© 2021, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/xge0001069

Outlier Exclusion Procedures Must be Blind to the Researcher's Hypothesis

Quentin André

(Forthcoming, Journal of Experimental Psychology: General)

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, Nicholas Reinholtz, Uri Simonsohn, Joe Simmons, Leif Nelson, Zoé Ziani-Franclet, the Associate Editor, and two anonymous reviewers 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 results.

Outlier Exclusion Procedures Must be Blind to the Researcher's Hypothesis

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 and in a re-analysis of existing data that by-condition exclusions can result in false-positive rates as high as 43%. I finally demonstrate that by-condition exclusions are a specific case of a more general issue: Any outlier exclusion procedure that is not blind to the hypothesis that researchers want to test may result in inflated Type I errors. I conclude by offering best practices and recommendations for excluding outliers.

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 and remove 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 psychology 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)[1].

[1] While past research has highlighted that the z-score is not an appropriate metric of dispersion for identifying outliers (Leys et al., 2013), it is still very commonly used in psychology, and is therefore discussed here.

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[2].

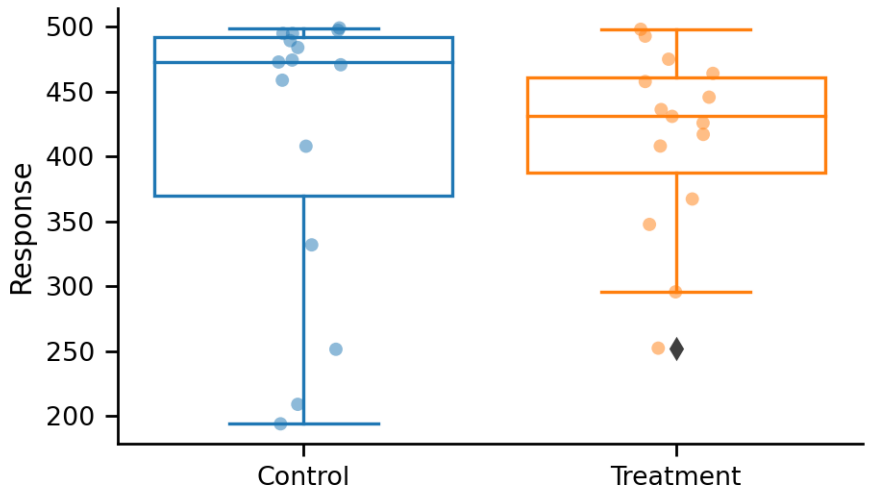

Figure 1.

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*

[2] 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.

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 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[3].

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 outlier exclusions, 9 of them are excluding outliers within the different experimental conditions[4].

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-

[3] 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.

[4] 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 procedure. The spreadsheet summarizing this search is available on the OSF repository of the paper.

hypothesis significance testing, and present evidence that this practice leads to inflated false-positive rates, both in simulated and actual data.

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[5] 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.

[5] The boxplot definition of an outlier is any data point that is lower than 1.5 times the IQR below the 25th percentile, or greater than 1.5 times the IQR above the 75th percentile.

