type.

Example 2: Identical design, but the researchers are only interested in the interaction effect between the workshop type and the difficulty of the puzzles.

The researchers can again choose to apply a different exclusion threshold to the “hard puzzle” and “easy puzzle” groups. Alternatively, they can apply a different exclusion threshold to the “mindfulness” and “relaxation” groups (if, for instance, they expect the effect of the

workshop type to be larger than the effect of puzzle difficulty). However, they cannot decide to exclude outliers within each of the four conditions: If they did so, their exclusion procedure would no longer be blind to the interaction between the two factors that they want to test.

Example 3: The same design again, but researchers are interested in all effects: The main effect of the workshop, the main effect of difficulty, and the two-way interaction.

Since all factors are interesting to the researchers, they have to exclude outliers across all four conditions.

Example 4: A within-subject study of reaction times: Participants are tasked, over many repeated trials, to find happy faces and angry faces in a crowd (Becker et al., 2011). Researchers want to test the hypothesis that happy faces are found faster than angry faces.

It is appropriate to exclude outliers at the participant-level: Doing so would account for the between-participant heterogeneity in reaction times, and facilitate the identification of "noisy" trials in which the participant was distracted. However, they should not exclude outliers within the "happy faces" trials and the "angry faces" trials separately.

REFERENCES

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, *16*(2), 270-301. https://doi.org/10.1177/1094428112470848

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. Wiley.

Becker, D. V., Anderson, U. S., Mortensen, C. R., Neufeld, S. L., & Neel, R. (2011). The face in the crowd effect unconfounded : Happy faces, not angry faces, are more efficiently detected in single- and multiple-target visual search tasks. *Journal of Experimental Psychology: General*, *140*(4), 637-659. https://doi.org/10.1037/a0024060

Cao, J., Kong, D. T., & Galinsky, A. D. (2020). Breaking Bread Produces Bigger Pies : An Empirical Extension of Shared Eating to Negotiations and a Commentary on Woolley and Fishbach (2019). *Psychological Science*, 0956797620939532. https://doi.org/10.1177/0956797620939532

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

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods : An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591-601. https://doi.org/10.1037/0003-066X.63.7.591

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd Edition). Chapman and Hall/CRC.

Ghosh, D., & Vogt, A. (2012). *Outliers : An Evaluation of Methodologies*. 6.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). Springer.

Kim, S.-J. (1992). The metrically trimmed mean as a robust estimator of location. *The Annals of Statistics*, *20*(3), 1534-1547.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573-603. https://doi.org/10.1037/a0029146

Leys, C., Delacre, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Psychology*, *32*(1), 5. https://doi.org/10.5334/irsp.289

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. https://doi.org/10.1016/j.jesp.2013.03.013

McClelland, G. H. (2014). *Nasty data : Unruly, ill-mannered observations can ruin your analysis.*

Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the Power of Your Study by Increasing the Effect Size. *Journal of Consumer Research*, *44*(5), 1157-1173. https://doi.org/10.1093/jcr/ucx110

Miller, J. N. (1993). Tutorial review—Outliers in experimental data and their treatment. *The Analyst*, *118*(5), 455-461. https://doi.org/10.1039/AN9931800455

Ng, A. W. Y., & Chan, A. H. S. (2012). Finger Response Times to Visual, Auditory and Tactile Modality Stimuli. *Hong Kong*, 6.

Nickerson, R. S. (2000). Null hypothesis significance testing : A review of an old and continuing controversy. *Psychological methods*, *5*(2), 241.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, *9*(1), 6.

Pain, M. T., & Hibbs, A. (2007). Sprint starts and the minimum auditory reaction time. *Journal of sports sciences*, *25*(1), 79-86.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological bulletin*, *114*(3), 510.

Schwertman, N. C., & de Silva, R. (2007). Identifying outliers with sequential fences. *Computational Statistics & Data Analysis*, *51*(8), 3800-3810. https://doi.org/10.1016/j.csda.2006.01.019

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. https://doi.org/10.1177/0956797611417632

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *The Quarterly Journal of Experimental Psychology Section A, 47*(3), 631-650.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *46*(3), 431-439.

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic press.

Wu, M. (2006). *Trimmed and Winsorized Estimators*.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika, 61*(1), 165-170. https://doi.org/10.1093/biomet/61.1.165