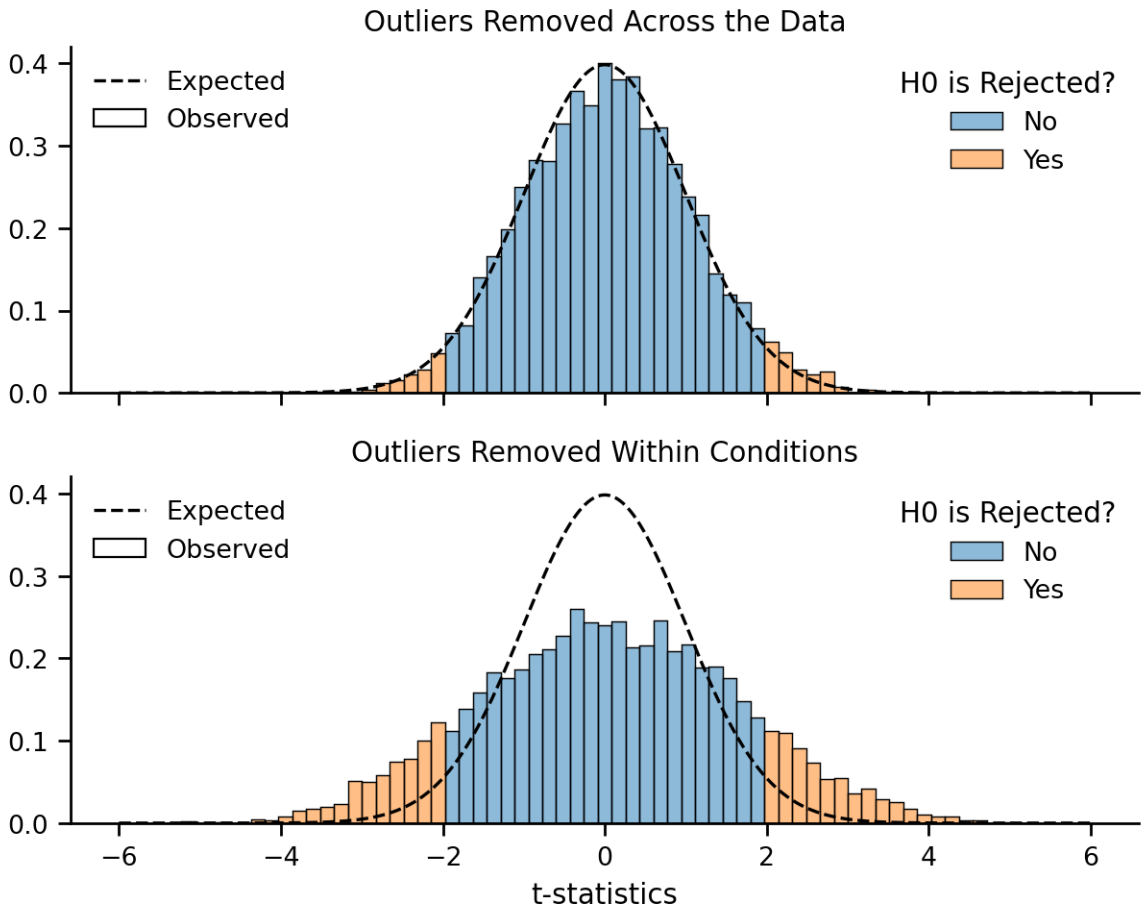

Figure 2

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 automatically violated once outliers are excluded within conditions: Each of the samples was submitted to a different data filtering procedure that *amplified* any pre-existing difference between them.

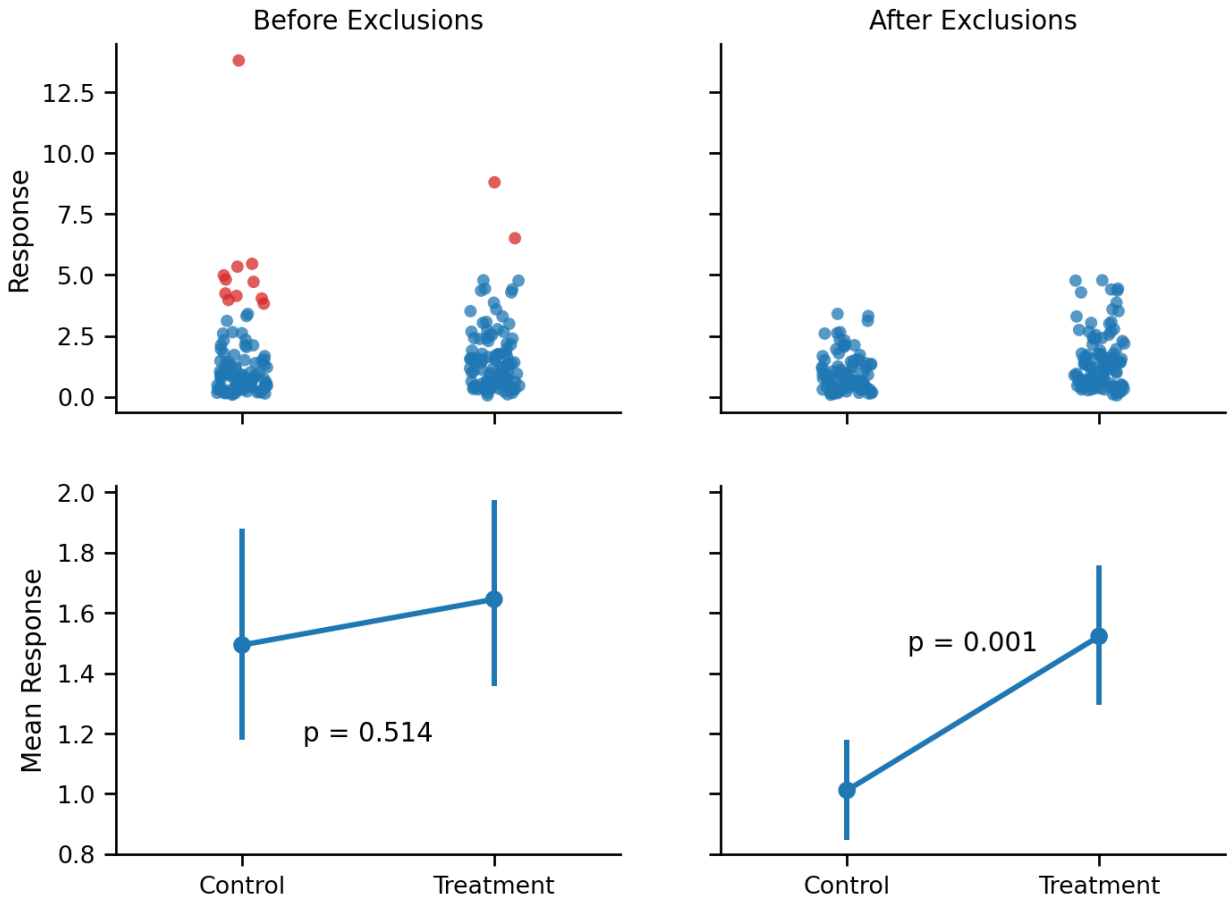

Figure 3.

Figure 3 provides one illustration of this amplification of small differences in the researchers' experiment. The two groups were drawn from the same distribution but, by chance, the values in the "Control" condition are slightly lower than the values in the "Treatment"

condition. Because of this minute difference, the same high values are flagged as outliers (red dots) by the boxplot in the "Control" condition, but not in the "Treatment" condition, and the difference between the two conditions became 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 ($p = .001$) after 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[6], 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

[6] 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)[7].

To obtain a smooth distribution of the potential outcomes, I generated 20,000 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 present the results collapsed across sample sizes and data types in Figure 4 below, and split by exclusion rules and exclusion cutoffs. The full breakdown of results is reported on the OSF repository of the paper.

[7] 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…

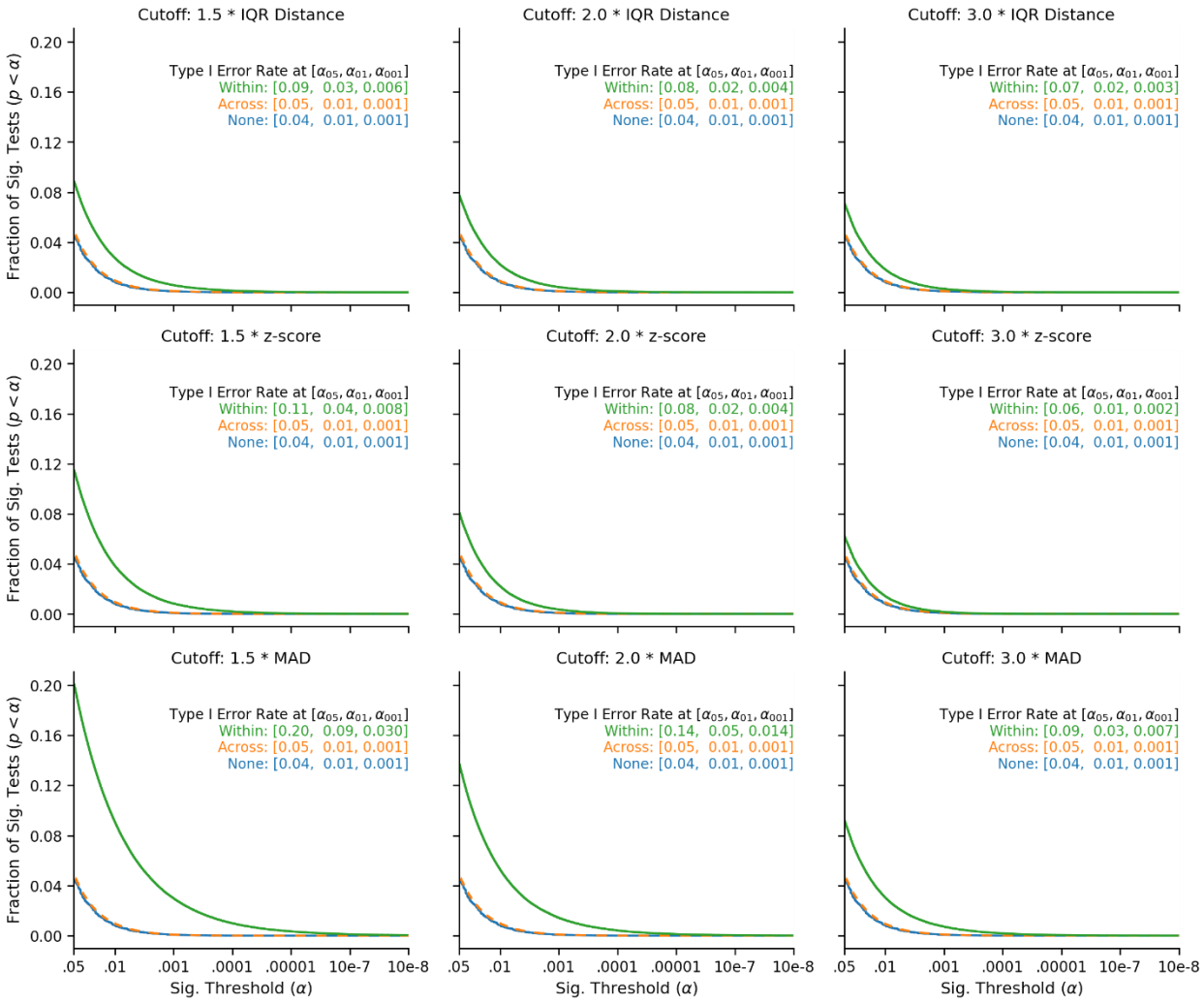

Figure 4.

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, on top of the blue line), 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 28%. 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.

AN EXAMPLE FROM RECENT DATA

The results presented so far paint a problematic picture: Analysis of simulated data suggest that excluding outliers will magnify any minute difference between conditions, and lead to high 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 by-condition exclusions can affect the conclusions of real experiments. 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 multiple reasons: It is one of the most

recent paper in a major psychological journal in which outliers were excluded within conditions, and the authors have (following best scientific practices) pre-registered their exclusion criteria and made the raw data of their paper available.

This paper describes 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. In support of the hypothesis that sharing a meal facilitates cooperation, the authors find in both experiments that dyads who were assigned to the “Shared Eating” condition have a higher Pareto efficiency than dyads who were assigned to the “Separate Eating” condition.

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. Such iterative procedure 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).

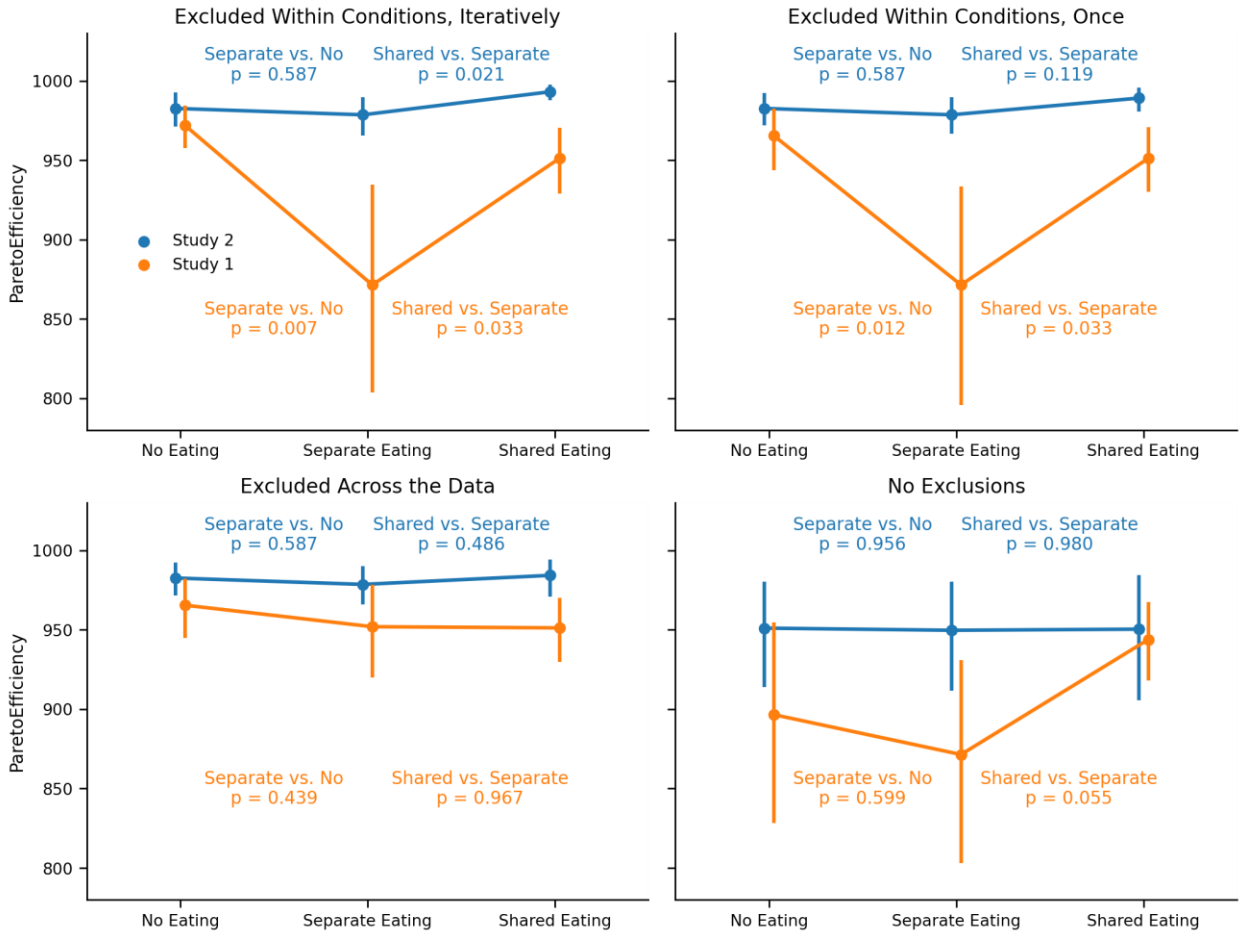

Figure 5

Thanks to the authors' sharing of their raw data, I was able to compare the results obtained in each study under different exclusion strategies (Figure 5). The upper-left panel presents the results reported in the paper: When outliers are iteratively excluded within conditions, the difference between the "Separate Eating" and the "Shared Eating" condition is large and significant. However, this difference is attenuated when only round of exclusion is performed within conditions (upper-right panel), and attenuates further when outliers are excluded across conditions (bottom-left panel) or when no exclusions are performed (bottom-right panel). This pattern illustrates the result presented earlier: Excluding outliers within conditions can magnify minute differences that were originally present between conditions.

This analysis does not necessarily mean that the result reported in the paper is a false-positive: To reach this conclusion, one would need to know whether the null hypothesis is true, and there is in fact external support for the hypothesis that sharing a meal promotes cooperation (Woolley & Fishbach, 2019). However, one can estimate, in the context this dataset, the likelihood of observing a false-positive result under different outlier exclusion strategies.

To do so, I generated 20,000 simulated samples from the data[8] by shuffling the residuals across the observations (Ter Braak, 1992). In each simulated sample, I draw two "conditions" at random, and compare their Pareto efficiency using a Welsh's t-test. Since the two conditions are drawn at random from the same data, the null hypothesis is true, and any significant difference found between the two groups constitutes a false-positive.

[8] The results reported here are based on the data of Study 2 (which was pre-registered). Similar results (reported on the OSF repository of the paper) are observed in Study 1.

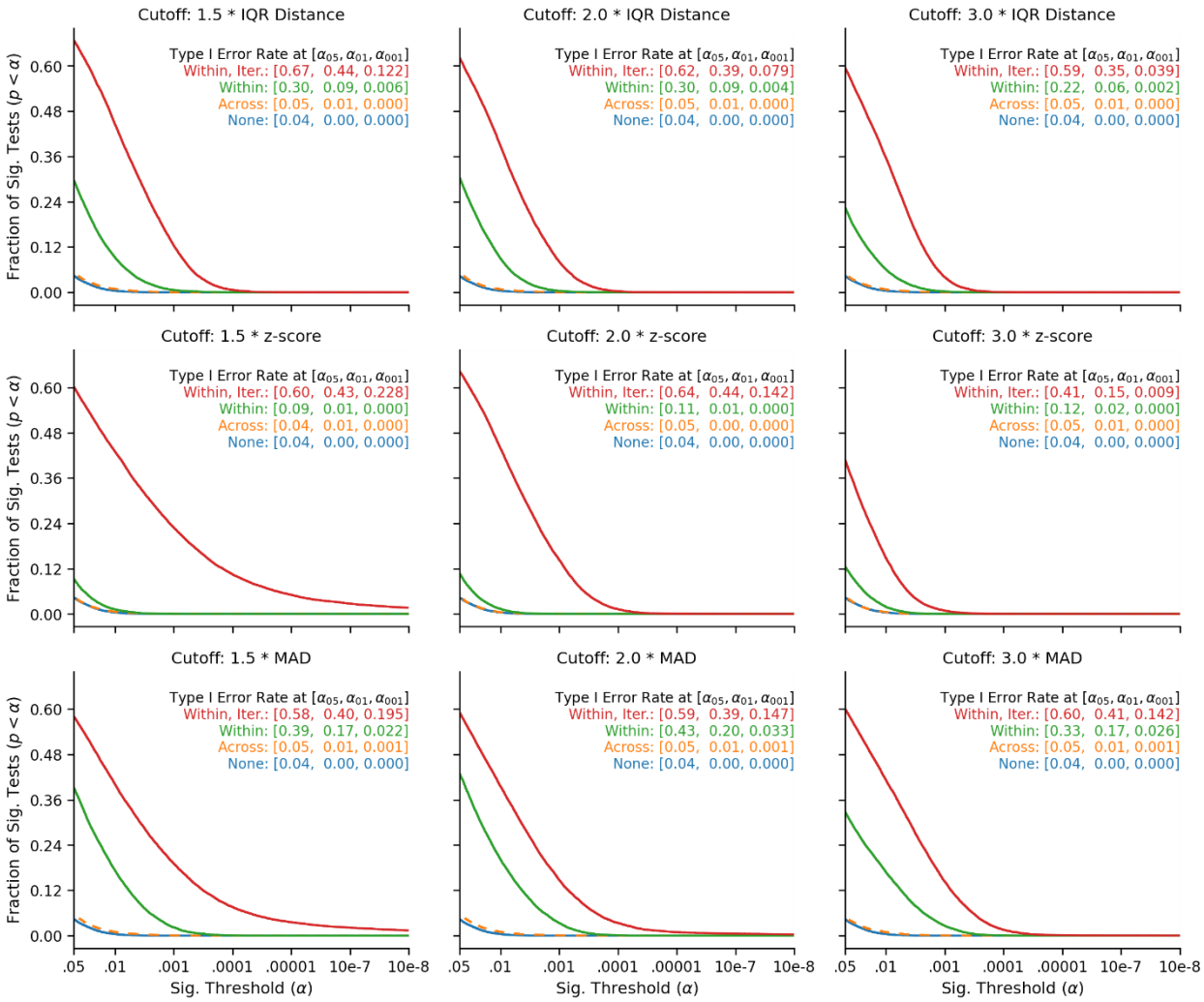

Figure 6.

Figure 6 replicates the Type I error inflation previously observed in simulated data (the exclusion criteria pre-registered in the paper, 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, on top of the blue line), the likelihood of a false-positive result increases sharply when outliers are excluded within conditions (the green line): It is always higher than 9% and can be as high as 43%.

Finally, this figure shows that the increase is even stronger when the outliers are iteratively excluded within conditions (the red line): Such iterated exclusions are causing the two conditions to further diverge, 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 59%, which translates into a false-positive rate of 29.5% for the directional hypothesis that the "shared eating" condition will have higher Pareto efficiency than the "separate eating" condition.

OUTLIER EXCLUSIONS MUST BE BLIND TO HYPOTHESIS

The present article has so far covered outlier exclusion methods based on descriptive statistics (i.e., z-score, IQR, or MAD-based cutoff) in simple one-factor experiments, and demonstrated the risks of excluding outliers within conditions (rather than across the data). In this final section, I demonstrate that by-condition exclusions are a special case of a broader issue: Outlier exclusion procedures that are not blind to the hypothesis that researchers want to test may result in inflated Type I error rates.

In the context of more complex designs (e.g., models involving interactions, continuous predictors, repeated measures…), a general approach to identifying and excluding outliers is to use the standardized (or studentized) residuals obtained from a linear model (Cohen et al., 2002; Judd et al., 2017). With this approach, any data point with a residual greater than some pre-determined threshold is excluded from the data. However, an important subtlety of this procedure is often overlooked: The model from which the residuals are computed must be blind to the hypothesis of interest.

To illustrate, consider a researcher who wants to test whether a predictor X is associated with some outcome Y, controlling for other covariates W that are known to influence the outcome. The linear model testing this hypothesis is $y_i = \alpha + \beta X_i + \gamma W_i + \varepsilon_i$, and the null hypothesis is $\beta = 0$. The researcher is concerned about the presence of outliers and will exclude any data point with a standardized residual greater than 2.

I consider two possible models from which those residuals can be computed: A "hypothesis-blind" model (in which the residuals are computed from the partial model $y_i = \alpha + \gamma W_i + \varepsilon_i$ that omits the hypothesized factor X), or a "hypothesis-aware" model (in which the residuals are computed from the full model $y_i = \alpha + \beta X_i + \gamma W_i + \varepsilon_i$ that includes the key predictor X).

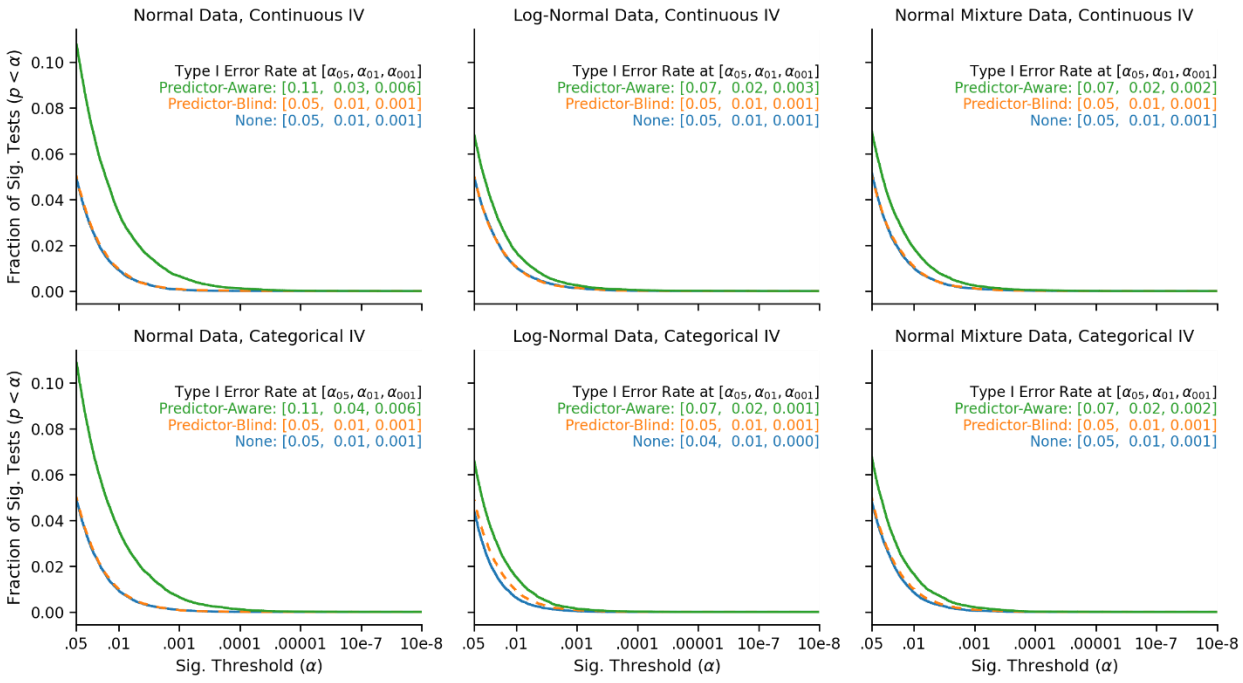

Figure 7

Figure 7 displays the false-positive rates that would be obtained with each of those two approaches, across a variety of possible datasets[9]. While Type I error rates are maintained at a nominal level when no outliers are excluded (blue line), or when the outliers are excluded using the residuals of the "hypothesis-blind" model (orange line, on top of the blue line), we see that the null is rejected too often when outliers are defined and excluded from residuals of the "predictor-aware" model (green line). Again, this result has an intuitive interpretation: When observations are excluded based on how much they deviate from a statistical model, we are effectively helping the data match the model. If the model includes the predictor of interest, this procedure will therefore amplify any relationship between the predictor and dependent variable, even when this relationship is coincidental.

SUMMARY AND RECOMMENDATIONS

A survey of the recent literature, and the common practice of splitting boxplots by conditions, both 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 and increase the likelihood of false-positive results. This pattern is observed for parametric and non-parametric tests, different exclusion criteria, different cutoffs, different sample sizes, and in both simulated and real data. Finally, I have demonstrated a more general result: Any outlier exclusion procedure that is not blind to the hypothesis that researchers want to test may result in inflated Type I error rates.

[9] I ran 20,000 simulations in each of 18 possible datasets, orthogonally varying the type of predictor X (continuous vs. discrete), the sample size (N = 50, 100, or 250) and the error structure of the data (normal, log-normal, or normal with outliers). The full breakdown of results is available on the OSF repository of the paper.

It is noteworthy that unlike most other "questionable research practices" that inflate false-positive rates (Simmons et al., 2011), the problem described in this manuscript is not driven by undisclosed flexibility in researchers' degrees of freedom. In this case, it is the procedure itself that is stacking the deck against the null hypothesis, independently of any decision from the researcher. For this reason, researchers should be particularly cautious when pre-registering their analysis: They should not only be very explicit in how they will handle outliers in their experiments (Leys et al., 2019), but also make sure that the outlier exclusion procedure that they are pre-registering is blind to their hypothesis. If researchers accidentally pre-register an outlier exclusion procedure that is not blind to their hypothesis, they will face an inflated Type I error, no matter how detailed and inflexible their pre-registered criteria were.

How Should Researchers Deal with Outliers?

First, it is worth repeating that identifying outliers and removing outliers are separate questions: The presence of outliers in one's data does not necessarily mean that those outliers should be excluded. I refer the reader to other articles that have extensively covered this distinction (e.g., Barnett & Lewis, 1994; Hawkins, 1980; Leys et al., 2019; Miller, 1993), and will assume for the sake of brevity that the researcher has already reached the conclusion that the outliers reflect erroneous or anomalous responses that should be excluded from the data.

The simplest recommendation would be to always exclude outliers in a way that is blind to the hypothesis of interest (i.e., across all the data, or based on the residuals of a model that omits all hypothesis-relevant predictors). As shown in the simulations presented in this article, this practice does not cause a Type I error inflation. If researchers (or reviewers) are uncertain about the appropriateness of a particular exclusion procedure, I would recommend that they

simulate data under the null, apply the exclusion procedure and analytical strategy to this data, and verify that the false-positive rates that they observe are nominal.

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 About Statistical Power?

The simulations reported in the paper show that, in certain circumstances (e.g., when using large cutoff values), the impact of hypothesis-aware exclusion procedures on false-positive rates might be small (e.g., from 5% to 7%). One may then conclude that such small increase in Type I error is an acceptable price to pay for a greater chance to detect an effect, and therefore decide to use hypothesis-aware procedures to exclude outliers. I believe that this conclusion is misguided for multiple reasons.

First, the connection between power and Type I error rates is a general property of statistical testing: A researcher who sets $\alpha = .1$ will detect true effects more often than a

researcher who sets α = .05, and an "hypothesis-aware" exclusion procedure (that falsely rejects the null more often than 5% of the time) will detect true effects more often than an "hypothesis-blind" procedure (that correctly rejects the null exactly 5% of the time). This apparent increase in power, however, is in part illusory: It reflects that the exclusion procedure makes it easier to obtain significant results, regardless of the truth of the null hypothesis.

Second, the small increase in false-positive rates observed in certain simulations should not obscure an important result of the paper: The Type I error rate inflation is highly heterogeneous across data types, statistical analysis, and outlier exclusion strategies, and the results presented in the simulations are not necessarily predictive of the Type I error inflation that a researcher would observe when using "hypothesis-aware" exclusions. A direct consequence of this heterogeneity is that researchers cannot determine how much Type I error they are sacrificing and cannot be certain that Type I error increase will be small in the context of their experiment.

Third, the Type I error inflation at α = .05 is a symptom of a more general issue: Since hypothesis-aware exclusions violate the logic of null-hypothesis significance testing, the p-value no longer represents "the likelihood of a result as extreme as that of the experiment under the null." To illustrate this point, consider the upper-right panel of Figure 7. This simulation shows that "hypothesis-aware" exclusions result in a relatively small Type I error inflation when setting α = .05 (40%; from 5% to 7%), in a larger Type I error inflation when setting α = .01 (80%; from 1% to 1.8%), and in an even greater inflation when setting α = .001 (140%; from .1% to .24%). This distortion of the entire distribution of the p-value not only limits our ability to interpret the amount of evidence presented in a single study, but most importantly invalidates the conclusions of meta-analytical tool that relies on p-values as a continuous measure of evidence (e.g., the p-curve; Simonsohn et al., 2013), Ultimately, the cumulative risk for false discovery across

multiple studies and papers is greater than what the Type I error inflation at $\alpha = .05$ would alone suggest.

Finally, there are superior alternatives available to increase statistical power in outlier-prone experiments. As mentioned earlier, researchers may first consider using "trimmed" estimators (e.g., Kim, 1992; Wilcox, 2011; Wu, 2006; Yuen, 1974) that increase statistical power while keeping Type I errors at a nominal level. Alternatively, researchers may consider adopting a different alpha level, as proposed by Lakens and colleagues (2018). This approach also trades off false-positive rates for power, but does so in a transparent way (i.e., the exact Type I error rate of the experiment is known to the reader and to the researcher) and does not disturb the distribution of the p-value under the null.

What if the Pattern of Responses Does Differ as a Function of the Predictor?

It might be tempting to justify the practice of excluding outliers by conditions by the observation that the different conditions *do look* different: One condition appears to have a higher mean, or a higher dispersion, than the other(s). As mentioned earlier however, this justification is a paradox: If we assume that the pattern of responses differs 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. The same logic applies to more complex designs, or to continuous predictors: If a predictor is known to influence the outcome, there is no need to test for its significance, and it is appropriate to rely on

it to identify outliers. On the contrary, if researchers want to test whether a predictor influences the outcome, then the procedure used to identify outliers must not rely on this predictor.

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. Again, 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 860% (from 5% to 43%), 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 amplification of small 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 to the frequentist confidence interval) contains zero more frequently when exclusions are performed within-condition exclusions (vs. across the data).

CONTEXT OF THE RESEARCH

This paper is part of a stream of research aimed at improving scientific practices in psychology and behavioral science. While working in this area, the author realized that false-positives might not only result from uncontrolled researchers' degrees of freedom, but also from the inappropriate application of statistical procedures and statistical tests. The problem of outlier exclusions presented in this manuscript is one such example. The author hope that the results and recommendations presented in this paper will help correct a common misconception on how outliers should be handled.

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

André, Q. (2020). Outliers Exclusion Procedures Must be Blind to the Researcher's Hypothesis. *OSF Repository*. https://osf.io/3tz76/

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

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3e édition). Routledge.

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.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data Analysis : A Model Comparison Approach To Regression, ANOVA, and Beyond, Third Edition* (3e édition). Routledge.

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-Curve : A Key to the File-Drawer. *Journal of Experimental Psychology: General*, 14.

Ter Braak, C. J. (1992). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related techniques* (p. 79-85). Springer.

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.

Woolley, K., & Fishbach, A. (2019). Shared Plates, Shared Minds : Consuming From a Shared Plate Promotes Cooperation. *Psychological Science*, *30*(4), 541-552. https://doi.org/10.1177/0956797619830633

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